Universiteit
Leiden
The Netherlands

The lazy mindreader : a humanities perspective on mindreading and multiple-order intentionality

Duijn, M.J. van

Citation
Duijn, M. J. van. (2016, April 20). *The lazy mindreader : a humanities perspective on mindreading and multiple-order intentionality*. Retrieved from https://hdl.handle.net/1887/38817

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/38817

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/38817 holds various files of this Leiden University dissertation

Author: Duijn, Max van
Title: The lazy mindreader : a humanities perspective on mindreading and multiple-order intentionality
Issue Date: 2016-04-20

Chapter 1

The bigger picture: language, narrative, and social cognition

Any mentioning of *intentionality* or *intentional states* comes with an interesting assumption: that we can speak meaningfully about the inner lives of others and ourselves. Indeed, everyday language is packed with "mentalistic" expressions of the type: "I *know* what you're after", "he *thinks* she's married", "she *believed* that he *feared* nothing more than that", etcetera, and in numerous contexts it is perfectly unproblematic to speak and reason in this way. At the same time, shifting from a run-of-the-mill perspective to one of philosophical and scientific inquiry, one may legitimately ask: what *do* we know about someone else's beliefs, thoughts, intentions, desires, fears, and so on? And what ways do we have available to form an understanding of this? A different question may be: why would we bother at all?

These questions have been central to the research into *social cognition*, the sort of cognition required for living in groups structured by social bonds and networks. As mentioned in the Introduction, researchers from a wide array of disciplines have contributed to this area, most notably psychologists, philosophers, anthropologists, ethologists, and neuroscientists. Most attention has been focused on the skill referred to as "mindreading" (also variously called "theory of mind", "mentalising", or sometimes "folk psychology"; see below), the capability to assess others' intentions, knowledge states, motives, etcetera—in short: their intentional states. As also set out in the Introduction, the main endeavour of this thesis consists in analysing the nature of the complexity involved in dealing with multiple intentional states that are mutually linked and/or embedded, as required by various aspects of our social and cultural lives, and investigating ways in which we handle such complexity linguistically and cognitively. Evidently, properly addressing the issue of handling *multiple* intentional states first requires knowledge of what it entails to form an understanding of just *one* intentional state. This chapter will start with a discussion of several possible views on this question, as given in the wider area

of research into mindreading. Next, a brief status quaestionis of research into multiple-order intentionality will be provided. The chapter will end by considering the links between, on the one hand, mindreading and multiple-order intentionality, and, on the other hand, language, narrative, and various aspects of social interaction more widely, as discussed throughout the literature on (primate) sociality and social cognition.

1.1 Mindreading and intentionality

1.1.1 Mindreading

Research into mindreading easily attracts attention, though not always for the right reasons. It all too often conjures up associations with myths, fairy tales, science-fiction stories, or even with fortune tellers and crystal gazers of the suspicious sort. In modern science there is of course a complete consensus that there is no magic involved in the way we form understandings of others' inner lives—however, anyone trying to come to grips with the extensive literature on mindreading that has emerged over the past decades might well form the suspicion that this is indeed the *only* consensus. To give a (rough and preliminary) impression: some research traditions have pictured a dedicated mindreading "module", forming the quintessence of the human mind (see e.g. Saxe, 2006). Others, by contrast, have conceived of mindreading rather as an "umbrella term" for a set of diverse tricks, strategies, and mechanisms that we use to make sense of the behaviour of ourselves and others around us (e.g. Apperly, 2011). Some have emphasised the role of brain functions specialised for mindreading (e.g. Carruthers, 2004), others have suggested that we use only general cognitive skills (e.g. Heyes, 2014). Some are particularly interested in the aspects of mindreading that are uniquely human, others emphasise their deep roots in our primate (or even mammalian) nervous systems (e.g. De Waal, 2013). Some maintain that mindreading relies on innate competencies (e.g. Fodor, 1983), whereas others stress that the most important parts of mindreading are learned in the course of growing up in our typically human socio-cultural

environments (e.g. Heyes and Frith, 2014). According to some researchers, mindreading is highly "enactive" and performed by using our entire body for modelling someone else's perspective (e.g. Gallagher, 2008), while others have suggested that we run simulations of what people around us feel and think using "mirror systems" in our brains (e.g. Gallese and Goldman, 1998). At the same time, defenders of an "inferentialist" understanding of cognition have suggested that we form representations of and theories about the inner lives of others instead of running simulations of any form (e.g. Gopnik and Wellman, 2012). Advocates of the "narrative practice hypothesis", in turn, have argued that mindreading most often relies neither on simulation nor theorising, but rather on structural and semantic knowledge of folk-psychological narratives (e.g. Hutto, 2008). Several of these positions will be detailed and built on below.

At the outset of his monograph *Mindreaders* (2011), Apperly provides a comprehensive overview of the main questions and debates that have occupied researchers of mindreading over the past decades. His book focuses mostly on explaining how mindreading works in terms of its underlying mechanisms, which means that it operates for the largest part on Marr's H-level ("how", "through which mechanisms"; see the Introduction). In line with the purposes of this thesis, the discussion in this chapter is mainly focused on the W-level ("what"): it aims at setting out a workable "task model" of mindreading by discussing the elements and stages of its process and the conditions under which it operates. Nonetheless, this chapter also contains several sections pertaining to the mechanistic and physical levels of explanation, as parts of introducing the larger field of research.

Stripped down to its basic outlines, the task model set out here features five elements:

(i) the mindreader;
(ii) the mindreadee;
(iii) cues;
(iv) intentional states (which can be called "mindreads" once the mindreading process has taken place); and
(v) an inferential process through which (iv) is derived from (iii)

I will begin by discussing an introductory example, exploring some of the issues and terms that will be revised and built upon in the sections that follow. Consider the following photograph:

Figure 1

A normally developed adult person standing in the position of the photographer (the mindreader) will most likely feel inclined to give the person on the staircase (the mindreadee) a helping hand. The term used for the basis of this inclination by Frans de Waal (2005) is *emotional contagion*:[7] we see the facial expression and posture of the person carrying out a heavy task and due to the deeply-rooted empathic tendencies we have as primates, we cannot even help but feel some of the burden ourselves, which triggers the impulse of providing targeted help (more details and alternatives will be discussed below in Section

[7] "Emotional" is here to be understood not in the narrow sense of the "basic emotions", but rather as the broader category of feelings including, for instance, pain, grief, agitation, relief, sorrow, embarrassment, surprise, and so on (De Waal, 2005: 46-47).

1.1.4). Whether this help is in the end provided will clearly depend on many factors, such as individual features of the mindreader, relationship to the mindreadee, local cultural rules, and more. However, it is not hard to see that Figure 1 above depicts a situation in which a helping hand would *in principle* be appropriate. Now consider the following photograph:

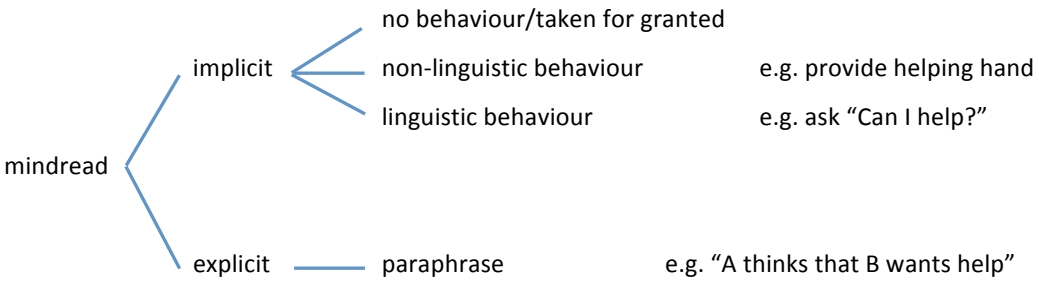
Figure 2

The weightlifter depicted in Figure 2 shows a posture and facial expression considerably similar to that of the person in Figure 1 and the tasks faced by both are also much alike: lifting a heavy object. Yet it is clear that this man would show much surprise, if not severe agitation, if the photographer or any of the other bystanders would offer a helping hand—and indeed, none of them shows the least inclination to come to his aid. (A caption making this point in a different way would be: "Why is this *not* an example of the bystander-effect?") Even if the weightlifter were to look up and ask for a helping hand, the mindreader would probably start mining the situation for opportunities to provide assistance with anything *other than* just lifting the weight bar, such as a

loose shoe lace or an unfastened spring collar at one of the bar ends—at least, that is, as long as the weightlifter gives the impression that he has the situation under control. This highlights another crucial component of the inferential process: *background knowledge*—not only emotional contagion influences the mindreader's decision on whether to take action or not, but also knowledge about the situation the mindreadee is in. In principle, one could imagine that someone completely unfamiliar with the context of a gym would hasten to help lifting the weight bar (some philosophers would suggest an empathic Martian; others would perhaps suggest an unworldly philosopher). He would pick up some of the burden felt by the mindreadee, while being unable to recruit the relevant background knowledge about what the possible *scenarios* are in this context. Someone who does know the context of a gym, by contrast, does have such scenarios available: the weightlifter wants to test or train his strength, or possibly show off to the bystanders. Clearly, in these scenarios help is highly unwanted. If he were lifting the bar from his car boot, though, a helping hand might again fit.

The decision about which behaviour is appropriate in the situations depicted by Figures 1 and 2 relies on what I here call a *mindread*: the assessment made of someone's intentional state in the context of a (real or imagined) social interaction event. Such an assessment is made on the basis of *cues*, which can be of virtually any nature. What does or does not count as a cue can only be defined from the perspective of the mindreader: it includes any observable aspect exhibited by the mindreadee and his or her "situatedness" in the context of the interaction that is used in the mindreader's inferential process. The cues are interpreted in the light of relevant background knowledge, recruited from the mindreader's memory. In practice, the resulting outcome, i.e. the mindread, can be made explicit or remain implicit, and can be taken for granted or factored into the planning of future behaviour. This behaviour can be linguistic (the mindread can guide form and content of an utterance or response) or non-linguistic, as would be the case when providing a helping hand to the person in Figure 1. Moreover, for purposes of analysis or reflection it is possible to form explicit paraphrases of a mindread, for example: "this person *intends* to lift the suitcase in order to get upstairs, so *will appreciate* a helping hand" or "A *thinks* that B *wants* help" in the case of Figure 1. As will be discussed in Section 1.1.3

below and in Chapter 2, it is important not to confuse such paraphrases with the cognitive processing required to make appropriate inferences, and with representations of intentional states as they appear "in the wild" of actual discourse. The distinctions now made can be summarised as follows:



Sections 1.1.2 and 1.1.3 will further elaborate on element (iv) of the list above: intentional states. In Section 1.1.4 element (v), the inferential process, will be discussed in more detail.

1.1.2 Intentional states

The concept of *intentionality* (not to be confused with the "intentions" we have when we want something to happen) has a rich history in scholarship and sciences of the mind. After its presumed origin in medieval scholastics,[8] the concept was most famously developed in the nineteenth century by Franz Brentano (1995 [1874]), as a part of debates now considered foundational for the emergence of psychology as an academic discipline, and in the twentieth century by Daniel Dennett (1971; 1987) and John Searle (1983), in work that was influential in the still-ongoing trend in psychology and the cognitive sciences to study mindreading. Brentano used the concept of intentionality to define the difference between mental and physical phenomena. In brief, his distinction boils down to the claim that physical phenomena have an autonomous existence, whereas mental phenomena necessarily are *about* something—they do not exist independently of their intentional object. In Brentano's words: "in

[8] According Chisholm's (1967) entry in *The Encyclopedia of Philosophy* a very similar concept was already present in Saint Anselm of Canterbury's 1078 treatise on the existence of God, but the term was coined later and goes back to the scholastic notion of "intentionalitas".

presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on" (1995 [1874]: 88). This "something" that is presented, judged, loved, etcetera, is the intentional object on which each intentional state depends for its existence.

Making a distinction between mental and physical phenomena in this way has implications for how the ontology of mental phenomena is construed. Therefore, work on intentionality has always been connected to fundamental philosophical debates regarding "dualism" and, more recently, the mind-body distinction and extendedness of cognitive processes. Brentano seems to accept a meaningful divide between the physical or material world, in which objects exist *as objects*, and the mental world, which includes non-material phenomena that are *about objects*.[9] From a non-dualist, "materialist" viewpoint this position is problematic, since mental phenomena are being viewed as a part of the physical and material world in no other sense of the words. Within the materialist view, a distinction can again be made between, on the one hand, theorists who conceive of cognitive processes as neurons in the brain "dancing" in a particular way, and, on the other hand, theorists who argue that cognition is mostly distributed over the entire body, or even beyond that: over the environment. Defenders of this latter position, which is known as "extended cognition", allow for combinations of, for instance, people, machines, books, and/or cultural practices to be included in their definitions of cognitive processes. Some more attention to the issue of embodied cognition will be paid in Chapter 2, where mindreading is considered in the context of drama and dialogue. In Chapters 4 and 5, I will argue that part of the burden of processing

[9] Note that this does not mean that he argues that physical phenomena can be experienced unmediated by our senses: he sees the world as being entirely mediated by subjective experience; however, some parts of our experience relate to physical phenomena and others to psychological phenomena (see Zahavi, 1992: 30). In this sense, Brentano can be called a dualist, but not a Cartesian. See also Dennett (1987: chapter 10).

complicated mindreading tasks can be alleviated by cultural and linguistic "thinking tools".[10]

Dennett, who is an explicit defender of cognitive materialism, bypasses much of this controversy by speaking of "intentional systems" (1971; 1983). Whether a cognitive system is construed as an immaterial mind, a group of interconnected neurons "dancing" in a particular way, an entire body, or as two people and a calculator, in all cases it can be seen as a *system capable of intentionality*, in the sense that it can enter a state in which it is *about* an object. This object should be taken in the broadest sense of the word, including, for instance, a ball in the mouth of a dog, a comic figure on a computer screen, a picnic in a short story by a price-winning writer, or the creator of the universe. Whether the object exists in some form outside the realm of the intentional state is, in this view, not part of the question: intentionality is seen as a property that a system can have, regardless of how this system is realised and regardless of the ontological status of the object. In this way, and as mentioned in the Introduction above, an intentional system can enter a state in which it is about *another* intentional system, potentially also exhibiting an intentional state. For example, a human mind can be about another human mind's intentional state, say, the other's desire for a glass of water, intention to cooperate, or understanding of her brother's love for his daughter. Such embedded or multiple-order intentionality will be discussed further in Section 1.2 below.

Characteristic of intentional states, then, is that they have a "dependent" or "extending" nature: when considering an intentional state, one necessarily also has to take into account the object this state is about. A traditional way of studying this is using "logical" propositions. Consider the following expressions:

(1) John believes that it is raining outside.

[10] See Dennett (1987: chapter 10) for a taxonomy and discussion of various theories about intentionality and their implications for the ontology of mental phenomena. Although explicit defence of a classical "Cartesian" dualist view is rare in modern philosophy and science, implicit assumptions referring to this view can be traced in many works and research traditions (see also Dennett, 1991: chapter 2 and 5). The latter point is also demonstrated in Sorem's (2010) discussion of Searle (1992): Sorem argues how Searle's "simple solution" (1992: 1) to the mind-body problem yields inconsistencies, in a way that is illustrative of how deeply rooted dualistic views are in everyday thinking as well as in specialised philosophy.

(2) Mary intends that John believes that it is raining outside.
(3) Mary talks to John and it is raining outside.
(4) It is raining outside and Mary intends that John has another cup of tea.

They can be rewritten into propositions as follows:

(5) A believes that p
(6) B intends that A believes that p
(7) p and q
(8) p and Mary intends that q

Note that clauses expressing intentional states are rewritten using the form "A [intentional expression] that p", whereas other clauses, referring to objects (in the broad sense, so including events, states of affaires, people, etc.), are rendered as single symbols (p or q). This reflects the structural property of intentional states discussed above: they are not independent, but reflect a *relationship* between an intentional being and a non-intentional object. In logical terms: a clause expressing an intentional state induces "referential opacity", as can be shown by the so-called "substitution test" (Dennett 1983: 344-345). In a proposition describing a particular state of affairs in the world, it is usually possible to substitute words with other words that refer to the same entity without consequences for the truth-value (or even referential value) of the sentence. In other words: "this rule is simply the logical codification of the maxim that a rose by any other name would smell as sweet" (Dennett, 1983: 344). To give an example: provided that *Macbeth* and *Hamlet* were written by the same author, it should be possible to substitute "the author of *Macbeth*" with "the author of *Hamlet*" in a proposition, without the truth-value and referential value being affected. The propositions (9) and (10) are thus either both false or both true:

(9) The author of Macbeth was born in Stratford-upon-Avon
(10) The author of Hamlet was born in Stratford-upon-Avon

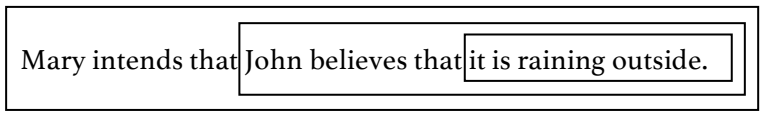
However, in the following two propositions this is not necessarily the case:

(11) John believes that the author of Macbeth was born in Stratford-upon-Avon

(12) John believes that the author of Hamlet was born in Stratford-upon-Avon

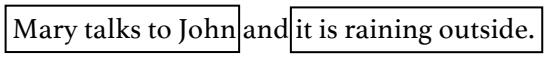
After all, what John does and does not believe is independent of the "real-world fact" that the author of both pieces is the same person. This is what Dennett calls referential opacity: "the terms in such clauses are shielded or insulated by a barrier to logical analysis, which normally "sees through" the terms to the world the terms are about" (1983: 345).[11]

The take-home message from the substitution test is really that there is a relation of dependency between the intentional agent and the non-intentional proposition. In the case of "John believes that it is raining outside", the intentional agent (John) and the intentional expression (believes that) are the responsibility of the speaker, the one who asserts the proposition, whereas the non-intentional proposition (it is raining outside) is placed under the responsibility of the staged intentional agent (John). As a consequence, it is "insulated", in Dennett's words, "shielded from logical analysis" (1983: 345). In "Mary intends that John believes that it is raining outside", there are two such dependency relations: "John believes that it is raining outside" falls as a whole under the scope of "Mary intends that", and "it is raining outside" falls under "John believes that". Schematically:

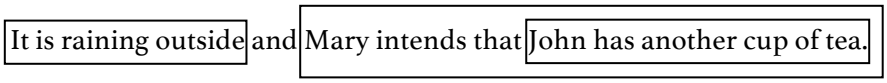


[11] There is a significant difference between this "logical" approach and the natural-language view mostly taken throughout the rest of this thesis. Instead of looking at references to the "real world", in this latter view the focus is on a speaker having a certain rhetorical goal (in this case presumably informing his interlocutor that John believes that the author of a particular piece was born in Stratford-upon-Avon). With a view to achieving this goal the speaker invites his interlocutor to consider the perspective of John, which he does (in accordance with local linguistic conventions) through the usage of a complementation construction (see Chapters 4 and 5 in particular).

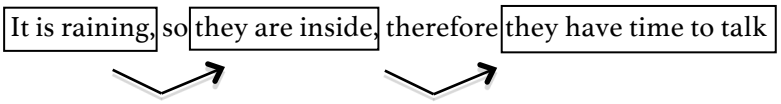
By contrast, in the case of a conjunction of two non-intentional propositions, there is no dependency in this sense:



Clearly, when an intentionality proposition is combined with a non-intentionality proposition, the latter is independent of the first:



Another aspect of intentional relations is what could be called their *non-transitivity*.[12] The following proposition features three clauses that are causally related:



In principle, it is possible to leave out the middle clause without violating the chain of causality expressed by this proposition:

In other words, if a proposition expresses "p so q therefore z", it follows that "p therefore z" is also true—causal relationships could therefore be called "transitive". By contrast, if one clause is left out of a chain of intentionally related clauses, it does not follow that the produced clause has the same truth-value: "B believes that A believes that p" does not entail that "B believes that p". The points made in this section will be of practical use when analysing questions from mentalising experiments in Chapter 6. Their theoretical importance will become clearer in Section 1.2 below and in the Chapters 2 and

[12] In formal logic transitivity is a property of certain relational predicates, such as ancestry. If A is an ancestor of B, and B is an ancestor of C, then it follows that A is also an ancestor of C (see e.g. Forbes, 1994: 275).

3; in fact, it can be said that the non-transitive, dependent nature of the relationships exhibited by multiple-order intentionality propositions is one of the core aspects of the problem dealt with in this thesis.

1.1.3 The intentional stance

So far in this thesis, different intentional relationships have been categorised using mentalistic expressions from everyday language, such as thinking, knowing, believing, desiring, intending, and so on. How can we be sure that these terms are appropriate? Do they correspond to the *actual* intentional states held by others around us? Or, for that matter, to those held by non-human animals? According to Dennett (1983; 1987) we do not need to be sure. He argues that in order to understand phenomena in the world, one can adopt various strategies or "stances", corresponding to different levels of theorising (partly overlapping with Marr's levels of explanation, as will be pointed out below). For example, for understanding why an analogue alarm clock rings, one could take the "physical stance" and aim at figuring out how, given the laws of physics, its springs exert particular forces on a system of cogwheels and axes, eventually triggering a clapper that hits a bell, which causes movement of air at particular frequencies, and so on. One could also take the "design stance", looking at what the clock was designed to do when setting it to a particular time and switching on the alarm function. Alternatively, one can adopt the "intentional stance", not towards the alarm clock itself, but to the intentional system who set it: one can question why someone has set it to this particular time and what he or she *intended* hearers of the alarm sound to *think*. If we decide to adopt the intentional stance, this means that we try to make sense of an intentional system's behaviour by assuming that it was underlain by intentional states. In this view, the intentional states are really "in the eye of the beholder", and their use is to understand a phenomenon in the world better.

Intentional systems can be humans, but also other animals, for example: "that fox digs a hole because it *wants* to build a nest" or "bird X *believes* that bird Y is hiding food". Dennett argues that usage of everyday language is not problematic in such cases, as long as one keeps to the appropriate level of theorising. To use an adapted version of his own example: when researchers

interested in the behaviour of a particular bird decide to call a certain berry in the bird's environment "food", they abstract from all kinds of biological and chemical details of nutrition and digestion. Biologists interested in such details might choose to refer to the same berry in terms of its composing sugars, acids, proteins, etcetera. Even if the latter research were still in an early stage and little would be known about the biochemical details of nutrition from this berry, those interested in the foraging behaviour of the bird could safely refer to it as "food" in their theories. Similarly, one can perfectly well make use of everyday mentalistic vocabulary as long as one is dealing with questions of some beings' behaviour in their social environments, and not with the "lower-level" mechanisms and physical processes underlying social living.[13]

In some ways this is reminiscent of the distinction made between "explicit" and "implicit" mindreads at the end of Section 1.1.1. Normally developed human adults can surely reason about mental concepts in an explicit way: they can think or talk about themselves or others in terms of their beliefs, desires, fears, etcetera, thereby explaining or motivating particular behavioural moves and choices. However, this does by no means entail that mindreading in the practice of social interaction constantly uses explicit mental concepts. Apperly's (2011: 3) analogy in the physical domain is the curve described by a ball thrown into the field: although we can in principle reason about angles, velocity, friction, and so on, in order to predict where it will land, this hardly reflects how we manage to make a catch in practice. The explicit mental concepts and linguistic paraphrases could be seen as the formulas dealing with angels, velocity, and friction: although they can in principle be used to form a mindread, they hardly reflect our reasoning in most cases of everyday interaction (more on mindreading and linguistic explications will follow in Section 1.2.2).

Note that Marr's levels of explanation can again do useful work in this context: both adopting the intentional stance ("the fox *wants*...") and engaging in explicit mindreading ("A *thinks* that B *intends*...") are unproblematic on

[13] See Dennett (1983: 344) for his version of this example. Related to this is the debate over the question whether the everyday mentalistic terms should be seen as temporary placeholders, used only until their "real" neurological correlates are figured out, or whether they have a different status. See Dennett (1987: ch. 10) for a discussion of his own perspective and various alternatives.

Marr's W-level, the level of "what the task is". Terms such as thinking, knowing, desiring, and so on, work fine when used to describe what a mindreading process is about and why it is taking place. However, this may change as soon as one is dealing with questions on Marr's H-level of the mechanisms at work or the physical level describing what machinery is used: it is likely that intentional terms such as thinking, knowing, or desiring have no role in, say, what drives a fox to dig a hole or what makes a bodybuilder lift a heavy weight bar on the physical level.

In the next section, I will offer a brief discussion of various hypotheses operating at the level of how the inferential processes underlying mindreading are carried out in practice, thus switching to the H- and physical levels.

1.1.4 The inferential process: theory, simulation, narrative practice

The concept of "theory of mind" goes back to discussions about the cognitive features and limitations of our close primate relatives in the 1970s. The foundational paper on this concept was published by Premack and Woodruff in 1978. They discussed experiments in which a chimpanzee, Sarah, was shown videotaped attempts of a human actor to solve particular problems, such as trying to grab a banana that was placed out of reach. Sarah (who, incidentally, was said to be familiar with the video screen from watching "commercial television") had little trouble matching the videotaped problems with photographs that pictured the "right" solution out of several options, such as using a long stick to bring the banana closer. Interestingly, it seemed to matter to her that the actor in the videos was her favourite trainer, for whom she clearly felt affection, since when the same part was played by another, less well-known acquaintance of hers in subsequent test rounds, she quite consistently chose photographs showing "bad" alternative solutions leading to "untoward outcomes" (Premack and Woodruff, 1978: 521).

It is worthwhile looking into this early paper in some detail, as it already addresses a few important issues that have since been discussed extensively in the literature on mindreading, and are mostly still under debate. After establishing that the results could not be explained by mere "physical matching" of objects, the authors discuss three possible explanations of how

the chimpanzee could have managed to match the videotaped problems to the right solutions: "associationism", "theory of mind", or "empathy". According to the first explanation, she would have solved the problem on the basis of familiarity with relevant sequences of action; in the authors' words: "when shown a sequence that one recognizes, but that is incomplete, one chooses the element that has the effect of completing the sequence" (1978: 516). The authors point out that Sarah most likely had similar experiences from her own daily life in the lab, but that she was not familiar with the exact sequences of action used in the experiment and that there was sufficient reason to assume that the presented problems contained at least some novel elements for her. They therefore grant the "associationism" explanation some credibility, but regard it insufficient to account for the totality of their findings.

The second explanation, "theory of mind", is described by Premack and Woodruff as follows: "In looking at the videotape, [the chimpanzee] imputes at least two states of mind to the human actor, namely, intention or purpose on the one hand, and knowledge or belief on the other" (1978: 518). They thus suggest that, according to this explanation, Sarah somehow went through the following strand of reasoning:

(i) "the human actor *wants* the banana and is struggling to reach it";
(ii) "the actor *knows* how to attain the banana"; and
(iii) that will lead to the situation depicted in photograph X rather than photograph Y (1978: 518; italics added).

The authors consider the third explanation, "empathy", to be identical as far as step (i) is concerned, but different for step (ii) and (iii). After imputing to the actor the intention to grab the banana (step (i)), according to the "empathy" explanation Sarah would put herself "in the place of the actor" (1978: 518) and choose the alternative consistent with what she would do in that situation.

Given that it mattered to Sarah's choices whether she saw her favourite trainer or a more removed acquaintance on the video screen, Premack and Woodruff favour the "theory of mind" explanation over the "empathy" one: after all, if she would picture *herself* in the presented situations, the actor's identity should not matter. They argue that this does not exclude "associationism" to play a role as well, and conclude as follows:

> in highly familiar situations, one's expectancies are based on existing associations. [...] In novel situations, however, one's expectancies are generated, we think, from theories, and are not the product of associative generalization. [...] There may also be developmental and inter-species differences in this regard. Young children and lower species may form expectancies by associative mechanisms, the former having yet to build any theories and the latter probably unable to build them; whereas adults and higher species may largely generate them from theory. (1978: 518)

Extensive debates followed this early discussion, partly revolving around questions of which primate species had such theory-of-mind abilities and to what extent, and for another part focusing on analogous competences in humans. Later on, important contributions were made by developmental studies and research addressing certain psychopathological disorders. It seemed that some people suffering from disorders along the autistic spectrum were well-described as having impaired theory-of-mind abilities, which was generally taken as a strong indication that there must indeed be some part or network in the human brain responsible for theory of mind (after all, "if it can break, it must be there"). Another boost to the field was given by the advances made in the neurosciences during the 1990s and 2000s, including also the discovery of mirror neurons, neural networks that are involved both when an action is performed and when this same action is observed in someone else.[14]

Notwithstanding the importance of all these subsequent findings and contributions to the debate, support for all three explanations given by Premack and Woodruff for how the inferential process works persists to the present day in one form or another. The aim in the remainder of this section is to offer a typology of the dominant positions in the current field. By way of illustration, work of proponents of each of these positions will be referenced and discussed in brief, but these discussions must by no means be taken to be exhaustive. For more comprehensive overviews, providing more extensive lists

[14] See Di Pellegrino et al. (1992) for one of the initial papers on the discovery of mirror neurons; for a full discussion see Pineda (2009). For an overview of neuroscientific research into mindreading see Frith and Frith (2006). For an overview of literature on mindreading and psychopathology see Baron-Cohen et al. (2013); Bird and Viding (2014). For mindreading across the primate world see Byrne and Whiten (1988; 1997); Rosati and Hare (2010); Whiten (2013).

of references to advocates of the different positions see Apperly (2011: especially chapters 2 and 7), and see Hutto (2008: especially chapters 8 and 9) for a critical perspective.

Theory-theory

Proponents of what is referred to as the "theory-theory" hypothesis suggest that mindreading relies on folk-psychological theories, generally held to comprise knowledge of rules and patterns of how social beings behave, and why they behave like they do, with a certain degree of *abstraction*. So mindreading competence is hypothesised to rely not (just) on knowing a collection of sequences of action that can be used as "exemplars" when making sense of new cases, but rather on more abstract rules and patterns that can be used to generate predictions of future intentional states and/or behaviour. There are multiple versions of the theory-theory hypothesis in circulation, primarily differing in two dimensions: "specialised versus domain-general", and "innate versus learned". High scores on both specialisation and innateness can be associated with, for example, Fodor's work in this area. Roughly, his idea is that humans are born with "innately cognized propositional contents" (Hutto, 2008: 144, citing Fodor, 1983: 85), which can be understood as specialised modules containing the basic rules of folk psychology. In the practice of social interaction, these rules can be applied to representations of someone's beliefs, desires, and other intentional states, in order to yield predictions of someone's future intentional states or behaviour, much in the fashion of steps (i), (ii), and (iii) above, as suggested by Premack and Woodruff. According to this view, mindreading capabilities are in place from birth, but they are initially "masked": infants lack the ability to exploit their innate understandings of intentions, desires, beliefs, and so on, until they improve general skills such as selecting and processing information, and applying it appropriately (for a more recent defence of a position along these lines see Leslie, Friedman, and German, 2004).

Other researchers suggest that folk-psychological theories are constructed rather than inherited genetically. Most notably, Gopnik and Wellman (e.g. 2012) argue that children use "data" gained from experience with their own and

others' actions in the social world in order to infer causal structures using forms of statistical learning. In the course of growing up, they may test their theories via "informal experimentation" through play, and further refine them through imitation and pedagogy. According to this view, forming theories about others' inner lives and social behaviours is done using the same mechanisms as forming theories about other aspects of the world (such as behaviour of physical objects). Gopnik and Wellman are thus situated at the other end of the theory-theory spectrum: according to them, mindreading is learned using more domain-general mechanisms, instead of relying on innate and specialised structure.

Simulation theory

The second dominant view on mindreading is known as "simulation theory". The essence of this view is that the inner lives of others can be *modelled* using one's own mind. Like in Premack and Woodruff's early discussion of what they call "empathy"[15], the basic idea is that one reasons "as if being in the other's shoes". However, in more recent accounts of simulation theory a distinction is being made between, on the one hand, low-level simulation of actions, bodily expressions, and basic emotions, and, on the other hand, the high-level simulation of intentional states providing the motivations and conditions behind these actions, expressions, and emotions (see e.g. Gallese, 2001; Goldman, 2006; for a broader discussion and critique see Gallagher, 2012). Roughly, the opinion among simulation theorists is that the low-level component is present in infants and other primates, whereas the high-level is unique to humans and develops throughout childhood. The low-level component is argued to rely primarily on activation of mirror neurons,

[15] There are many different usages of the term "empathy" around in the literature on social cognition. Sometimes it is used as a synonym for mindreading, sometimes it is framed as a process underlying mindreading, and sometimes it is argued that the two can do without each other (see e.g. Gallagher, 2012, for a discussion of different positions among simulation theorists). In general, I think that the meaningful categories in the domain of mindreading and social cognition are covered in this section, though sometimes in a simplified form. It is possible to discuss some of these categories in light of the term "empathy", or even to re-label some categories with terms such as "affective empathy", "emotional empathy", or "cognitive empathy" (e.g. De Waal, 2005; Uzefovsky et al., 2015), but this is rather a different way of cutting the same cake than an extension yielding genuine conceptual enrichment.

providing a very direct sense of another's body movements, facial expressions, and (through that, probably) basic emotional states (see Note 7 above; see also Knoblich and Sebanz, 2006, for what they refer to as "common coding"). When given the example of the lady with the suitcase on the stairs discussed in 1.1.1 above, simulation theorists would probably explain the "contagious" effect of this scene in terms of low-level simulation processes making one "take over" part of the burden—after all, the mirror-neuron view predicts that some of the same networks in the brain are activated when *executing* the action of lifting the heavy suitcase as when *seeing* someone else do this. Simulation theorists would probably go on to add that the higher-level component of simulation is needed to distinguish between the situation on the stairs and the one in the gym: after all, only after running a full imaginative simulation of both situations ("putting oneself in the shoes" of the two mindreadees in the pictures), can one become aware that the lifting of the heavy object has a different goal in either case and thus is underpinned by a different set of motivations.

Narrative-practice and two-systems approaches

The third and fourth views that have come to prominence in the literature on mindreading are the "narrative-practice" hypothesis (Gallagher and Hutto, 2008) and the "two-systems" approach (Apperly and Butterfill, 2009; Apperly, 2011: chapter 6 and 7). These two views can be characterised by their attitudes towards the first two positions: defenders of the narrative-practice hypothesis tend to argue that *neither* theories nor simulations can account for how we understand the inner lives of others, whereas two-systems thinkers generally grant the importance of elements of *both* theory and simulation. However, this difference is one of emphasis rather than of essence, since upon a closer look the two approaches have much in common. Both start from the view that the foundations of social interaction lie in the mutual coordination of actions and body movements, allowing for such "embodied" routines as mimicry, alignment, and imitation, and both approaches suggest that these processes are predominantly automatic, in place from early infancy, to some extent shared with other primates, and that the mirror-neuron system might play an important role. In fact, their suggestions of how the foundations of social

interaction work, are much in line with the low-level part of simulation theory. However, it is characteristic of the narrative-practice approach to emphasise that social interaction at this level has nothing to do with assessing intentional states. According to Gallagher and Hutto, the most prominent adherents of this view, the entire understanding of others around us in terms of intentional states, is an *a posteriori* dimension which we apply to the social world and its inhabitants using our experience with folk-psychological narratives. It is only because children are *told* (and adults keep telling each other) what people think, intend, desire, etcetera, under which conditions they do this, and how this is linked to behaviour, that we make sense of interaction events by referring to "underlying" intentional states (see Gallagher and Hutto, 2008; Hutto, 2008; Gallagher, 2012). Although there is debate over the precise implementation of this view in the practice of actual social interaction, an important part relies on the "matching" of previously collected exemplars with the case at hand.[16] In this sense, this view comes closest to what Premack and Woodruff referred to as "associationism", with the important difference that chimpanzees of course only have their own experiences and observations available as exemplars, whereas the crucial advantage for humans is our access to culturally accumulated experiences and observations through narratives—we can gain experience through others' eyes, as it were.[17] A related though more specific suggestion regarding the acquisition of exemplar cases through language and narratives will be made in Chapter 4 and 5 of this thesis, and be integrated in the synthesis developed in the Conclusion.

Two-systems thinkers share with advocates of the narrative-practice hypothesis the view that narratives and social schemas are highly important for our understanding of the social world and the inner lives of its inhabitants, but narratives and schemas have a different place in their model. According to two-systems theory it is pivotal to recognise that social conventions have a high degree of normativity, and that the settings of (in)formal instruction through

[16] Incidentally, the building of *abstractions* on the basis of these exemplars is here not being excluded by the authors, which definitely blurs the sharp distinction with theory-theory they make elsewhere (e.g. Gallagher and Hutto, 2007).

[17] I discuss this view in more detail in Van Duijn (2015, in Dutch). A position that has much in common with this view, but does not put as much emphasis on narratives, can be found in Heyes (2012) and Heyes and Frith (2014). See also the summary of my position in this debate the Conclusion.

which children are familiarised with these conventions are aided by the practice of telling narratives (Apperly, 2011: chapter 6; see also Warneken and Tomasello, 2006). However, where narrative-practice thinkers have a tendency to downplay the importance of mindreading altogether, replacing it by non-representative forms of "bodily" coordination on the one level, and narrative competences on the other, the two-systems approach rather uses bodily interactions, narratives, social norms, and schemas to explain how mindreading at various levels is *possible*. The basic idea here is that, in order to serve the actual practice of social interaction, mindreading has to be quick and flexible at the same time (Apperly and Butterfill, 2009). If it were the case that for every word, gesture, coordinated movement, helping hand, etcetera, a full mindreading process had to run, based on the totality of cues, represented intentional states, knowledge of folk-psychological rules or simulation of the others' position, and so on, fluent interaction would be impossible. Therefore, two-systems theory suggests that part of the burden is taken away by quick, deeply-rooted, mostly automatic, bodily interaction routines, and that knowledge of social norms, schemas, and narratives can help a great deal in constraining the amount of information that has to be processed, and selecting what is relevant in a particular context, thus making the mindreading task tractable.[18] However, all of this may come at the cost of the flexibility needed when one is confronted with mindreading tasks that go beyond bodily routines, general schemas, and so on. Therefore, according to two-systems theory, there is at least one other system available on top of the basic, "quick" system, which can deal with non-straightforward cases in a more explicit, flexible, though slower and more cognitively demanding way, possibly using elements of both theorising and simulation (see Apperly and Butterfill, 2009; Apperly, 2011: chapter 6 and 7; see also Kahneman and Tversky, 1982).

[18] In fact, Chapter 2 works out this line of thinking for stories that appear to comprise highly complex mindreading tasks: I argue that a combination of "expository strategies" makes these tasks tractable, so that the audience in principle need not more than basic mindreading skills to be able to follow the plot.

Blind men and the elephant

By way of concluding this section, I will discuss a little thought experiment adapted from a study by Kahneman and Tversky (1982; also discussed by Gallese and Goldman, 1998) leading to a brief summary and synthesis. Imagine that two travellers share a taxi on their way to the ferry port. Each of them has to take a different ship, however, both ships are scheduled to depart at the same time. During the taxi ride they are confronted with unexpectedly heavy traffic and, on top of that, when they are nearly there, the driver takes the wrong highway junction. As a consequence, they arrive at the port an hour late. Traveller A finds out that his ship has left thirty minutes ago, at the scheduled time. Traveller B is told that his ship was delayed by twenty-five minutes and left only five minutes ago. Who of the two travellers will be more upset? In Kahneman and Tversky's study, nearly all participants agreed that this would be traveller B. How did they arrive at this conclusion? A theory-theory explanation would stress the role played by abstract (innate or acquired) intuitions of how different intentional and behavioural states relevant to this scenario are causally related. Given that the desire to catch the ship is equally present in both travellers, the difference must be explained from the dissimilarity between A's belief that, given the heavy traffic, he would have missed his boat anyway, and B's belief that he would still have caught his boat if only the traffic had been just a little less chaotic, or the driver had not missed the junction. Using abstract knowledge of the rules governing folk psychology, one can reason from these represented belief states to the expected degree of "upsetness" and judge whether the answer is A or B. Simulation theorists, in contrast, would suggest that there is no need to apply abstract, folk-psychological rules, or even to represent desires and beliefs, since one can simply compare one's own degree of imagined upsetness in either situation. Defenders of the narrative-practice hypothesis would argue that one's sense of how the projected intentional states in this scenario link together can be correlated with similar scenarios one has acquired previously, leading to one outcome rather than the other. Two-systems thinkers, finally, would leave space for pragmatic combinations of these explanations.

When providing a brief overview of research into mindreading, like I have done in this section, it is unavoidable that each approach is reduced to a simplified sketch giving but an impression of the line of thinking behind it. Clearly, all these approaches have long histories (which also became clear when discussing Premack and Woodruff's 1978 paper) and build on substantial foundations of philosophical inquiry and empirical evidence. Interestingly, it seems that the more papers one reads by (self-)proclaimed defenders of each of the camps, the clearer it becomes that much of the contrasts and controversies are rooted in conceptual and terminological incompatibility, or in dissent with respect to what the relevant questions are, rather than in disagreement over the answers to given questions.[19] One might be reminded of the well-known Indian parable in which six blind men are for the first time confronted with an elephant, and report to one another what this magnificent creature must be like: the one who feels the trunk says an elephant is like a flexible tree branch, the one who feels a leg says it is like a soft pillar, the one who feels the ear says it is like a hairy pancake, and so on. The overarching concept which I have here been referring to as "mindreading" can be found throughout the field under the labels "theory of mind", "folk psychology", "(lower-order) mentalising", "cognitive empathy", "second-order intentionality", and more, all with slight differences in what exactly is meant—not even to mention how this applies to the whole range of adjacent concepts and terms, such as "social cognition", "simulation", "affective empathy", "emotional empathy", "folk-psychological narratives", etcetera. And even abstracting from terminology, there is ample variation in the phenomena and behaviours which are considered to be of interest. In my view, as with the elephant in the parable, real progress will require researchers from different backgrounds to "talk to each other" and cooperate beyond disciplinary borders.

To add just one more example to those already discussed (based on Apperly, 2011: 114-116): imagine a study in which an experimenter sits behind a table that has two boxes on it. A participants sits down at the other end of the

[19] Apperly makes a similar diagnosis, leading to his pragmatic approach of considering a wide variance of existing studies and insights on their merits before laying out his own "two systems"-model. For additional reflection on terminological controversies surrounding "theory of mind" see Schaafsma et al. (2014).

table. The information is provided that there is a piece of chocolate in one of the boxes. Next, the experimenter looks either to the right or the left box, after which the participant is asked to judge whether the experimenter *thinks* there is a piece of chocolate in the box he is looking at. In the first condition, the instructions are such that the participants have to deduce what the experimenter believes and intends in order to locate the chocolate. In the second condition, the instructions are the same, except that one piece of information is added that makes it possible to skip any reasoning about mindstates and simply use the experimenter's gaze as a cue to infer where the piece of chocolate is located. Does this mean that this study is only in the first condition about "mindreading" and in the other condition about, say, "following eye gaze", or "using behavioural cues"? This position is problematic, since in the second condition participants may use either (or both) mindreading and following eye gaze before formulating their answer. Once again this issue can be avoided by using Marr's distinction between the W- and H-levels. On the what-level, it is safe to say that the entire experiment is about mindreading: subjects are asked about where they *think* the experimenter *believes* the chocolate is located. However, on the how-level, generalisations must be made with great caution: it may well be the case that there are differences from one condition to the other, and between subjects, regarding the strategies and mechanisms used to complete the task. This is a point that applies already to Premack and Woodruff's (1978) early paper: after all, they suggest that familiar problems are more likely to be solved through "association", while novel ones require "theorising"—in other words, they already allow room for the possibility that one and the same task involving predictions of others' intentions and behaviours, may or may not require "theory of mind" depending on individual factors and context (see also Apperly, 2011: chapter 6 on this point).

The next two sections provide a further introduction to embedded mindstates or multiple-order intentionality. The focus will no longer be on the H-level of how inferential processes could be carried out when reading minds (as was the case in this section), but will shift back to the conceptual W-level of "what it is" that needs to be explained.

1.2 Embedded mindstates

1.2.1 Multiple-order intentionality

As stated in the Introduction, researchers from various disciplinary backgrounds and convictions have made a case for the importance of the ability to deal with multiple, interrelated intentional states at various levels of complexity. The common way to conceptualise this complexity is using *orders of intentionality*. Originally, the scale of orders of intentionality figured in debates on primate cognition from the 1960s and 1970s. Dennett, who was himself an important contributor to these debates, explained the scale of orders of intentionality as follows:

> A *first-order* intentional system has beliefs and desires (etc.) but no beliefs and desires *about* beliefs and desires [...]
>
> x *believes* that p
>
> x *wants* that q
>
> where "p" and "q" are clauses that themselves contain no intentional idioms. A *second-order* intentional system is more sophisticated; it has beliefs and desires (and no doubt other intentional states) about beliefs and desires (and other intentional states) – both those of others and its own. For instance
>
> x *wants* y to *believe* that x is hungry
>
> x *believes* y *expects* x to jump left
>
> x *fears* that y *will discover* that x has a food cache
>
> A *third-order* intentional system is one that is capable of such states as
>
> x *wants* y to *believe* that x *believes* he is all alone
>
> A fourth-order system might *want* you to *think* it *understood* you to *be requesting* that it leave. (Dennett, 1983 [1962]: 345)

Although this way of counting orders is not as straightforward as it may seem (see Chapter 6), Dennett's explanation does provide a good impression of the

logic of thinking underlying the scale of orders of intentionality.[20] This logic can be detailed in propositions as follows:

P_0	[It is raining outside]	0^{th}-order
P_1	Bill believes that [it is raining outside]	1^{st}-order
P_2	Mary believes that Bill believes that [it is raining outside]	2^{nd}-order
P_3	Peter believes that Mary believes that Bill believes that [it is raining outside]	3^{rd}-order
P_4	John believes that Peter believes that Mary believes that Bill believes that [it is raining outside]	4^{th}-order
P_5	Sally believes that John believes that Peter believes that Mary believes that Bill believes that [it is raining outside]	5^{th}-order
P_n	$Name_n$ believes that P_{n-1}	n^{th}-order

Table 1 – The square brackets indicate that "it is raining outside" is here seen as a fact of the world, independent of a subject having an intentional state about it. In P0 there is no such subject, in P1 there is a subject (Bill) exhibiting first-order intentionality (by having an intentional state about the fact that it is raining), in P2 there is a subject (Mary) exhibiting second-order intentionality (by having an intentional state about Bill having an intentional state about the fact that it is raining), and so on.

The logic of counting orders of intentionality in this way has inspired a vast amount of research in experimental psychology and cognitive neuroscience, ranging from the development of tests to assess individuals' performance on reasoning tasks involving varying orders of intentionality, to a focus on typical and atypical development, involved brain areas, and formal models of the

[20] Incidentally, below I will discuss a problematic side of this logic that seems to some extent prompted by the very term *"order(s)"* of intentionality. The idea of it being "orders" evokes questions such as "How many orders can a species/individual process?" or "What is the number of orders involved in this task/event/story?" I will argue that intentional states are most of the time not "piled up" (as orders), but interlinked in various kinds of ways. In this Section and in other parts dealing with or building on the long tradition of research on this topic, I will retain the term multiple-order intentionality. In Chapter 3 I will discuss my alternative concept of the "thoughtscape" in more detail.

mechanisms underlying the ability to deal with embedded mindstates.[21] In addition, linguists, literary theorists, archaeologists, anthropologists, and researchers from a handful of other fields have used the concept of embedded orders of intentionality in their frameworks (various examples will be discussed in Sections 1.2.2 and 1.3 below). In this thesis, the totality of research that has implemented the logic of the orders of intentionality in some form will generally be referred to as research within the *mentalising paradigm*, named after the tests used for the assessment of one's competence to reason with embedded intentional states, the so-called "mentalising tests".

In mentalising tests, subjects are asked to read or listen to short stories describing a particular sequence of social interactions, such as the organisation of a surprise party. The story is followed by questions of the form "Did A know that B wanted C to come to his party?", or "Did C know about the party?", or "Did B want A to think that C should know about the party?". By using three to five such stories, each followed by around ten questions of differing orders of complexity, a score indicating "mentalising capability" can be calculated for each individual participant. In a range of studies, scores from this test have been shown to be associated with various sorts of measures of people's social capabilities and real-life social functioning. For example, a number of studies have indicated that mentalising scores correlate with estimates of social network size, suggesting that those participants who perform better at mentalising tests have, on average, more people in their social networks. Another study has indicated that participants with higher mentalising scores were less likely to attribute causes of negative events to others: they appeared to be, as it were, less "distrustful" of others' intentions in a social context. Other studies have investigated the relations between mentalising, empathy, and executive functioning, or mentalising skills and language competence. Also, a version of the mentalising test adapted for children showed an association between test scores and general social aptitude as assessed by their teachers. Another perspective was added by various studies in the field of social

[21] For a discussion of research into typical and atypical development see Baron-Cohen et al. (2013); for involved brain areas see, among others, Frith and Frith (2003) and Rushworth, Mars, and Sallet (2013) and Mars et al. (2013); for formal models see, for example, Behrens et al. (2009) and Yoshida et al. (2008; 2010). See also Note 14 above.

neuroscience: higher mentalising scores have been shown to correlate with higher amounts of grey matter in cortical areas important for social functioning.[22]

All these statistical associations may be taken to indicate that mentalising tests *do* tap into at least some skills and properties relevant to actual social life and interaction. However, authors presenting mentalising research themselves, as well as critical outsiders, have stressed that it is still to a large extent unclear *why* these associations exist, or in other words: little is known about which mechanisms are targeted by these tests and how precisely they relate to real-life social interaction. In addition, discussions have arisen over ecological relevance and methodological soundness of the questionnaires, but the tests have been improved over the years and researchers have found ways to control for factors such as general memory capacity or language ability.[23] Throughout this thesis, questions pertaining to the mentalising tests will return in various forms, and a detailed analysis will be offered in Chapter 6.

1.2.2 The roles of language

Earlier in this chapter, the option of formulating explicit mindreads has been discussed (Section 1.1.1), along with the possibility to describe and categorise different mindstates and their mutual relationships using linguistic propositions (Sections 1.1.2 and 1.1.3). With this, however, only one of three roles of language in relation to mindreading has been addressed. The current section

[22] The "mentalising test" (also sometimes referred to as the "Imposing Memory Task" or "IMT") was originally designed by Kindermann, Dunbar, and Bentall (1998), for a study in which they investigated the relation between test scores and causal attribution of negative events. Afterwards, the test was revised, updated, and adapted several times. Stiller and Dunbar (2007) demonstrated a positive correlation of mentalising scores with estimates of social network size, which was replicated several times (see Lewis et al., 2011; Powell et al., 2014; Launay et al., 2015). All these studies suggest a better performance among women. For research showing associations between mentalising performance and volume of the orbital prefrontal cortex see Powell et al. (2010), Lewis et al. (2011), and Powell et al. (2014). See Launay et al. (2015) for mentalising in relation to empathy and executive functioning. For mentalising in children see Liddle and Nettle (2006) and in adolescents see Haddad (under review). An elaborate analysis of these studies is offered in Chapter 6 of this thesis.

[23] See Launay et al. (2015) for a general discussion, O'Grady et al. (2015) for a critical review of the methodology and an alternative testing method using movie clips, and Oesch (2015) for mentalising and language competence.

distinguishes these roles and points forward to the chapters in which they will be discussed in more detail.

The first role of language is thus the *representation* of mindstates and mindreading tasks. This can itself be subdivided into *formal* or *propositional representation*, where mindstates and their mutual relationships are made explicit for the purposes of investigating them and assessing their complexity (as in, for example, Table 1), and *natural representation*, the way in which mindstates and their relationships are rendered and managed in various genres of natural discourse, including novels, plays, newspaper texts, radio reports, conversations, etcetera. Language in the role of representing mindreading, both propositionally and naturally, is important throughout this entire thesis. In Chapter 2 and 3, the focus will be on (literary) narrative language, in Chapter 4 and 5 more everyday forms of language usage (newspapers, conversations) will enter the stage, and Chapter 6 will deal with linguistic representations of mindreading tasks in the context of psychological experiments.

The second role concerns the *conceptual support*, *scaffolding*, and/or *training* that language can provide for our mindreading skills, even when the actual reasoning is performed implicitly and/or non-linguistically. For example, various researchers have suggested that children around the age of 3-4, who learn to deal with embedded sentences (e.g. "Snoopy thinks that the candy is in the box"), not only acquire a way to *communicate* about mindstates and perspectives, but also learn a formula for *thinking* about them in the first place (see e.g. Lohmann and Tomasello, 2003; Milligan et al., 2007). In other words, the matrix structure of such sentences may not only provide a new "label" for an existing reasoning process, but also add a new strand of reasoning to a child's thought repertoire. In a similar way, stories can be argued to form a natural training environment for one's mindreading skills. They offer insight in the fictional minds of characters, thereby enabling one to experience "what it is like" to be inside someone else's head, and they provide a mode for projecting hypothetical social scenarios, thereby avoiding the potential costs of trying these out in real life. On top of these ways of support, scaffolding, and training

for mindreading, as identified by various authors,[24] I suggest an additional one, which pertains to the *structural* properties of narrative language usage. In short, as I will argue in Chapter 2, narrative language features all kinds of strategies for fleshing out perspectives and mental states, and for mutually coordinating them in a natural and comprehensible way. Learning to deal with narrative may therefore hone one's "real-world" capabilities of switching between multiple perspectives, understanding situations in terms of the underlying perceptions, intentions, motives, etcetera, and mapping behavioural patterns on particular mental states. Viewed this way, narrative is not just a way of speaking, but also a way of *thinking*, which is at least partly governed by the conventions of narrative language that we acquire in the context of learning to understand and tell stories (see Van Duijn, 2015). More details on this idea will be worked out in Chapter 2, 4, and 5.[25]

Not only does language thus serve to represent mindstates and mindreading tasks (first role), nor is it just likely to provide implicit support and scaffolding for our mindreading abilities (second role), it is also in important ways *itself dependent on* and *building upon* mindreading. This third role of language makes things complicated: after all, if all three are considered together, it is implied that language and mindreading must have a relationship of mutual dependency and "cosupport" in developmental terms, and one of "coevolution" in evolutionary terms—which is precisely what I will assume throughout this thesis, and argue for in various ways. Such arguing is necessarily incomplete and to some degree speculative, given the issue's enormous psychological complexity and evolutionary depth of hundreds of thousands of years. Nonetheless, I hope to provide convincing arguments and evidence at various points that it is the *only* possible way of construing the relationship between language and mindreading. With an eye on that, it is important to briefly introduce some concepts from the study of human interaction, with which I will round off this section.

[24] For a variety of views in the broader area of literature and (social) cognition see, among others, Zunshine (2006); Boyd (2009); Vermeule (2011); Oatley (2011); Nussbaum (2011); Djikic et al. (2013); Carney et al. (2014).
[25] Note that this idea is reminiscent of the "narrative practice hypothesis" (e.g. Gallagher and Hutto, 2008), but only partly overlaps with it; see also Section 1.1.4 above.

Human interaction is, broadly, the context in which language usage takes place. In the default version it happens face-to-face between a S(ignaller) and A(ddressee) who reverse roles with every turn taken, using a multimodal stream of auditory, visual, and palpable cues—all other interaction forms, such as writing, phone calls, text messaging, and so on, are ultimately variants of or derivatives from this default setting (Fillmore, 1981). Interaction is by no means always linguistic: humans can manage each others' behaviour, share information, make friends, play tricks and jokes, and interact in all kinds of other ways without ever saying or writing a word. This is known to anyone who has ever been "lost in translation", trying to get around in a place where no one speaks one's language. Or, another good example of how rich interaction can be without the aid of language is provided by the game of charades: players often manage strikingly well in getting complex meanings across, even though all conventional, mostly linguistic symbols are banned (except for a few ones specific to the game). However, any player of charades or anyone being lost in translation also realises how strained and impoverished communication without language is. To use an amended version of Scott-Phillips' (2015: 16) words: language is not what makes interaction possible, but what makes it powerful.[26] After all, as will be discussed in more detail in Chapter 5, the conventions of a language can be seen as "supercues", coagulated local solutions (i.e. within one cultural-linguistic community) to the coordination problems that arise when interacting. In this view, every lexical item and grammatical procedure ultimately is the result of generations of language users trying to coordinate their mindstates in interaction with each other and the environment, thereby converging on solutions that are communicatively effective, physically and cognitively efficient, and learnable for new generations of language users (see also Verhagen, 2015; Mesoudi, 2011; Tomasello, 2008: chapter 6).

In his analysis of the distinctive properties of human interaction, Levinson (2006) introduces the concept "Schelling mirror world". Schelling was an economist who studied a specific species of coordination problems: the ability of subjects to arrive at a solution together in the absence of

[26] The full version of Scott-Phillips' quote will be discussed in Chapter 5 (and contested on an important part not cited here).

communication. For example, if they are told that they have to meet someone else in Moscow the next day, but not exactly where and when, and they know that the other has had precisely the same instruction, they can perform much better than chance would permit by (implicitly or explicitly) asking themselves what the other will think, and what the other will think that they will think. A "Schelling point" (Schelling, 1960) high above the odds in Moscow is probably "12 noon at the Red Square, in front of the clock tower besides the Kremlin". If one has to meet in a theme park, this point would probably be the entrance, or in a crowded department store it may be the "lost-and-found" desk. Converging on such Schelling points, according to Levinson, requires not only a special way of reflexive thinking (about what the other will think one will think, etcetera), but also a notion of mutual knowledge or *common ground*, including a sense of mutual salience: "what leaps out of the common ground as a solution likely to independently catch our joint attention" (2006: 49, referring also to Clark et al., 1983, and Clark, 1996). He argues that these same ingredients are also requirements for human communication: reflexive thinking and common ground, including a mutual sense of salience. After all, as has been described by many linguists and philosophers of language, there are thousands of possible ways in which a particular meaning can be expressed, while at the same time, every expression can have many different meanings.[27] Only through the same combination of reflexive thinking and common ground, including a sense of mutual salience, can humans coordinate their mindstates while interacting, or in Levinson's words: it is through these factors that "meetings of the mind" can occur in the "Schelling mirror world" that underlies human interaction (2006: 49; for an experimental approach see Stolk, Verhagen, and Toni, 2016; Stolk, 2014).

Grice (1957) was the first to present a fundamental study of how communicative meanings can arise despite the indeterminacy of linguistic

[27] This can be demonstrated using nearly any utterance, but consider the example of me saying to a friend: "hey, there is Ann". If we are standing outside a music venue, and Ann has our tickets, this probably means something to the effect of "all right, we can go inside". However, if Ann is my friend's ex-girlfriend and we are about to enter a bar for a drink, it can mean "let's go somewhere else"—unless my friend has just told me that he hasn't seen his ex-girlfriend in a while and would be interested in a conversation with her, in which case it probably means "what a coincidence, let's go inside"... etcetera. For a discussion see, among many others, Keller (1995), Sperber and Wilson (1995), Clark (1996), and Scott-Phillips (2015).

expressions as such. According to his theory of meaning, "a signaller S communicates z by behaviour B if S intends to cause an [addressee A] to think z, just by getting [A] to recognise that intention" (Levinson, 2006: 49; "recipient" in original replaced by "addressee"). Sperber (1994; 2000) and Scott-Phillips (2015) have reformulated this insight in terms of a multiple-order mindreading problem, suggesting that for any full-blown linguistic interaction event:

S intends
that A should recognise
that S intends
that A should believe
that z

The precise nature of this mindreading problem, assumed to be at the heart of language usage, will be detailed (and contested) in Chapter 5. The version at which I will eventually arrive, building on Clark's (1996) and Verhagen's (2015) notions of *common ground* and *joint intentionality*, suits the lazy mindreader by being much more economical in terms of the assumed amount of cognitive complexity. In short, it turns the argument upside down: instead of suggesting (following Sperber and Scott-Phillips) that interaction works because interlocutors (somehow, implicitly) take the steps spelled out above in order to "meet" each other at five orders of embedded intentionality, I argue that *as a rule* they start off having already met—and instead of suggesting that it is necessary by default, I suggest that it is only in exceptional cases that such steps need to be taken (for example, as will be discussed in Chapter 5, in order to work out and repair a misunderstanding: "Ah! I thought you intended me to think that….etc.").

Put differently: *in theory* it is possible for interlocutors to reflect on the communicative situation in the way suggested by Sperber and Scott-Phillips, but in practice it is rarely necessary. Normally, a signaller "tosses" a particular behaviour (typically a string of sounds, gestures, and facial expressions) into the Schelling mirror world, assuming that the addressee will be able to figure out what the signaller means by it. In nearly all instances of communicative interaction there are several principles and mechanisms at work that save the

signaller and addressee from having to apply multiple-order mindreading. Summarised in brief:

- Common ground/joint intentionality: interlocutors always start from a set of shared beliefs or "common ground" (Clark, 1996) instead of having "join" their individual sets of intentional states each time they interact;
- Ready-mades/packages: for many expressions, occurring in particular contexts, we may have existing meaning associations stored in our memory that are either shared between speaker and addressee in particular, or among members of the cultural-linguistic community more widely. Such associations can be easily retrieved, compared, adjusted, and used as ready-made blueprints or frames in interaction, without having to establish complex meanings "from zero" (as worked out for examples such as "allegedly" and "accidentally" in Chapter 4).
- Interactive structure/alignment: in interaction we do not have to sort everything out by default and right away—in every communicative turn we seem to build representations that are "good enough" for the interaction to keep going, but no better (cf. Apperly, 2011: 114-119). If required, interlocutors can work out a particular point in more detail, aiding and steering each other in the desired direction turn by turn. Many conversations do not have "signal-response" as their basic structure, but rather "testing-adjusting-retesting" (Levinson, 2006).
- Relevance: driven by the need for communicative efficiency, signaller and addressee have both learned from their experience as communicators to become geared towards choosing maximally relevant solutions. This means that, in most cases, what the signaller has to do is pick the first expression that comes to mind, while the addressee has to pick the first interpretation that comes to mind. If this does not work, they can try the second-most relevant expression or interpretation; thus both speaker and addressee in practice work downwards on the gradient of relevance (cf. Apperly, 2011: 115-116, referring to Sperber and Wilson, 2002; and Chapter 5).
- Ratchet effect of linguistic items: not only are signaller and addressee experienced in choosing the most relevant cues and interpretations, the

> linguistic tools they have available also store a wealth of such accumulated "experience". After all, they have emerged as a result of numerous instances where generations of signallers have tried to get particular meanings across to addressees, in settings that have for at least some important parts not changed (cf. Chapter 4 and 5).

As I will argue at several places throughout this thesis, in many cases of daily interaction these mechanisms work so well, that mindreading in the "full" form as suggested above is hardly ever needed in order to communicate—it is only in exceptional cases, such as when trying to repair a misunderstanding, playing a sophisticated pun, or reflecting on the very act of communication, that participants in a communicative setting are incited to go "all the way down" and work out what the other intends that they understand that the other wants...etcetera. In other words: language is what makes human interaction so powerful not just because it can represent mindstates and their relations in efficient ways (first role), nor just because it may support mindreading implicitly (second role), but also because it can work as a "mindreading-avoidance tool". Mindreading is indeed necessary for communicative interaction (third role), however, various mechanisms and principles that are part of, mediated by, or closely tied to language save interlocutors the trouble of having to process all steps suggested by Grice, Sperber, and Scott-Phillips by default.

1.3 The social brain

1.3.1 Early primate roots

Up to now this chapter has been concerned with the fundaments of what intentional states are like, how mindreading can work, and how both relate to language. However, as stated in the first section, there is another basic question: why is it that we bother about mindstates of others at all? The context in which an answer to this question can be provided (and, indeed, the context in which this question itself becomes relevant) is offered by research surrounding the

social brain hypothesis. At the core of this hypothesis lies the idea that the complex social environments in which primates have lived in their evolutionary past were the primary drivers behind the emergence of their increasingly large and powerful brains—or rather, *our* large and powerful brains, since humans are of course included in the primate order.[28]

The briefest version of the story of primate evolution goes as follows. In the geological period known as the Palaeogene or Lower Tertiary, a bit over 50 million years ago, certain mammals on the African continent started foraging in groups (see Shultz, Opie, and Atkinson, 2011, for key evidence supporting this scenario). This development was probably driven by a transition from nocturnal to diurnal activity, which increased the risk of being attacked by predators while moving around in search of food. Although living in groups lowered vulnerability to predation, at the same time it posed some very particular challenges for the ancestral primates, including finding new ways of organising reproduction and care for offspring, resolution of conflicts arising over access to resources within groups, avoidance of the costs inflicted by freeriders, and coordination involved in moving collectively or protecting the group against risks from outside. In response to these challenges, various species have evolved different solutions over millions of years of time, as reflected in the different forms of social organisation and complexity that can be found throughout the primate world today (for references and more detailed overviews of the early episodes in primate history see Dunbar, 2014: chapter 2 and 4; Gamble, Gowlett, and Dunbar, 2014: chapter 3 and 4).

An important characteristic of social organisation found throughout the entire primate world is the tendency to form intense social bonds and coalitions

[28] The suggestion of the association between brain size and social complexity was originally made by Jolly (1966) and Humphrey (1976), before it was addressed in Byrne's & Whiten's volume *The Machiavellian Intelligence Hypothesis* from 1978. Dunbar further developed the social brain hypothesis and was the first to test it systematically (1992; 1998), discovering the correlation between social group size typically formed by a primate species and its neocortex size, and subsequently presenting numerous findings supporting and/or refining the hypothesis. He has written a vast amount of publications on the topic, several of which play important roles throughout this thesis. For this introductory section, I will to a large extent follow the line of his recent overview book *Human Evolution* (2014), along with the co-authored volume *Thinking Big. How the Evolution of Social Life Shaped the Human Mind* (Gamble, Gowlett, and Dunbar, 2014), only at key points referring to the original papers. In the rest of this thesis, the original papers will be used.

that last over longer periods of time. These bonds and coalitions directly or indirectly protect individuals against the challenges and costs of living in groups and defuse the stress that comes with it. Primate bonds have an emotional and a more cognitive component. The first is primarily mediated through endorphins triggered by specific social activities. The latter can be defined in terms of having a mutual sense of trust and obligation, and, in some cases, willingness to provide help and support, all of which require some form of cognitive coordination (cf. Dunbar, 2014: chapter 2). I will return to the cognitive demands of social living below, but first focus briefly on the emotional, endorphin-mediated component.

1.3.2 The "bonding gap"

The main activity associated with endorphin release in non-human primates is social grooming, the process where one individual sifts through another's fur to remove small bits of debris and inert skin. Besides hygiene benefits, this mildly painful treat triggers the release of endorphins in the brain, which alleviate stress and pain levels and presumably underpin the feelings of emotional closeness that we know from friendship and love. This mechanism is still at work in humans, as can be sensed when receiving a massage or engaging in light stroking and cuddling. However, in our case, time has seen the addition of other mechanisms of maintaining intense social relationships, which have taken over much of the heavy lifting (Dunbar, 2014: chapter 1 and 8).

According to Dunbar the transition from social grooming to other bonding activities constitutes one of the main threads in the evolutionary story of our lineage. Simply put, when group sizes increased, our ancestors must have run up against time limits: since grooming is an inherently time-consuming, one-on-one activity, it works fine for smaller groups, but will put high pressures on time budgets in larger groups. A partial solution found in some species of primates is to invest in a strong relationship with a few core social partners (instead of weaker ties with many or all group members), thus breaking up the larger group into interlinked and partly overlapping coalitions. This structure was most likely found in the groups formed by our hominid ancestors, and

arguably is still visible in present-day human social life.[29] However, it was estimated that if ancestral human societies had relied on strictly one-on-one bonding activities such as grooming, their members would have needed to spend over forty per cent of their day doing this, which would have conflicted severely with the time budgets reserved for foraging and resting (Dunbar, 2014: chapter 7; Gamble, Gowlett, and Dunbar, 2014: chapter 5). In other words, it seems likely that the amount of free time left for social activities *after* foraging and resting, put a constraint on the maximum number of social ties individuals could maintain. Therefore, in order to be able to break through various glass ceilings of maximum group sizes, more time-efficient bonding activities had to emerge in our lineage, bridging the "bonding gap" (Dunbar, 2008) between the groups of 40-60 individuals, in which our early ancestors lived, and the groups of around 150, as formed by anatomically modern humans—and this is precisely what Dunbar and colleagues have argued: activities involving for instance laughter, dance, music, and, of course, language have become our alternatives for social grooming.

Both dance and laughter have the capacity of fairly straightforwardly triggering endorphin release, thus supporting emotional social bonding in a direct way. A similar case can be made for singing together. Also, importantly, dancing, singing, and laughing can be done together with several others at the same time, greatly increasing the effectiveness of time spent socially.[30] The same, of course, holds true for talking. However, the links between language and social bonding are more complex. Talking and listening *as such* do not seem to be contributing much to social bonding: rather, language contributes indirectly through such activities as gossiping and sharing jokes, myths,

[29] For this social group structure, referred to as the "fission/fusion-model" see Dunbar (2003).
[30] For an overview see Dunbar (2014: chapter 2, 3, and 8). For social bonding in relation to dance and moving "in synchrony" more widely see Tarr, Launay, and Dunbar (2014); for laughter and social bonding see Dunbar et al. (2011); for singing see Pearce et al. (forthcoming). Research shows that laughing, even today, is typically done in intimate cliques of two to four individuals, rather than in larger groupings. Interestingly, these cliques are similar in size to the groups people tend to form in natural conversations (see Dunbar, 2015).

religious stories, and fiction.[31] However, with the possible exception of telling (some sorts of) jokes, these activities all require a highly sophisticated form of language to be in place, capable of representing at least some abstract concepts, referring to events outside the here-and-now, and coordinating multiple referents and possibly their mindstates (cf. also Tomasello, 2008: chapter 6). The emergence of such sophisticated language forms is typically assumed to be of a relatively recent date in our evolutionary history (possibly only with the arrival of *Homo sapiens* around 200,000 years ago or even later; see e.g. Fitch 2010; Perreault and Mathew, 2012), whereas our social group sizes, and thus our need for efficient social bonding mechanisms, have shown important increases much earlier, going back probably around 2 million years (Gamble, Gowlett, and Dunbar, 2014: chapter 5). This would suggest that other factors (such as dance, music, and laughter) were at first more important in bridging the bonding gap, while various forms of gossiping and storytelling came in later. In addition and related to this, as discussed in Section 1.2.2 above, it is argued by some researchers that mastering a language "as we know it" requires powerful mindreading capacities, which is also considered to be a reason for why the emergence of such sophisticated language must be dated to our more recent history.

The first of two main ways in which social group size can be linked to cognition, and to mindreading in particular, can thus be summarised as follows: bonding larger groups may require language and storytelling skills, which rely on mindreading capacities, which again rely on large and powerful brains. However, one of the objectives of this thesis is to rethink the way in which mindreading and language are related, so this topic will be continued (and the argument partly challenged) at various places, especially in Chapter 5 and in the Conclusion.

[31] Besides contributing to endorphin-mediated bonding indirectly, language clearly has other important advantages in the context of keeping increasingly large communities together: talking can "time share" on other activities (such as walking, eating, and cooking) and it can be used to share information about the social network (gossip) in a much more efficient way than by personal observation (see Dunbar, 2014: 227).

1.3.3 Cognition and primate social life

Besides the emergence of new bonding activities, a second main thread in the story of human evolution identified by Dunbar is brain size. The social brain hypothesis projects that if an animal's social life is more complex (and thus demands more sophisticated social behaviour), it will have more grey matter in the brain areas associated with social cognition. Support for this relationship has been found throughout the entire animal kingdom, both at the level of one species compared to another species and at the level of individuals within the same species. Dunbar (2014) discusses research showing that it holds even in social insects: species with more complicated social structures show increased "brain" volume (or relevant neural network size), compared to species of social insects with less complex social structures. In a similar vein, queen bees have a significantly more sophisticated social life than their worker sisters, and also show more relevant brain volume. Species of birds forming pairbonds, hence needing to be able to maintain intense, long-term relationships with their partners, have larger brains compared to birds who have more flexible mating systems. Primates with more diverse repertoires of social behaviour, for example involving deception or alarm calls, tend to have relatively larger neocortices. The same is true for primates living in larger social groups, where they have to maintain higher numbers of social relationships and/or exhibit more diverse repertoires of social behaviour. This is reflected in the correlation between mean group size and neocortex size as plotted in Figure 3:[32]

[32] Note that the social brain hypothesis is thus essentially about social complexity, and not about the number of relationships an individual can deal with per se. "Qualitative" factors are equally important: for example, mating strategies, deception rates, or coalition complexity all correlate with relative brain size irrespective of total group size (Shultz and Dunbar, 2014: 49-50; cf. also Dunbar, 2008).

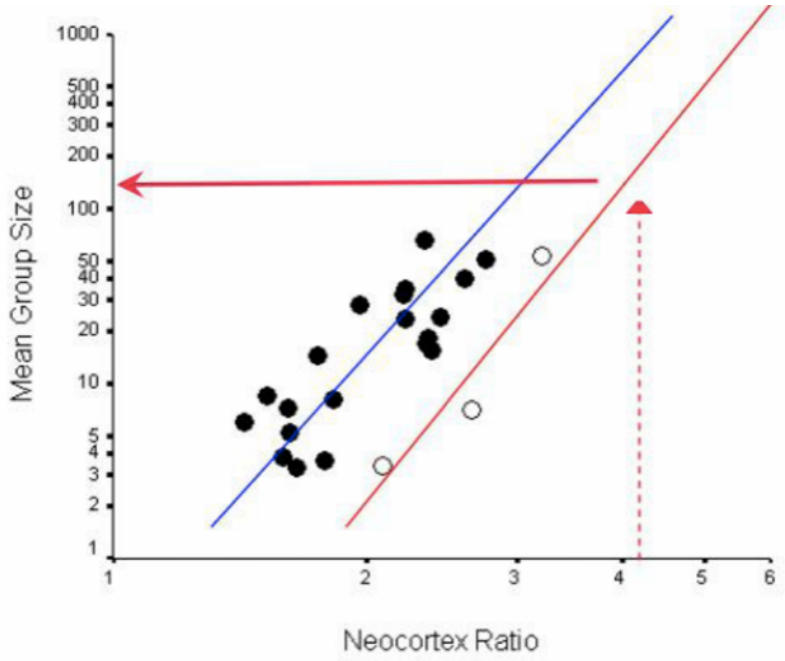


Figure 3 – The filled dots represent species of prosimians and monkeys, showing a robust correlation between mean group size and neocortex ratio, the size of the neocortex divided by the total brain size. The open dots represent the three species of great apes, from left to right orangutans, gorillas, and chimpanzees. The red arrows indicate the predicted mean group size for modern humans, based on their neocortex ratio factored into the ape equation: ±150 individuals, known as "Dunbar's number" (Dunbar, 2003).

Group sizes in our primate ancestors were not only constrained by the availability of social time for grooming and/or alternative bonding activities (which may or may not have involved mindreading, as discussed in the previous section), but also by cognitive limits in a more direct way. This is the second of the two main ways in which cognition, and mindreading in particular, can be linked to social group size. In order for a group not to break down into chaos, individuals need to *coordinate* their behaviour with respect to one another: from the very basic capacity to move in accordance with others' movements, a skill apparent in for example bird flocks or ungulate herds, to sophisticated empathic, cooperative and strategic behaviours, as in for example targeted helping, conflict resolution, group hunting, deception, or consolation. Mindreading competence at various levels of complexity has been argued to

underpin different aspects of such coordination of individuals within groups, each of which will be discussed in the subsections below.

1.3.4 Mindreading, coordination, and group size

In the case of birds flying in a flock, coordination of behaviour from an individual's perspective comes down to adjusting to a few others that happen to fly near. The apparently sophisticated movements emerging on the level of the flock thus do not require birds to bother about anything beyond the movements of a handful random neighbours: there is no need to distinguish between them individually, and relationships with them do not have to persist beyond the coordination task itself (David-Barrett, 2014).

Figure 4 – Individual birds in a flock only coordinate with a handful of neighbours, but do not need to have a representation of the behaviour of the flock as a whole (David-Barrett, 2014).

As a consequence, little cognitive investment has to be made for successful group membership and the cognitive costs do not add up when group size increases. This stands in stark contrast to the situation in primates, where individuals in principle maintain a constant relationship with every other individual in their social group. When interactions of some form occur, this existing relationship functions as the basis, instead of the interaction being the

basis for an ad hoc relationship (as in flocks and herds). Therefore, primates have to be (and, indeed, *are*, Parr and De Waal, 1999) able to distinguish between individual group members and, to a certain extent, they have to keep track of previous interactions they have had with them. This is an important reason why primate group life causes cognitive load: when groups grow larger, their members have to tell apart more individuals and remember the current status of their relationship with each of them.

Moreover, there is an even more weighty reason why primate sociality is cognitively demanding. In order to fit into a primate group, it is not sufficient to know one's own relationship to all other group members: one has to keep track of "third-party relationships" between all of the other group members as well. For this reason, the number of relationships primates have to be able to distinguish and keep track of, in theory, can be shown to increase exponentially with every extra group member added (David-Barrett and Dunbar, 2013). Evidence from observations of social structures in many different primate species, in relation to their relative brain sizes, supports this idea (Shultz and Dunbar, 2014).

However, all of this does not yet warrant the importance of the capability to deal with some form of mindreading: a primate individual can in principle monitor social relationships within its group without having to deal with intentional states. Mindreading seems to come in as soon as *prediction* of others' behaviour and intentions enters the stage—yet in this case it is a possibility, not a necessity. For example, when making decisions about whether or not to act cooperatively towards another individual, one will have to predict whether the other will reciprocate this approach, or exploit it. The scenario of deciding whether to cooperate or not has therefore been linked to the ability to deal with multiple orders of intentionality. Yoshida et al. claim that such a decision is necessarily based on "recursive representations of another's intentions, since if I decide to [cooperate], I must believe that you believe that I will cooperate with you" (2010: 10744). There seems to be ground here for an arms race inflicting a constant pressure on individuals to stay ahead of their peers. Advantage can be gained if one can from time to time make the other believe that one intends to cooperate, while in fact one is about to exploit him. At the same time, it is important to be able to detect when the other intends one to believe that he will

cooperate while in fact he will not… (Byrne and Whiten, 1988, have emphasised the importance of such "Machiavellian" cheating and cheater detection in the evolution of primate social intelligence; see also Tomasello, 2014: 37-38).[33]

Note that I wrote purposefully that it *can* be the case that mindreading is needed when prediction of others' behaviour and intentions enters the stage, but that it is not a necessity. A point that often seems to be overlooked in studies that link mindreading to cooperation is that calculating whether the other believes that one intends to cooperate is one way of predicting the other's reliability as a cooperation partner, but surely not the only way. Another option would be to make this prediction on the basis of past behaviour of the other, either towards oneself or towards others. All that is needed is the ability to tell individuals apart and a memory for previous interactions, but no mindreading (see Wilkinson, 1984, for an example in vampire bats). A third way of predicting another's intention to cooperate would be using emotional or situational cues: is he nervous, are there more group members around to watch his behaviour, etcetera. A fourth way would be to make an assessment of the "rationality" of the task: what can the other gain from cooperating, and what does he have to invest? In fact, it would make most sense to use combinations of these ways (and potentially even additional strategies), and it may well be that this is what individuals do in practice. In short, deciding whether or not to cooperate with others may involve mindreading, but not necessarily so.[34]

---

[33] Another way in which cooperation and mindreading can be linked is by factoring in third-party opinions: the question for X whether or not to cooperate with Y can also depend on predictions of what Z would think of this alliance. This adds complexity in terms of the number of intentional states involved in making a decision for X, without these intentional states necessarily being *embedded* (Y thinks… and Z thinks…). The issue of interlinked (but not embedded) intentional states will also return in various forms throughout the following chapters.

[34] In some way this comes down to saying, more generally, that not all mindreading tasks are solved through mindreading; or more precisely: some tasks generally considered to involve mindreading (such as decisions about cooperation) can be solved *both* in ways generally considered to be mindreading (e.g. placing oneself in the other's shoes, reasoning about the other's possible motivations) and in ways generally not considered to be mindreading (e.g. extrapolating from the other's previous behaviour). Note that this is the same point as made at the end of Section 1.1.4., and that this again highlights the importance of distinguishing between different levels of explanation. Sometimes it can be said on the W(hat)-level that "X decided to cooperate because he *thought* that Y *intended* to cooperate as well", whereas on the H(ow)-level this decision came down to (for example) mere extrapolation of X's experience that Y always cooperated on previous occasions. Is this a case of mindreading? Yes on the W-level, no on the H-level. (Cf. also Kümmerli et al. 2010).

Apart from the issue of deciding whether or not to cooperate, does cooperation itself involve mindreading? For example, chimpanzees seem to hunt cooperatively (e.g. Boesch, 2005), which could be argued to require A to *understand* that B *wants* to move around the tree, so that B *intends* A to *understand* he should take the other side, and so on, implying mindreading at multiple levels of complexity. However, it has been pointed out that a more likely scenario is that all chimpanzees participating in a hunt try to maximise their own chances of catching the prey, which results in a situation that only *seems* to be coordinated intentionally from the perspective of an outside observer (similar hunting is found in hyenas, lions, and wolves; see Tomasello, 2008: 173-175; Dunbar, 2014: 244). Such hunting "alone together" does not seem to require much mindreading, apart from again the possible Machavellian twists of misleading others or anticipating potential misleading by others. In that sense, this form of cooperation may on the level of its underlying mechanisms well come closer to the bird flock than it seems at face value.

1.3.5 Mindreading and social learning

It has further been claimed that mindreading underpins living in social groups through facilitating effective learning mechanisms. Whereas some theorists have suggested that most of the important learning takes place through imitation, and therefore does not rely on taking others' perspectives (see Heyes, 1993; 2012a), Tomasello and others have suggested that it is precisely because we, humans, are able to picture ourselves in someone else's shoes, that we can learn "through" them (Tomasello 1999; 2008; 2014). In this way, thanks to our mindreading competences, cultural conventions can reliably spread through a group at a fast pace, since learning not only takes place from parents and caretakers to a new generation ("vertically"), but also "horizontally" between peers:

> The form of social learning required here is not just imitation, but role reversal imitation, in which each initiate to the convention understands that she can use the convention toward others as they have used it toward her, and vice versa—so that both producer and comprehender

> roles are implicitly present in both production and comprehension.
> (Tomasello, 2008: 221-222, referring to Tomasello, 1999)

Groups of humans (and potentially some of our hominid ancestors and ape relatives)[35] having such a mechanism in place through which conventions spread, turn into cultural communities where coordination between group members works in a highly effective way: when conventions are mutually shared, there is (as it were) a supra-individual order capable of *orchestrating* behaviours and interactions in all kinds of domains. This saves huge amounts of negotiation and trial-and-error costs—time, risk, energy, cognitive power, and so on—otherwise borne by individual group members. This will be elaborated further in Chapter 5, where the notions of "joint intentionality" and "common ground" are introduced (following Clark, 1996, and Verhagen, 2015) and where linguistic items will be viewed as coagulated solutions to coordination problems occurring when interlocutors try to update a set of shared beliefs.

1.3.6 Mindreading, language, and narrative

Apart from (but clearly related to) cultural learning, the ability to deal with multiple orders of intentionality has been argued to enable and support language ("third role" in terms of Section 1.2.2 above), and thereby activities important for living in social groups, such as gossiping and storytelling, as discussed above. Dunbar (2014) and others consider the latter activities highly important factors in how our hominid ancestors could break through glass ceilings of group size and brain capacity. The final section of this chapter will be concerned with Dunbar's view on the role of language and stories in the context of the social brain hypothesis. This brings the discussion back to the

[35] There is evidence that chimpanzee groups also have some form of cultural conventions that spread both horizontally and vertically. However, compared to the human situation, there clearly is an enormous difference in the amount to which these conventions modify and enhance the chimpanzee ways of living. See Whiten et al. (1999) and Whiten (2011).

core issue of this thesis: the relation between mindreading, language, and narrative.[36]

As stated earlier, the basic idea advocated by Dunbar is that our lineage, over time, exhibited increasingly better mindreading competences. In brief, it is assumed that our current capacity comprises five "levels" of intentionality (Kinderman et al, 1998; Stiller and Dunbar, 2007) and that the last ancestor we shared with our closest relatives in nature, chimpanzees and bonobos, could, like them, achieve at most two of such levels. Smaller-brained monkeys are assumed to be capable of only one level. Combined with the mentioned neuroimaging experiments suggesting that, in human subjects, mindreading competence is correlated with brain mass in areas relevant to social cognition, a function can be hypothesised expressing brain size in terms of achievable level of intentionality (Powell et al., 2011; Dunbar, 2014: chapter 7). When brain sizes of our ancestral hominids, estimated on the basis of fossil skull bones, are factored into this function, this yields the following graph:

[36] This thesis will not explicitly address religion, but it is clear that religious traditions rely for an important part on the exchange of stories. Therefore, much of what will be said in this thesis about stories in relation to mindreading is also relevant to building and maintaining religious communities. For a discussion of religion in relation the orders of intentionality, see Gamble (2010) and Dunbar (2008; 2014: chapter 8).

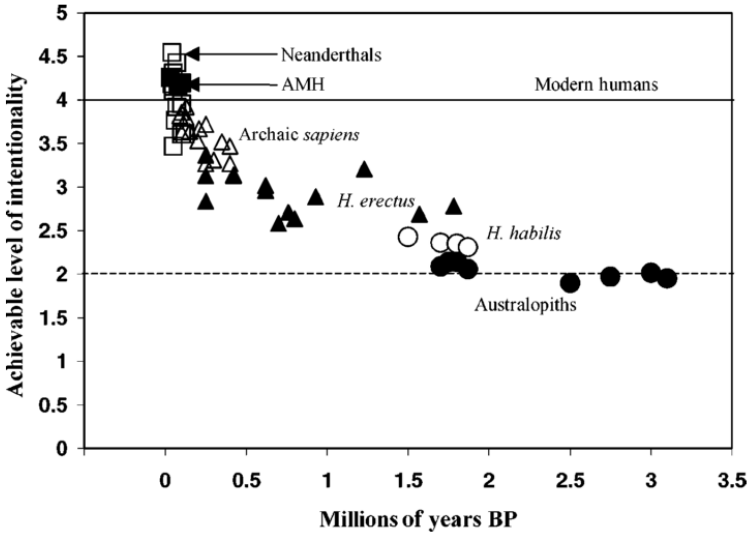


Figure 5 – AMH stands for "Anatomically Modern Humans"; BP for "Before Present" (source: Dunbar, 2003). Note that more recent insights suggest that Neanderthals could achieve four orders rather than five as indicated here (see Dunbar, 2014: 241-244).

Dunbar's fundamental assumption, then, is that brain size was the factor limiting how many levels of intentionality could be processed. Given that the archaeological record shows an increase in cranial volume in our lineage over the past 3-4 million years, this leads to the claim that our ancestors were able to handle increasingly more levels of intentionality across this period. Next, Dunbar argues that the maximum achievable level of intentionality put a limit on the sophistication of the language that could be developed at any particular stage, and thus the complexity of the activities that this language could support. For example, imagine that an early form of language existed in *Homo erectus* that was useful for coordination purposes in the here-and-now, but not sophisticated enough to support social bonding through gossip or telling stories. In that case, our ancestors at that time would not have been able to use this language for bonding (much) larger communities, as is assumed to have been the case in later stages (see Section 1.3.2 above). In this way, according to Dunbar's model, brain size limits achievable level of intentionality, which then limits sophistication of the language that can be developed, which in turn limits

the number of individuals that can be bonded in a coherent social group. Or phrased differently: increases in brain size over time "released" additional mindreading capabilities, which enabled more sophisticated forms of language, which in turn allowed for larger communities to be maintained in a coherent way.

As discussed in relation to the "third role" of language (Section 1.2.2 above), some researchers claim that it involves the capability to work at a handful of levels of intentionality to entertain human language "as we now know it", even when producing the most basic utterances (e.g. Sperber, 2000; Scott-Phillips, 2015). Combined with Dunbar's model set out above, this position necessarily entails that such language arrived late in our evolutionary history. After all, according to this model, the required ability to operate at such higher orders of intentionality was only available in anatomically modern humans. Earlier hominids may have had language, but this must then be assumed to have been of a lower degree of sophistication, given their limit at second- or third-order intentionality (for a discussion see Dunbar, 2014: chapter 7).

As said, in the chapters that follow I will first develop a perspective of economy, not so much looking at the limits of our mindreading capacity, but rather focussing on the *minimal* amount of mindreading needed for using language and dealing with stories. Chapter 6 then addresses the implications of this perspective for the practice of assessing mindreading experimentally. Finally, in the Conclusion I will return to the bigger picture set out in this first chapter and sketch the contours of how it should be updated in the light of the points developed throughout this thesis: after all, if the relationship between mindreading, language, and narrative is construed differently, this has potential consequences for the chronology of events assumed in the story of human evolution, and for the way these events are causally related.