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

Duijn, M.J. van

Chapter 6

The Mentalising Test Revisited

In several parts of this thesis I have referred to "mentalising tests", versions of the instrument used by experimental psychologists for assessing people's ability to reason about intentional states at various levels of complexity. In the Introduction and in Chapter 1, various key findings of research using these tests were listed, such as the correlations between individuals' mentalising scores and the sizes of their social networks, or the amount of grey matter in particular brain areas. In Chapter 2 a story used in a version of the test was cited (about Emma trying to get an increase in wages with her job at the greengrocer) and an example of a question was discussed (Jenny *wanted* Emma *to believe* that her boss *thought*…etcetera), after which the distinction was made between *situations* involving multiple orders of intentionality and their linguistic *representation* in the form of a proposition or narrative. Chapters 2, 3, and 4 discussed representations of multiple-order intentionality across plays, novels, journalistic discourse, and spoken language. In Chapter 5, I have argued that, when interacting, reasoning about orders of intentionality enters the stage especially when we need to determine how individual perspectives *differ* from shared knowledge or common ground. The present chapter first discusses key issues of mentalising tests in general, in the light of insights from the previous chapters, followed by a detailed investigation of a selection of stories and questions from three mentalising studies.

6.1 Five central conclusions of the mentalising-test paradigm

As discussed in the Introduction and in Chapter 2, in mentalising tests "complexity" has always been conceptualised as the number of embedded intentional states featured in each question, following Dennett's scale of the orders of intentionality (see Chapter 1, Section 1.2.1). As an illustration, consider the following two true/false-statements[86] from the version of the mentalising test used by Brown (more details of her study follow in 6.2 below):

(1) Sam wanted to buy a stamp
(2) Henry knew Sam believed he knew where the Post Office was

After a short story was read out twice, participants had to answer "true" or "false" for twenty of such statements. Ten were "intentionality questions" and concerned intentional states of characters featured in the story; the other ten were "memory questions" and concerned factual details, for example:

(3) The Post Office was closed because it had moved to Bold St

In the case of intentionality questions, the level of complexity for each question is determined by counting the number of embedded intentional states, whereas in the case of the memory questions, complexity corresponds with the number of factual details included in the statement. In this way, statement (1) has complexity level 2, since the participant has to work at second-order intentionality following Dennett's scale: the intentional system (the participant) has to $know_1$ whether or not Sam $wanted_2$ to buy a stamp. Statement (2) has complexity level 4, since the participant has to $know_1$ whether or not Henry $knew_2$ Sam $believed_3$ he $knew_4$ where the Post Office was. In a similar way, statement (3) is a memory question at complexity level 2, since it has two factual

[86] Some of the studies used true/false statements and others (such as the original study by Kinderman et al. (1998) or the recent one by O'Grady et al. (2015) discussed below) presented two alternative statements from which participants had to choose the right option.

elements which have to be checked against the story: *the Post Office was closed*$_1$ because *it had moved to Bold St*$_2$.

The original finding by Kinderman, Dunbar, and Bentall (1998), who used a test comprising five such stories and sets of questions, is that error rates go up steeply at complexity level 6 (when counting as explained above)[87] in intentionality questions, but not in memory questions. The following graph presents this result:

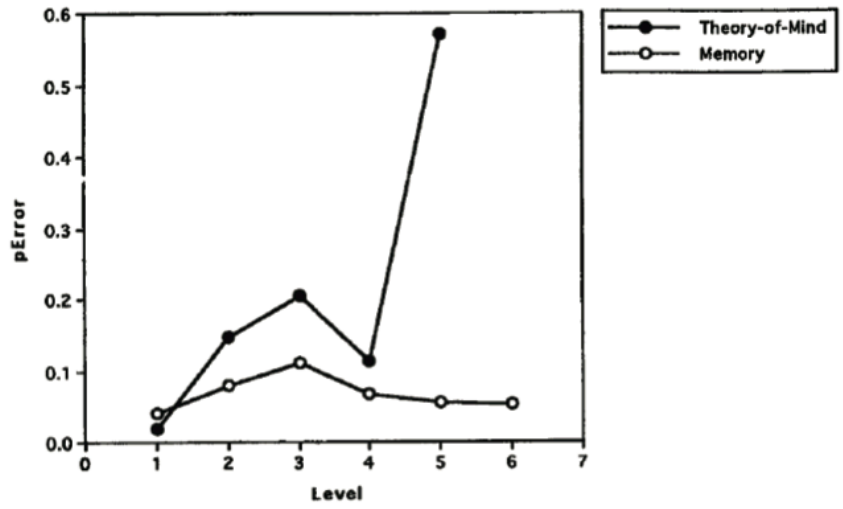

Figure 1 – Graph from the original paper by Kinderman et al. (1998). The proportion of incorrect answers (error rate) is indicated on the y-axis; the level of complexity on the x-axis. Mentalising questions are labeled "Theory-of-Mind"; questions about factual details "Memory"; n=77 participants.

The *asymmetry* between error rates at higher levels in intentionality versus memory questions, along with indications of significant variance in performance *between* participants, as found in this study, were interpreted as

[87] It is important to note that in this paper the number of orders was counted starting from level 0, which means that their level 1 would be referred to as level 2 in terms of the counting method used in this thesis and in by far the majority of cases throughout the literature. In the counting of Kinderman et al. (1998), error rates thus went up at level 5 instead of level 6. The rationale behind the common way of counting is given in Chapter 1, Section 1.2.1: in for example (1) above, the participant has to work at second-order intentionality, since s/he has an intentional state about Sam having an intentional state.

the first evidence for three of the total of five central conclusions of the mentalising-test paradigm that I will distinguish:[88]

conclusion (i) In mentalising questions, mean error rates remain constant or increase to a limited degree with each level of complexity added, until a steep increase takes place at level 6 (i.e. level 5 counting according to Kinderman et al.), suggesting the presence of a natural "limit" at that level.

conclusion (ii) There is significant between-subject variance in error rates at the different levels of complexity, suggesting that some individuals have their "limit" at level 4, some at level 5, some at level 6, and some at an even higher level of mentalising complexity.

conclusion (iii) Performance on mentalising and memory questions is related (participants making more mistakes in memory questions, also tend to make more in intentionality questions), but the results cannot be explained in terms of memory alone: beyond remembering multiple, mutually related details from the story (which is necessary for answering both types of questions), reasoning about embedded intentional states adds an extra challenge, as reflected in the differences in error rates at especially the higher levels of complexity.

Another issue investigated in the paper by Kinderman et al. concerns correlations between scores on the mentalising questionnaire and scores on another test, known as the "Internal, Personal and Situational Attributions

[88] I will refer to research using versions of the discussed method as studies within the "mentalising-test paradigm". The original title of the questionnaire was the "Imposed Memory Task" or "IMT", and it is sometimes referred to using this title across the literature.

Questionnaire" or "IPSAQ". This latter test was designed to measure participants' tendency to attribute negative and positive experiences in social settings to either themselves, others, or situational circumstances. A (strongly) increased tendency to see others as responsible for negative social events is associated with psychopathological disorders (most notably, paranoia). In a healthy subject population, there is nonetheless individual variation in how such attributions are made. Kinderman et al. showed that this variation in part correlates negatively with mentalising performance: those individuals who exhibited less ability to reason about intentional states had a higher tendency to attribute negative social events to others (or, put bluntly, "were somewhat more paranoid"). This was interpreted as the first evidence for what can be identified as the fourth central conclusion of the mentalising-test paradigm:

conclusion (iv) Mentalising performance is relevant to people's actual social lives: an individual's mentalising score tends to be reflected in indicators of this person's social life and general aptitude in the social domain.

The patterns found by Kinderman et al. have been replicated in an array of subsequent studies among different participant populations. These studies used improved versions of the mentalising questionnaire and introduced new measures against which the mentalising scores were tested. In this way, they have yielded additional evidence for the conclusions (i) – (iv), and added a fifth one situated in the neuroscientific realm (see below). The most important additions include findings pertaining to several domains of social and cognitive functioning, further supporting conclusions (ii) and (iv). First, performance at higher levels of mentalising complexity was found to be lost in patients suffering from bipolar disorders (Kerr, Dunbar, and Bentall, 2003). Secondly, estimates of people's personal social network size turned out to be associated with their mentalising scores. Stiller and Dunbar (2007) found that people who exhibited higher mentalising performance, on average tended to have a larger "support clique" (defined as the innermost circle of social contacts, from which

one receives emotional support).[89] Thirdly, studies using a version of the mentalising test adapted for children indicated that higher scores were associated with higher social competence as independently rated by their teachers (Liddle and Nettle, 2006). Moreover, associations were found between mentalising performance and personality traits (Nettle and Liddle, 2008; Van Duijn et al., 2014), schoolgrades (Van Duijn et al., 2014), empathy, and aspects of executive functioning (Launay et al., 2015).

The fifth and final central conclusion of the mentalising-test paradigm can be formulated as follows:

conclusion (v) Mentalising performance is related to brain size, in particular to the amount of grey matter in the orbital prefrontal cortex in humans.

This conclusion forms an important background to the idea, discussed earlier at the end of Chapter 1, that our hominid ancestors gradually, over many generations, evolved an increasing capacity for mentalising, going from a limit at around complexity level 2 in our last common ancestor with chimpanzees and bonobos, through a limit at level 3 in homo erectus, one at level 4 in archaic humans, and eventually a limit at level 5 or 6 in anatomically modern humans (see Chapter 1, Section 1.3.6 above). This idea rests on the pillars of two findings. Firstly, monkeys with smaller brains seem capable of only first-order intentional attributions, whereas monkeys and apes with relatively bigger brains can (under certain circumstances) handle second-order intentionality (see e.g. Call and Tomasello, 2008). Secondly, as expressed in (v), there is evidence in humans that the subjects in test populations who perform better at mentalising tests, have more brain volume in the orbital prefrontal cortex, an area associated with various aspects of social functioning (Powell et al., 2010; Lewis et al., 2011). More precisely, the claim is that the amount of grey matter in the orbital prefrontal cortex is positively correlated with social network size,

[89] Even more interesting than the correlation with mean support clique size might actually be the possibility that mentalising competence imposes a *limit* on an individual's maximum possible support clique size—Stiller and Dunbar find some support for this suggestion in their data (see 2007: 98-100). The correlations of mentalising scores with estimates of mean social network size were replicated by Lewis et al. (2011), Powell et al. (2012), and Powell et al. (2014).

and that this relation is *mediated* by mentalising competence (Powell et al., 2012; Powell et al., 2014). Note that this comes down to a within-species version of the social brain hypothesis: bigger and more powerful brains allow for the management of larger and more complex social networks not only between different species of primates, but also between human individuals.[90]

All in all, the current status quaestionis is that individuals having relatively more brain volume in particular areas relevant to social functioning (through genetic or developmental factors, or both), possess some social competences that can be measured using the mentalising test. As a consequence of these social competences they can maintain a larger and more complex social network. What those social competences measurable with intentionality tests entail *precisely* is unclear, but they correlate with particular psychometric measures (causal attribution of negative and positive social events) and indicators of traits such as short-term memory performance, personality, empathy, and, potentially, executive functioning. I will get back to this interpretation below and in the Conclusion.

6.2 Testing mentalising competence

For non-human species, a broad range of behavioural tests have been developed to investigate their abilities and limitations in reasoning about intentional states. Examples include (to name just a few) interpreting informative cues (e.g. Premack and Woodruff, 1978; Tomasello, 2008), hiding and tracking food in competition with others (e.g Hare et al., 2006, for chimpanzees; Clayton and Emery, 2004, for corvids), or working together to gain access to food (De Waal, 2005; Yamamoto, Humle, and Tanaka, 2012). Such

[90] The same trend has been found in macaques, although evidence there seemed to indicate that the increase in grey matter in relevant areas was the *result* of living in a larger social network rather than a precondition (see Sallet et al., 2011). This interpretation seems also feasible in humans: living in more complex social environments (such as a larger personal social network) could increase social capability, which is then reflected in the amount of grey matter one has in the orbital prefrontal cortex and one's scores on a mentalising test. Note that these scenarios (more complex social life>more grey matter; more grey matter>more complex social life) are not mutually exclusive: "boot-strapping" or co-evolution is a likely possibility.

tests have broadly been combined with observations in the wild and other forms of "anecdotic" evidence (in the non-pejorative sense; see e.g. De Waal, 2005; Boesch, 2005).

For normally developed human adults, as well as for infants and several "special" subject populations (such as individuals suffering from psychopathological disorders), an even greater variety of tests attempting to provide insight into intentional reasoning abilities have been used (see Apperly 2011 for an overview). Roughly, these tests can be classified as follows:

	explicit/reflective	implicit/behavioural
1st- and 2nd-order	A	B
3rd-order and above	C	D

Table 1

Categories A and C include tests that are mediated by descriptive language of a kind one would not normally find in interaction. Such language usage is very similar to what is in Chapters 1 and 2 referred to as "propositional" or "formal". For example, the proposition "Mary intends that John believes that it is raining outside" is quite different in nature from the actual (linguistic and/or non-linguistic) behaviour one can expect Mary to exhibit in the described situation, which would probably consist of saying "Hey, it's raining", "Be prepared to get wet", or just handing John an umbrella (see also Section 6.2.3 below on this point). In a similar way, classic false-belief tasks (such as the original version of the Sally-Ann test, see Apperly, 2011: chapter 2) ask of participants to *reflect* on a situation using explicit descriptions of mindstates and behaviour, rather than participating in such a situation oneself, which makes them fall under A in the above matrix. For the same reason, classic mentalising tests (following Kinderman et al., 1998) fall under C, since participants have to read a story describing a social situation and then are asked to reflect on this situation by judging statements as true or false.

By contrast, implicit/behavioural tests, indicated by B and D in the above matrix, ask of participants to take part in a (controlled) social situation. Examples of such tests include those in which children have to provide a

particular kind of help to an experimenter who tries to solve a task (after which conclusions can be drawn about their ability to judge the experimenter's goals; see Tomasello, 2008) or a "Schelling game" in which participants have to ask themselves what the other would do, what the other will think they would do, and so on. Note that in practice, A/C and B/D are extremes on a scale rather than clear-cut categories. For example, Baron-Cohen's (2001) Reading the Mind in the Eyes-test could be located somewhere in the middle, as it does use descriptive language to label the moods that participants are supposed to pick up from the pictures they are shown (more A), but also involves the "implicit" act of reading the eyes (more B).[91]

In what follows, the focus will be on three studies that can be located on the imaginary scale ranging from C to D. One is a "classic" mentalising test, close to the Kinderman et al. (1998) original. The other two broadly maintain the format, but replace stories and/or questions by dialogues or movie clips, marking a move from C (somewhat) towards the direction of D in the matrix above. I will refer to each study using the surname of its primary investigator. The studies can be outlined as follows:

Brown — This is a classic mentalising study (C in the matrix above) that was performed using a questionnaire adapted from the original by Kinderman et al. (1998). It was run by Rachel Brown as a pilot for subsequent neuroimaging studies (leading to the publication of Lewis et al., 2011) and remained unpublished. It featured 25 participants (18 female; age 21-76) to whom the stories were read out twice, after which the same was done with each question (they could not see the story or the questions in written form so they had to go by their ears). Each story was followed by around 20 questions, typically 10 intentionality and 10

[91] Note that this scale is also to some degree meaningful in the case of tests for non-human animals. Even though such tests are of course non-linguistic, they can still be more reflective or more behavioural: for example, Premack and Woodruff's (1978) chimpanzee Sarah had to choose which tool use could solve a task using video recordings (which is, relatively, more A), whereas chimpanzee's in an experiment by Yamamoto et al. (2012) had to actually hand the right tool to a peer themselves (which is B).

memory ones in random order. Participants had to answer "true" of "false" using two buttons after each question.

Haddad

This study, run by Anneke Haddad, had the same procedure as Brown's, but stories were replaced by dialogues that were recorded using different voices for each character (thus arguably moving slightly on the scale from C to D, given that the dialogues mimic real-life interaction more closely than descriptions of social interaction do). However, the questions were still presented in the classic format. There were two participant groups: adolescents and adults. Slightly different versions of the dialogues were used to attune to each participant group: "colleague" was replaced by "class mate" for adolescents, and so on. The results are currently under review; I thank Anneke Haddad for permitting me to use her study here.

O'Grady

This study, run by Cathleen O'Grady, was recently published in *Evolution and Human Behaviour* (O'Grady, Kliesch, Smith, and Scott-Phillips, 2015) and comprised four conditions: (i) stories acted out and presented using movie clips, followed by a series of pairs of alternative continuations of the story also in acted movie clips, one of which was "true" and the other "false" (called the "implicit-implicit" condition by the authors); (ii) acted movie clips followed by pairs of alternative propositions read out by one "actor" facing the camera (implicit-explicit); (iii) stories read out by one actor facing the camera, followed by pairs of alternative continuations of the story in acted movie clips (explicit-implicit); and (iv) stories read out followed by propositions read out (explicit-explicit). Participants were allowed to watch each item as often as they wanted, but could not go back once

> they had gone to the next item. The conditions can be placed on the scale from C to D in reverse order: (iv) is closest to C, (iii) and (ii) are in between, and (i) is relatively closest to D. However, none of the conditions is "D proper", given that even the implicit-implicit condition, featuring movie clips mimicking real-life interaction to some degree, yields a reflective rather than a behavioural test. Participants were 66 students (41 female) who declared that they did not know the actors in any of the movie clips.

Below I will discuss selected examples from these studies; more questions and stories can be found in the Online Appendix (see note 95). Given that Haddad's and O'Grady's studies are derived from the type of study represented here by Brown, I will discuss this latter one in detail and use it as a benchmark against which the other two can be compared.

6.2.1 Narratives and propositions

Brown's study is a good representative of what could be called the "classic" mentalising study. It used a questionnaire close to the Kinderman et al. (1998) original, which was also used (abstracting from some various smaller revisions) in many other studies over the years, for example: Stiller and Dunbar (2007), Powell et al. (2010), Lyons, Caldwell, and Shultz (2010), Lewis et al. (2011), Powell et al. (2012), Powell et al. (2014), and Launay et al. (2015). I will first discuss some issues I consider to be general for al these studies, using the first story and question set of Brown's questionnaire.

(4) WHERE'S THE POST OFFICE?

> Sam wanted to find a Post Office so he could buy a Tax Disc for his car. He was already late buying one, as his Tax Disc had run out the week before. Because traffic wardens regularly patrolled the street where he lived, he

was worried about being caught with his car untaxed. As Sam was new to the area, he asked his colleague Henry if he could tell him where to get one. Henry told him that he thought there was a Post Office in Elm Street. When Sam got to Elm Street, he found it was closed. A notice on the door said that the Post Office had moved to new premises in Bold Street. So Sam went to Bold Street. But by the time he got there, the Post Office had already closed. Sam wondered if Henry, who was the office prankster, had deliberately sent him on a wild goose chase. When he got back to the office, he asked another colleague, Pete, whether he thought it likely that Henry had deliberately misled him. Pete thought that, since Sam had been anxious about the Tax Disc, it was unlikely that Henry would have deliberately tried to get him into trouble.

After hearing the story twice, participants were presented with the questions listed below (answer and level are added here; during the test questions were presented in mixed order). They received the instruction: "do not guess, if you think you cannot answer the question on the basis of (what you remember from) the story, choose 'false'".

	Intentionality questions	**Answer**	**Level**
1	Sam wanted to buy a stamp	F	2
2	Henry wanted to play a trick	F	2
3	Henry thought that Sam knew he was a prankster	F	3
4	Pete suspected that Henry was playing a prank on Sam	F	3
5	Henry knew Sam believed he knew where the Post Office was	T	4
6	Sam thought Henry knew he wanted a Tax Disk	F	4
7	Sam thought that Henry knew the Post Office was in Bold Street and hence that Henry must have intended to mislead Sam	T	5
8	Sam believed that Pete thought the Post Office was in Elm Street and hence that Pete must not have intended to mislead Sam	T	5
9	Pete wanted Sam to know that Henry believed that the Post Office	T	6

	was on Elm Street and hence did not intend to mislead him		
10	Pete wanted Sam to know that he believed that Henry had intended not to mislead him	T	6
	Memory questions		
11	Sam needed a Tax Disc from the office	F	1
12	The Post Office was closed and Sam's insurance had run out	F	2
13	The Post Office was closed because it had moved to Bold St	T	2
14	The Post Office in Elm St. had a notice on the door	T	2
15	Sam left Bold Street, then went to the office and spoke to Pete	T	3
16	Sam found the Post Office closed and couldn't buy a tax disk for Pete	F	3
17	Pete, the man who worked at the same place as Henry, and who knew that Henry was the office prankster, was Sam's cousin	F	4
18	Sam asked Henry, and did not ask Pete or the traffic wardens, about where the Post Office was in order to buy a Tax Disk	T	4
19	Sam who worked with Pete and Henry did not know where to buy a Tax Disk because he was new to the area	T	4
20	Henry, the man that Sam spoke to about where to buy a Tax Disk after he realized he needed to buy one soon, was a colleague of Pete's	T	4

Table 2

Looking at this test, several observations can be made. First of all, some of the formulations of the questions are quite puzzling: it could well be that the participants did understand the story correctly, even remembered the relevant details about who-knew-what, who-wanted-what (etc.), but then got lost when dealing with the questions. As discussed in Chapter 2, propositions of the form used for the higher-order questions (such as questions 7, 8, 9, and 10) do not exist in the "wild", so language users cannot rely on experience when assessing

them.[92] The stories communicate the same information in a somewhat more natural way: in fact, they offer another demonstration that in natural communication "narrative takes over" when more than three perspectives are involved. This leads to the first general issue:

issue (i) Classic mentalising tests use *narratives* to present a social situation, but use *propositions* to present the questions. Especially above complexity level 3, such propositions are a very unnatural way of representing intentional states in discourse. With the analysis from Chapter 2 in mind, I suggest that the unnatural presentation of the questions can be a factor limiting performance on especially higher levels of complexity.

On top of this, the propositions used in the questions vary in complexity, measured in terms of the number of *embedded* mindstates they present. However, in the stories mindstates may not only be embedded into one another, but also connected in different ways. As an example, consider the following two sentences from the story cited (4) above:

(5) Sam wondered if Henry, who was the office prankster, had deliberately sent him on a wild goose chase. When he got back to the office, he asked another colleague, Pete, whether he thought it likely that Henry had deliberately misled him.

Looking at this passage in the way used to analyse the narrative texts in Chapter 3 and the news-paper excerpts in Chapter 4, one must conclude that it

[92] An indication that such sentences are very infrequent or even non-existent at least in spoken discourse is that the Corpus of Spoken Dutch (CGN; 10 million words) features not a single sentence with four or more embedded intentional clauses (see also Chapter 2). Note that from this indication that these sentences are idiosyncratic in the context of everyday interaction it does not automatically follow that the test cannot be used to assess some aspects relevant to such interaction—quite generally, tests do of course not always have to mimic their target subject as closely as possible to be valid. However, when interpreting test results it is crucial to know in what respects the test differs from "real life", and in what follows I suggest that, in the case of studies investigating mentalising using the test described in this section, this has not always been in clear focus.

coordinates a thoughtscape of mutually interlinked, but not necessarily embedded, intentional states. First of all, there is a narrator providing insight in the perspective of Sam. Using a form of indirect thought ("wondered if…"), the reader (or, in the experimental setup, hearer) is informed that Sam wants to know whether or not Henry had deliberately sent him to the wrong location, in order to play a prank. Within the scope of Sam's thoughts, there are thus two alternative mindreads of Henry's intentions: either Henry intended to provide the right location but did not know that the post office had moved, or he did know that the post office had moved, but intended to provide the wrong location because he thought that this would be funny. Eventually, Sam checks with Pete what he thinks Henry had intended. Readers end up with the knowledge that Sam still has two alternative mindreads of Henry to choose from, plus Pete's opinion on which one is the most likely.

Put more compactly, the second issue is:

> issue (ii) Narrative language usage features all kinds of cues that prompt and mutually coordinate intentional states of characters. As the analysis in Chapter 3 showed, a thoughtscape emerging in this way is easily misrepresented by propositions featuring only embedded mindstates. This suggests a *structural discrepancy* between the nature of the complexity presented in the stories and in the questions.

Related to this, an observation that can be made repeatedly in especially higher-order questions in Brown's study is that the chains of embeddings are "broken". Consider question 7 as an example:

> 7. Sam thought that Henry knew the Post Office was in Bold Street **and hence** that Henry must have intended to mislead Sam[93]

[93] Given that the questions from Brown's study are already numbered in Table 2, when citing them I will not continue my regular numbering between brackets: (1), (2), and so on.

The method of schematic depiction of "narrative spaces" introduced in Chapter 3 is once again a useful tool here. The following configuration of narrative spaces can be drawn on the basis of question 7:

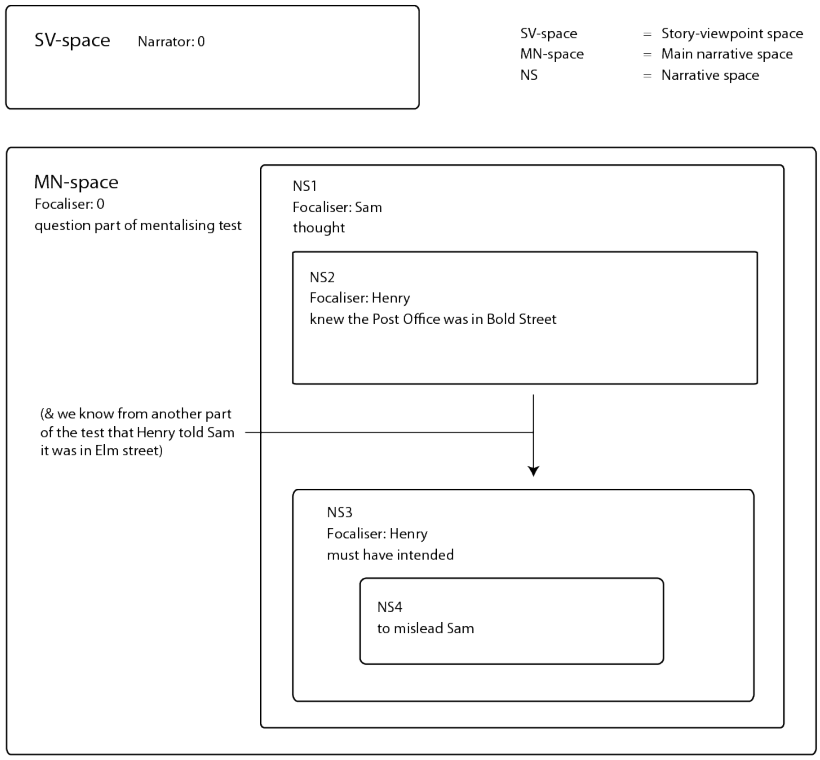

Figure 6 – Schematic depiction of the narrative-spaces configuration prompted by question 7. NS1 and NS2, together with the background knowledge gained earlier when reading the story, work as premises for the conclusion drawn in NS3 and NS4. Sam is the main focaliser in NS1 and all spaces within this space (NS2, NS3, and NS4); Henry is an embedded focaliser in NS2 and NS4. Note that NS4 contains a viewpoint package ("mislead") that could be unpacked into further spaces (see Dancygier, 2012, and Chapter 3 above for more details about narrative spaces).

Question 7 is not composed of a continuous string of four embeddings, but instead of a proposition exhibiting two embeddings ("Sam thought Henry knew the Post Office was in Bold Street"), coupled to another proposition ("Henry must have intended to mislead Sam") using a combination of connectives

marking causality ("and hence").[94] This is different in question 10, in which the chain is unbroken:

> 10. Pete wanted Sam to know that he [Pete, MvD] believed that Henry had intended not to mislead him [Sam, MvD]

Here the following schematic depiction can be drawn:

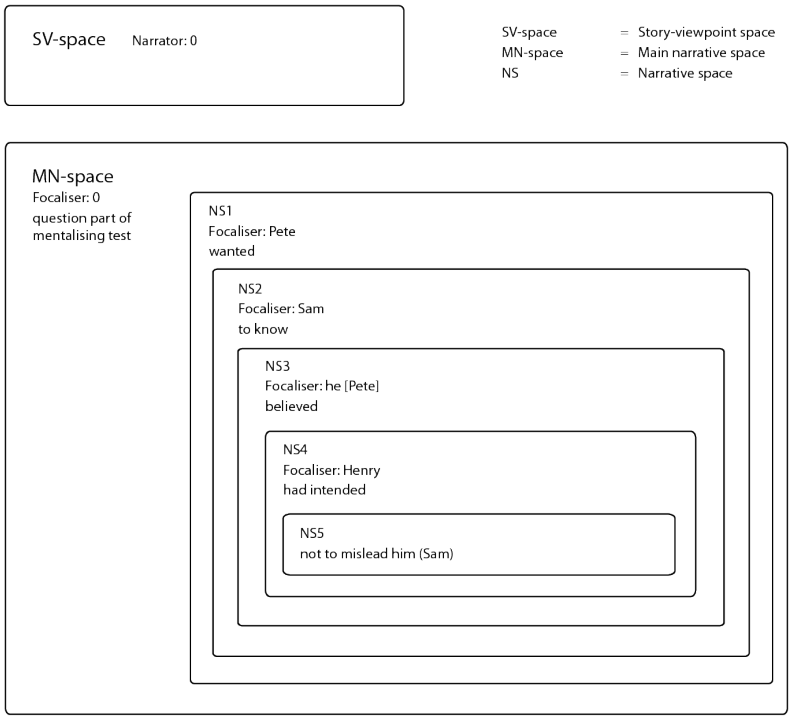

Figure 7 – Schematic depiction of the narrative-spaces configuration prompted by question 10. As in the case of Zunshine's paraphrase of *Mrs Dalloway* discussed in Chapter 3, each narrative space is embedded into the former one.

[94] Note that alternative interpretations of how the question should be read can lead to slightly different narrative-spaces configurations. The one drawn here would in fact be expressed more naturally by the sentence: "Sam thought that Henry knew the Post Office was in Bold Street and hence that Henry must have intended to mislead *him*"). Usage of "Sam" instead of "*him*" seems to suggest that the second proposition (NS3 and NS4) is not what Sam thought (so not part of NS1), but instead added by an external observer/narrator (hence best drawn as part of the MNS in Figure 2). However, this interpretation is again countered by the word "must" (suggesting an inference on Sam's part), meaning that the interpretation drawn here seems to fit best after all.

A similar broken-chain structure as in question 7 can be found in the questions 8 and 9 cited in Table 2 above, as well as in a handful other questions throughout the questionnaire.[95]

The paper presenting the results of O'Grady's study (see O'Grady et al., 2015) begins with a critical discussion of the questionnaire used in most previous studies, of which Brown's study discussed here also uses a version. The authors object to the use of broken-chain questions, stating that instead of testing a "single metarepresentational unit" of higher-order complexity, these test the conjunction of multiple lower-order mindreading tasks that can be processed as separate chunks. This means that participants can possibly use a short-cut to determine the right answer: in some cases they can check the separate parts against the "reality" of the story without ever having to considering the statement as a whole (more about this below). O'Grady et al. argue that including broken-chain questions therefore boils down to a methodological flaw. I agree with their critical stance in as far as *comparing* questions with and without broken chains is concerned: it may be methodologically tricky to assume that a, say, fifth-order question with a broken chain exhibits the same complexity as one without a broken chain. However, regarding the general *validity* of both types of questions I come to a different conclusion. Whereas O'Grady et al. suggest to avoid broken-chain questions because they do not test "real" higher-order mindreading, my analysis suggests they may in fact be better at testing how well participants understood the relationships between intentional states contained in the story. As pointed out in issue (ii) above, the story presents the intentional states not as being just embedded, but instead prompts a thoughtscape of intentional states that are interconnected in all kinds of ways. From that perspective, a question such as 7 above exhibits a structure that is more "realistic", compared to the narrative presentation of events in the story, than does question 10. Broken-chain questions, as it were, "burst" out of the artificial straitjacket of embedding-only propositions above a certain level of complexity and adopt a structure that leaves more space for expressing the viewpoint complexity

[95] The broken-chain/narrative structure can be found in questions 29, 30, 31, 48, 50, 67, 68, 69, 70, and 85. See the Online Appendix for the full questionnaire at http://liacs.leidenuniv.nl/~duijnmjvan/TLM/Appendix, password "thelazymindreader".

contained in the story in a nuanced way—once again narrative seems to be taking over.

6.2.2 Packages and inferences drawn from the common ground

Some of the viewpoint layers are fully spelled out in the story, i.e. prompted *compositionally*, to use the term introduced in Chapter 4. For example, "Sam wondered if Henry [...] had deliberately sent him on a wild goose chase" features the compositional construction of Sam's viewpoint using the verb "wondered" and the complement "if Henry [...] had deliberately sent him on a wild goose chase". However, there are additional layers contained in the story that are coordinated *holistically*: following the conditions defined in Chapter 4, the word "mislead" is a viewpoint package adding extra viewpoint layers from which the related content can be seen. The same goes for "prank(ster)" and arguably for "deliberately" and "sending someone on a wild goose chase". The entire possible mindread of *Henry knowing where the post office is, but deliberately telling Sam the wrong location because he thinks this is funny*, is nowhere spelled out but added holistically by the combination of these cues. It is possible for the reader or hearer to unpack (or *decompress*; cf. Chapter 4, Section 4.3.6) this mindread into single constituent propositions (as I just did in italics), but this is not necessary for following the story as such: as argued in Chapter 4, readers or hearers can take a viewpoint package on board holistically, and integrate only its relevant implications in their understanding of the situation.

The third and fourth issue I want to point out pertain to these dynamics of unpacking (or not unpacking) the situations and events presented by the story into single viewpoint layers. Consider again the passage cited in (5) above, this time in relation to question 2:

> 2. Henry wanted to play a trick

There is no direct evidence that Henry wanted to play a trick, but also no conclusive evidence to the contrary. Or more precisely: it is not written explicitly in the story that Henry wanted to play a trick (i.e. there is no compositional construction of this viewpoint), nor are there any other (holistic)

cues provided from which the inference can be drawn that Henry wanted to play a trick. All we know is that *Pete thought* that Henry did not want to play a trick, given Sam's anxiety about the tax disc. Although Pete's view sounds reasonable (and leads to the right answer according to the test: "false"), strictly speaking participants cannot know the answer to this question. They are best advised following the instruction (cited before Table 2 above) to choose "false" whenever the text provides no conclusive evidence. Out of Brown's 25 participants 23 indeed chose "false", thus answering correctly according to the test.[96]

A similar analysis can be made in the case of question 3:

3. Henry thought that Sam knew he was a prankster

There is no clear answer provided to this question either. In the story, there is a narrator inserting the comment "Henry, who was the office prankster" when reporting Sam's thoughts (see excerpt (5) above), suggesting that this is shared knowledge or common ground for everyone working at their office. In this sense, in the "reality" of the story, it is quite likely that Henry *does* know that Sam knows that he (Henry) is a prankster—after all, this is what he can derive from this being common ground. Question 3 thus presents a thought that Henry theoretically *could* have had, but, realistically speaking, only *will* have had if a context occurred in which it was relevant for him to derive this information from the common ground. Such contexts would be, for example: him *understanding* that Sam is anxious about the tax disc, and therefore *anticipating* that Sam might not trust him straight away, given that he is known as the office prankster; or: the context of Sam coming back from his failed attempt to buy a tax disc and confronting Henry himself (instead of Pete) with the suspicion that he had deliberately misled him, after which they could talk about the *misunderstanding* that occurred because Henry *had not anticipated* that Sam knew that he was the office prankster. However, the story features no evidence for any of these contexts.

[96] More information on the error rates can be found in the Online Appendix, see note 95.

All in all, question 3 presents viewpoint layers that Henry *could have unpacked* if there had been a context requiring this. Given that there is no conclusive evidence of the presence of such a context, the answer should be "false" following the same line of reasoning as with question 2. This is indeed what 24 out of Brown's 25 participants chose, thus answering correctly according to the test.

So far so good—now consider question 5:

> 5. Henry knew Sam believed he knew where the Post Office was

From the story it can be concluded that Sam did believe that Henry knew where the Post Office was, but there is no conclusive evidence whether Henry did or did not know that Sam believed this. The answer can again be "true" in theory, given that Henry had a conversation with Sam in which he told him where the Post Office was, and under normal circumstances Henry should be able to draw the inference that Sam believes what he told him. However, again there is no evidence of a context in which Henry would indeed have drawn this inference, so following the same line of reasoning as with questions 2 and 3 the answer should be "false". Yet according to the test the correct answer is "true", which is what 17 out of the 25 participants went for. The "failure" of the other 8 participants to provide the "correct" answer can be due to their inability to handle the fourth-order-intentionality proposition featured in question 5, but also to the fact that they followed the instruction to choose "false" whenever the story provides no conclusive evidence for the existence of particular intentional states, either by constructing them explicitly/compositionally or by providing a context from which they can be inferred.

Two further issues can be formulated with this analysis in mind:

issue (iii) Following the same line of reasoning leads sometimes to a correct and sometimes to an incorrect answer, suggesting that factors other than the amount of orders of intentionality involved in a question (which is of course the target variable of the test) can quite easily interfere with error rates produced by participants.

It is clear that the tests can be improved by checking the questions one by one for such inconsistencies (and some authors have done so, as examples below will show).[97] In addition, it may be advisable to add a third answer option apart from "true" and "false", which participants are supposed to tick when a question cannot be answered on the basis of evidence from the text. This avoids at least some of the ambiguity between cases where participants have failed to process the intentional reasoning task and those where they have failed to apply the instructions correctly.

A more important and fundamental issue, however, has to do with the unpacking of viewpoint layers where this is not relevant in the context of the story. I have above discussed this for the questions 3 and 5, and question 10 contains another example. From the story we know that Pete and Sam had a conversation about what happened to Sam, and that "Pete thought that [...] it was unlikely that Henry would have deliberately tried to get him into trouble". Question 10 asks whether "Pete wanted Sam to know that he believed that Henry had intended not to mislead him" (the answer is indeed "true"). The core issue questioned here is really whether Pete believed that Henry intended to mislead Sam, which is a fourth-order problem. That Pete wanted Sam to know this, follows logically from the fact that they have a conversation. In other words, the first two orders of the question in fact "unpack" what is naturally given in the discourse situation. In Chapter 3, I have argued that Zunshine unnecessarily starts to count from the author, suggesting that Woolf *intends* her readers to *believe* that Richard Dalloway *thinks*, and so on. However, just as these first two layers normally do not have to be processed by readers of a Woolf novel, I suggest that we are normally not concerned with processing that a speaker *intends* the hearer to *understand* that he *thinks*, and so on. In Chapter 5 I have argued, following Clark (1996; 2006a), that such viewpoint layers belong to the (infinitely large) category of inferences that can be drawn from the common ground. When Sam and Pete have a conversation about what Henry

[97] Also, it must be noted that the outcomes of the tests are averages produced by multiple questions (usually between five and nine) at each level of complexity and by mostly quite substantial samples of participants, which means that such inconsistencies are at least to some extent balanced out as part of regular "error variance" for a test like this.

wanted, it can be inferred that Pete wants Sam to know what he (Pete) thinks about what Henry wanted, just as it can be inferred that Pete wants Sam to know that he (Pete) wants Sam to understand what he thinks about what Henry wanted, and so on. Normally, such inferences are not drawn, since they state obvious truths in a complicated way, without adding anything to what both interlocutors consider to be common ground. However, as said above with respect to questions 3 and 5, it is possible to think of exceptional contexts in which drawing such inferences can be useful. Most notably, this is the case whenever it turns out that interlocutors do not understand each other or are, as it were, not "on the same page". For example, imagine the following conversation between Pete and Sam:

(6) Pete: Henry may be a prankster, but above all he is an empathic person.
Sam: What do you mean?
Pete: I want you to understand that I think Henry did not want to deliberately mislead you, given your anxiety about being too late.

Here, Pete first tries to share his thoughts in an indirect way, expecting that Sam will draw his conclusions on the basis of the information that Henry is an empathic person. However, when Sam makes clear that he does not know what to do with that information in this context, Pete "unpacks" and makes explicit what he wanted Sam to understand.

There is no evidence in the story for a situation in which Pete and Sam are not on the same page, which means that it is unlikely that any of them needs to bother about unpacking the discourse situation into separate viewpoint layers. Of course this does not mean that it is impossible for participants to do this when answering the questions. However, when looking at the questions I think it is important to make a distinction between viewpoint layers that are in some way relevant to the characters in the story and the development of the story's plot, and viewpoint layers that are "generated extra" by unpacking layers that would normally be obvious and/or unnecessary.

All in all, the fourth issue can be formulated as follows:

issue (iv) Participants are asked to reflect, *in the same way*, on:
- viewpoint layers that are relevant to the characters and the development of the plot (whether or not these are spelled out compositionally or cued holistically); and
- viewpoint layers that can *in principle* be inferred from the story, but do not have such relevance.

Potentially, including obvious though irrelevant viewpoint layers in the questions leads to confusion about whether the answer should be "true" or neither-true-nor-false, and thus "false" (see analysis of questions 2, 3, and 5 above). In addition, it may also introduce pseudo-complexity: in question 10 a fourth-order problem is preceded by two obvious layers, which in fact come "for free" with the information that Pete and Sam are having a conversation; it is irrelevant for the interlocutors to reflect on this, neither is it relevant for the development of the plot. It is unclear how a question like 10 compares to a question staging six viewpoint layers that do have such relevance.

6.2.3 Judging facts and intentional states

Another point is that judging whether a factual statement is true or false is *conceptually* a different task from judging whether an intentional statement is true or false. Being so-called control questions, the factual questions clearly have to be different—however, the problem may be that they are too different. Ideally, the only difference between intentional and factual questions would be that the first concern intentional states and the latter do not. This is not the case in Brown's study: another important difference here concerns what I have referred to as "(in)transitivity" in Chapter 1, Section 1.1.2. This property affects strategies available for assessing statements for truth-value, either with respect to "reality", or, in this case, with respect to a story: the "intransitive" nature of embedded intentional states make sure that participants have to process the statement as a whole, whereas the "transitive" strings of factual statements can be checked against the reality of the story element by element. As a consequence, there are often "short-cuts" to the answer available in factual memory questions. Consider the following two questions from Brown's study:

> 4. Pete suspected that Henry was playing a prank on Sam
> 16. Sam found the Post Office closed and couldn't buy a tax disk for Pete

The answer to both questions according to the test is "false". As a participant, in order to determine this for question 4, one has to think about what Pete thought about Henry's intentions—a task in which all elements of the statement are somehow involved. However, seeing that 16 is false is a lot easier: all one needs to know is that Sam was not going to buy something for Pete but for himself.

This means that *in theory* there is a structural gap between the complexity of intentionality questions and factual questions. In practice, as pointed out above, a few of the intentionality questions in at least Brown's questionnaire exhibit broken chains of embedding, sometimes also enabling short-cuts for participants (I will say a few more words about this in the next section). The fifth general issue can thus be formulated as follows:

> issue (v) Questions exhibiting unbroken chains of embedding have to be processed as a whole, whereas in questions containing conjunctions and/or causal links each constituting element can be checked against the story separately. Given that the first category contains only (some of the) intentionality questions and the latter all factual memory questions (and the rest of the intentionality ones), this may have affected the observed difference in performance on both types of questions.

A similar argument is put forward by O'Grady et al. (2015) in their critical discussion of the existing mentalising questionnaire. They make the general point that a part of the intentionality and factual questions of the classic mentalising questionnaire (in this chapter represented by Brown's study) can be answered without processing the entire statement. In addition, they raise two specific objections regarding the factual control questions. First, they demonstrate that the intentionality questions use syntactically more

complicated sentences by counting the number of embedded clauses (they report a significant difference between the overall median being 0 versus 2 in factual versus intentionality questions). Second, they argue that the factual questions are inappropriate controls in the first place, since they do not involve *conceptual* embedding. They suggest to eliminate all possibilities for short-cutting, and to match the syntactic and conceptual complexity of intentionality and factual questions at every level by making use of "non-mental recursive concepts", such as relationships of possession or localisations in space and time. They implement these suggestions in their version of the mentalising test. Consider the following three example questions from their questionnaire ((7) is a mentalising question, (8) and (9) are factual memory questions; participants had to choose between alternative options A or B):

(7) A. Victor knows that Amy knows that Sheila intends that John thinks that she hasn't realised that he likes her.
B. Victor doesn't know that Amy knows that Sheila intends that John thinks that she hasn't realised that he likes her.
(complexity level seven[98]; see story 4, intentionality question 6 from the online supplementary material of O'Grady et al., 2015)

(8) A. Michelle and Nick started dating after a walk in the park, when Nick was tipsy, in the afternoon, on November 22nd, before Thanksgiving.
B. Michelle and Nick started dating when Nick was tipsy, during a walk, in the morning, on November 29th, after Thanksgiving.
(complexity level six; see story 2, control question 5 from the online supplementary material of O'Grady et al., 2015)

(9) A. Shaun is Sheila's supervisor Mike's boss's PA John's best friend's girlfriend Amy's brother.

[98] Note that O'Grady et al. (2015) start counting from level zero, as do following Kinderman et al. (1998). According to their study, a participant judging what a character in the story believes is working at first-level intentionality (instead of second-order as counted in most other studies and throughout this thesis), so they would refer to (7) as being level six. See also note 2 above.

> B. Shaun is Sheila's best friend Mike's supervisor's boss John's PA's girlfriend Amy's brother.
> (complexity level seven; see story 4, control question 6 from the online supplementary material of O'Grady et al., 2015)

O'Grady et al. argue that in all three questions both the concepts and the syntax are recursively embedded. This is clearly the case in the intentionality question in (7), but how about the memory questions in (8) and (9)? In a way, it is true that in (8) the walk in the park is *conceptually* embedded in Nick being tipsy, which is again conceptually embedded in an afternoon, embedded in the day November 22nd, embedded in the period before Thanksgiving. In some way, it can also be argued that the syntax of clause(s) indicating when Michelle and Nick started dating exhibits a recursive structure: a noun phrase embedded in a noun phrase, embedded in yet another noun phrase, and so on ([a walk in the park, [when Nick was tipsy, [in the afternoon, [on November 22nd, [before Thanksgiving]]]]]).[99] However, if the aim is to match the form of intentionality questions as closely as possible, this type of recursion does not do the job: as I have argued in Chapter 1, Section 1.1.2, a distinctive feature of embedded intentional states is that they exhibit "intransitivity": "A thinks that B thinks that C thinks that X" means something quite different than does "A thinks that C thinks that X", and if the first is true it does not follow that the latter is true as well. This is not the case in chains of conjunct clauses or causally related clauses: if the proposition "A and B and C" is true, it follows that "A and C" is true as well, and if "A leads to B leads to C" is true, it follows that "A leads to C"

[99] This is the interpretation O'Grady et al. (2015) seem to suggest on the basis of a different example they discuss in their paper. I think it can be argued for, however, I doubt whether this is the most natural analysis, given that it asserts the possibility of inserting the entire (recursively formed) noun phrase elsewhere in the sentence: [It was after a walk in the park, [when Nick was tipsy, [in the afternoon, [on November 22nd, [before Thanksgiving]]]], that Nick and Michelle started dating]. This might yield a "grammatical" sentence in the strict sense, but certainly not one that language users would easily produce in practice. Alternatively, one could argue that the relevant part of (8) is not one recursively formed noun phrase, but a string of serially combined noun phrases: [after a walk in the park,] [when Nick was tipsy,] [in the afternoon,] [on November 22nd,] [before Thanksgiving]. Following this analysis, which I think is the more credible from a language usage point of view, the sentence would be more like an elliptic version of a "narrative" presentation in multiple sentences: Nick and Michelle started dating after a walk in the park. Nick was a bit tipsy. It was in the afternoon on November 22nd, before Thanksgiving... See also Verhagen's (2010) discussion of "tail versus true recursion".

is true. If one of the elements in a string of conjunct or causally connected clauses does not fit with the story, the entire statement is false. However, if the intentional clause "B *doesn't know* about X" does not fit with the story, it does not follow that "A *thinks* that B *doesn't know* about X" does not fit with it. In other words, all that participants need to know for picking the correct option in the case of (8) is the answer to any one of the following questions: did the couple start dating before or after Thanksgiving? Was it on the 22nd or on the 29th? Was it in the morning or in the afternoon? In the case of (7), though, they *do* need to know something about what Victor knows that Amy knows that Sheila intends that John thinks that she thinks about him liking her. (However, note that participants may be crucially "aided" by the forced-choice design here: in fact, they only need to know whether Victor *does* or *does not* know about all this. More about this issue in the next section).

All in all, I think that question (8) embodies an unsuccessful attempt of the authors to solve all three of their own objections against the old test: taking a short-cut is possible after all, the question exhibits a different form of conceptual embedding compared to intentionality questions, and it exhibits a different form of syntactic embedding (or, arguably, no syntactic embedding at all, see note 99). A total of 14 out of 28 control questions in the relevant condition ("explicit"; see below) of O'Grady's study are structurally similar to (8) (cf. the supplementary material of O'Grady et al., 2015).

The other 14 questions in the relevant condition take the form of the one cited in (9). Do they do a better job of eliminating those objections? Each single clause is dependent on the clause to its left: for example, Sheila's supervisor's boss' friend is not the same person as Sheila's friend or Sheila's boss' friend. However, there is also a difference. It can be the case that John thinks that it is sunny outside, while Mary thinks that John thinks that it is raining. Yet it is not possible that John's mother is Mary's boss, while at the same time Mary's boss is not John's mother. In other words, Dennett's substitution test, explained in Chapter 1, Section 1.2.1, does not work for intentionality statements, but it does work for possessive relationships. As a consequence, spotting the one element that is at odds with the story can here also be done without processing the entire string: in the case of (9) all one needs to know is that Mike is Sheila's supervisor and not her best friend in order to choose the correct option: A.

Intentionality questions like (7) and memory questions like (9) may thus be matched in the sense that the concepts and the grammar exhibit recursive embedding[100] (the second and third objections are eliminated), but the possibility to short-cut processing keeps haunting these questions.

In the third study I have introduced at the beginning of Section 6.2, carried out by Haddad, the two types of questions are also matched more closely compared to the original questionnaire. Here the philosophy was not to introduce conceptual and syntactic embedding in the memory questions, but to concentrate on matching sentence length. The questionnaire as a whole is indeed well balanced in this respect (see the Online Appendix as referred to in note 95). In addition, broken-chain intentionality questions were eliminated. However, in this study it is also still possible to use processing short-cuts in the ways described in this section for quite a few of the questions.

Whether all of this is a problem or not depends on what is expected from the control questions: in, for example, Kinderman et al. (1998) factual questions serve to determine whether memory for details from the story is a factor limiting performance. However, given that the intentionality questions are different in a variety of respects, it may be problematic to make more precise comparisons between these questions (as has been done in, for example, Powell et al., 2010, where factual questions have been used as a baseline task). In some studies this may have led to interpretation problems, given that it is hard to determine to which part of the difference between factual and intentionality questions the findings have to be attributed. Although this problem is also noticed by O'Grady et al. (2015), their attempt to match the factual questions closer to the intentionality question comes with new problems, as I have discussed above. The same is true for Haddad's improved control questions.

[100] Interestingly, from the perspective of language usage, the resulting statements are also equally idiosyncratic: for both (7) and (9) it is impossible to think of a real-life context where such sentences would be uttered (except, perhaps, a humorous context like the one in the Friends episode discussed in the Introduction of this thesis).

6.2.4 True and false statements

Answering a question without processing it as a whole, i.e. taking a short-cut, is structurally more likely to be possible in questions where the right answer is "false". Consider the following factual memory questions taken from Haddad:

(10) Sam asked about finding a Post Office so that he could send a birthday present [false]
(question 13 in Haddad's questionnaire, see note 95)

(11) Sam couldn't send the card because when he got to the Post Office, it was closed [true]
(question 15 in Haddad's questionnaire, see note 95)

Haddad's version of the story about the Post Office is slightly different: the office prankster Henry goes by the name Helen and instead of wanting to buy a tax disc, Sam wants to send a birthday card to his grandmother. All that participants need to know to determine that (10) is false is that Sam was going to send a card, not a present, but to see that (11) is true, they have to know that Sam found the Post Office closed *and* that he was going to send a card. In other words, in "false" questions spotting *one* element that does not fit with the story is enough, whereas in "true" questions participants have to determine that *all* elements fit with the story. The higher the order of complexity of the question, the more this misbalance is amplified: in a second-order question it is one false element against two correct ones, but in a fifth-order question this ratio is one to five.

This issue is not limited to factual memory questions. Consider the following example:

(12) A. Megan wants Lauren to know that she, Megan, knows that Stephen knows that Elaine knows that Bernard feels she doesn't know him well enough to date, so that Lauren asks Stephen out [correct]

> B. Megan doesn't want Lauren to know that she, Megan, knows that Stephen knows that Elaine knows that Bernard feels she doesn't know him well enough to date, so that Lauren doesn't ask Stephen out [false]
> (see supplementary material of O'Grady et al., 2015)

Question (12) can be answered by knowing the answer to the simple (second-order) question: does Megan want Lauren to go out with Stephen or not? The clauses containing the "conclusions" ("so that Lauren asks Stephen out" and "so that Lauren doesn't ask Stephen out") have the same effect as had the broken chains discussed above: by inserting one causal link into the string of embedded clauses, it became possible to process the "cause" and the "consequence" as separate chunks. If one of the two did not match the story, the entire proposition was false. Similarly, in (12) it is possible to process the premise (involving eight orders of intentionality) and the conclusion (involving two orders) as separate chunks. Given the forced-choice design, plus the fact that the two conclusions present opposing scenarios, it follows that one of the conclusions *has* to be inconsistent with the story. What is an "eighth-order" question according to the authors, can in this case be answered by simply comparing two second-order propositions.

From this follows a crucial difference between the true/false design used in Brown and Haddad and the forced-choice design used in O'Grady: in the former it should be structurally harder to answer questions where the answer is "true", whereas in the latter this effect does not occur.[101] After all, in a forced-choice design it is *always* possible to find the right answer by spotting a single false bell. Whether or not this is an advantage can be debated: it certainly does increase consistency within the questionnaire, however, if the overall aim is to test participants on their ability to handle questions at different orders of complexity, it may be rather disruptive, as became clear with question (12).

Note that there are also false statements in which spotting the inconsistent element is possible only by processing the statement as a whole.

[101] In studies using a true/false design it may be advisable to calculate mentalising scores of individual participants on the basis of questions where the answer is "true" only. This was done by Van Duijn et al. (2014) in their study of mentalising performance in relation to school grades and personality traits: the questionnaire featured both true and false questions, but only the true ones were used to calculate the scores.

This goes for statements that exhibit an unbroken chain of embedding and, at the same time, do not involve unfamiliar or implausible elements. Consider the following examples:

(13) Pete thought that Helen wanted Sam to know that she realised that the Post Office was no longer on Elm St. [false]
(question 7 in Haddad's questionnaire, see note 95)

(14) A. Megan knows that Stephen doesn't know that Elaine knows that Bernard feels that she doesn't know him well enough to date [false]
B. Megan knows that Stephen knows that Elaine knows that Bernard feels that she doesn't know him well enough to date [correct]
(see supplementary material of O'Grady et al., 2015)

Both questions consist of statements exhibiting unbroken chains of embedding. On top of this, (13) refers to a scenario that is false, but that does have some credibility, given that the entire confusion in the story is exactly about Helen (alias Henry in Brown's version) intending or not intending to send Sam to the wrong street. Likewise, both options in (14) are credible, given that the story is precisely about Megan finding out what Stephen does or does not know about Bernard and Elaine.

The observations discussed in this subsection can be summarised in the issues (vi) and (vii) as follows:

issue (vi) Some questions exhibit a disproportionate increase in complexity per level for questions where the correct answer is "true" versus those where the correct answer is "false": if the correct answer is "true" participants have to check *every* element for consistency with the story, whereas questions where the correct answer is "false" can be answered by spotting *one* element that does not fit. This issue seems to be best avoided in questions that exhibit an unbroken chain of embedding and that present

a scenario that is (about) equally credible when thought of as true or false.

issue (vii) Given issue (vi), there is a crucial difference between questionnaires using a true/false design and those using a forced-choice design: when using the latter, spotting a single "false bell" is possible in *all* questions (although it is more difficult in questions exhibiting an unbroken chain of embedding and presenting two equally credible answer options).

6.2.5 Structure of interaction

As mentioned in the brief description of the study at the beginning of Section 6.2, the central aim of O'Grady et al. is to increase ecological validity by introducing movie clips that feature acted-out versions of the stories and questions. Whereas I make a distinction between narrative and propositional presentation in this thesis, they make a distinction between "explicit" and "implicit" presentation. Confusingly, at least in this context, is that by "explicit" they mean *both* the narrated stories and the propositional questions as used in the classic mentalising tests (which I have argued to be very different in kind in Chapter 2), and by "implicit" they mean their novel acted-out stories and questions. In four conditions they cross narrated stories with propositional questions (explicit-explicit), narrated stories with acted-out questions (explicit-implicit), acted-out stories with propositional questions (implicit-explicit), and acted-out stories with acted-out questions (implicit-implicit). The example questions cited in (7), (8), and (9) above are all from the propositional/explicit condition. The scripts for their acted-out counterparts from the implicit conditions are as follows (cited again from the supplementary material of O'Grady et al., 2015):

(15) A. Sheila: Anyway, so I chatted to Amy about it at the office and she reckons it's a good plan to just keep letting John think I haven't figured it out.
Victor: Yeah, Amy came by and told me about it all.
B. Sheila: Anyway, so I chatted to Amy about it at the office and she reckons it's a good plan to just keep letting John think I haven't figured it out.
Victor: That's weird, I spoke to Amy today. I had no idea that she knew about this situation.

(16) A. Nick: We started dating before Thanksgiving, in the afternoon on November 22nd. I was a bit tipsy and we'd just got back from a walk in the park.
B. Nick: We started dating after Thanksgiving, in the morning on November 29th, while we were walking in the park. I was a bit tipsy at the time.

(17) A. Amy: Yeah, it's really complicated. So your best friend is Sheila's supervisor Mike's boss's PA, John, and Shaun is my brother.
B. Amy: Yeah, it's really complicated. My boyfriend is PA to Sheila's best friend Mike's supervisor's boss, John, and Shaun is my brother.

An important innovation here according to the authors is that they, by introducing their acted-out presentation form, have managed to present *conceptually* embedded information without using embedded sentences. Note that this was already done in the narratives used in the original tests. However, a novelty indeed is that they also have two conditions in which the *questions* are presented without using embedded sentences.

It is indeed true that (16) and (17) look a lot closer to normal language usage than their propositional/classic counterparts (7), (8). However, it should be noted that (17) retains the problem I have pointed out above: it is still possible to find the right answer just by knowing whether it was before or after Thanksgiving, on the 22nd or the 29th, or in the morning or afternoon—each of

which is literally given in the story that the participants have just heard or seen in acted-out form. Question (18) is only minimally disentangled compared to the embedded-sentence version in (9). Yet if anything, this has made it even easier to spot the element that does not fit the story. The same goes for the acted-out/implicit version of questions (12) (boldface added):

> (18) 7. A. Megan: So, I'm thinking that Lauren needs to know what I heard. Right? Because if she knows what I know right now, about Elaine's crush, and Bernard's rejection, and that Stephen knows the whole thing…she'll work up the guts to ask him out! **So I'm going to tell her tonight**.
> B. Megan: Well, if you think about, if Lauren knew what I heard today – and if she knew that Stephen knew all about it too, about Elaine's crush and Bernard's weird reason for rejection and everything – she'd ask Stephen out. But I don't want her to do that, **so I'm not going to tell her**.

As pointed out above, it was possible with (12) to bypass the processing of the eight-order string of embedded intentional states. This is made even easier in (19), especially in the implicit-implicit condition, given that the acted-out/implicit version of the story ends as follows (boldface added):

> (19) Megan: Right! So, Lauren doesn't want to ask Stephen out because she thinks he's into Elaine – but if she knew that Stephen knows that Elaine likes Bernard, and that Stephen knows that Elaine's not into him, she might work up the guts to ask Stephen out.
> Chris: I guess…so are you going to tell her?
> Megan: **Yeah, I'm going to tell her the whole thing tonight**.

All participants need to remember to answer the question that allegedly embodies the highest level of complexity in the test, are the ten final words of the acted-out story.

O'Grady et al. (2015) present as their central finding that participants performed strikingly well at all levels of complexity, especially in their implicit-implicit condition. Performance did not drop drastically at any level, as was

claimed on the basis of the classic tests. They argue that this is probably due to the high ecological validity of their stimuli: according to them our natural human social ecology is full of higher-order intentional processing tasks. By designing stimuli that mimic this ecology as closely as possible, they claim to have shown that participants can almost effortlessly handle tasks up to eight orders of intentionality. However, I think that their test is highly ecologically valid precisely because participants do *not* need to process long strings of embedded intentional states, as shown in my analysis, but instead can rely on simple cues that bypass such processing when deciding between alternative scenarios. Other than in the classic tests, where participants are at least sometimes forced to deal with idiosyncratic statements that have to be processed as single units, O'Grady et al. (2015) allow participants to be normal mindreaders—and that is: lazy mindreaders.

6.3 Concluding remarks

In this chapter I have distinguished five central conclusions brought forward by the mentalising paradigm. In short:

- error rates of intentionality questions show a steep increase at complexity level 6, suggesting a limit to the ability of participants to handle embedded intentional states;
- between-subject variation suggests that some individuals have their limit around level 4, others around level 5, and again others around level 6 or even higher;
- although scores on intentionality questions and memory-control questions tend to be related, the steep increase in error rates at level 6 cannot be explained in terms of memory performance only, suggesting that there is something especially challenging about reasoning with embedded intentional states;
- participants' mentalising scores correlate with other measures of these individuals' social aptitude;

- participants' mentalising scores correlate with the amount of grey matter in relevant brain areas.

These conclusions are based on a series of studies, in which the general pattern has been replicated multiple times. However, nearly all of these studies have used versions of the same questionnaire, featuring stories describing social events, followed by forced-choice or true/false questions with embedded intentional states and factual details. In this chapter I have discussed issues connected with this way of testing mentalising competence, pertaining to the discrepancy between a narrative presentation in the stories and use of embedded sentences in the questions, inconsistencies regarding viewpoint layers that are sometimes "unpacked" and sometimes left implicit, structural differences between questions exhibiting embedded clauses and those with clauses that are related in different ways (e.g. conjunct or causally related), structural differences between true and false questions, and the gap between the use of intentional reasoning in the test and in real-life interaction.

The recent study by O'Grady et al. (2015) has raised doubt regarding the conclusions from the mentalising paradigm, partly based on the same issues with the questionnaire that I have pointed out here. However, as I have also discussed in this chapter, their own updated mentalising test, while having promising aspects, is also partly haunted by some of the old issues and for another part comes with new problems. Therefore, more research is needed before anything reliable can be said about the consequences for the central conclusions of the paradigm.

For now, I think, there is another important puzzle still unsolved: how can it be explained that the mentalising tests, despite all issues, produce meaningful variation correlating with measures of participants' social lives and overall aptitude in the social domain? I see two non-exclusive explanations:

- There are problems with quite a few of the questions, but others (such as (13) and (14) above) work well in the sense that they genuinely force participants to process the entire task they pose as one single unit. The meaningful variation in mentalising scores between participants could be principally driven by these questions. I have argued throughout this thesis that the processing of such unbroken chains of embedded intentional states is something we do not do by default when interacting,

but rather in exceptional cases, such as anticipating or repairing a misunderstanding. This can explain correlations with measures of social aptitude: if filling out a mentalising test is like handling exceptional situations in social interaction, those participants who are better at the test are also better at trouble-shooting whenever an interaction threatens to break down—a skill that may well be a good indicator of general social aptitude.

- The way in which the questions, and to some extent also the stories, present the fictive social situations used in the mentalising tests may be unnatural and problematic, however, the test itself creates a new and real situation of social interaction: the one between experimenter and participant. The experimenter has designed the questions and decided what the correct answers are. Some participants may be better at estimating what the experimenter wanted them to do, and be more motivated to figure this out in the first place. They may pick up even the smallest cues (like those in questions 7 and 10 from Brown's study, see also note 94 above) directing them towards the answer that the experimenter had in mind. This is a point that may theoretically produce biases in any test involving questionnaires, but in this particular case the bias happens to overlap with the target variable: being better at estimating the experimenter's intentions may indicate greater general social aptitude.

Both these explanations thus suggest that the associations between mentalising scores and other factors relevant to participants' social lives were not produced because the tests accurately "mirrored" the complexity generally involved in human interaction, but because they assessed participants on factors indicative of their ability to deal with special (partly extreme) cases of such interaction. This does not mean that these associations are no longer meaningful, but it does shed a different light on the foundations of social cognition: although it may be possible to assess general social aptitude using a task that forces participants to deal with embedded intentional states, this task should not be used as a model for what people do on a cognitive level in everyday social contexts.