Children's Explanation of Action:
Desires versus Beliefs in Theory of Mind

Tim Paul German

MRC Cognitive Development Unit.

and

Department of Psychology,
University College London.

Thesis submitted for the degree
of Ph.D. in Psychology,
University of London,
Acknowledgements

My first and greatest thanks go to Alan Leslie. Alan manages to combine being both extremely clever and very funny with no trouble whatsoever. His ability to inspire and his contribution to my work were made even more salient and acutely missed after his departure for the US. Ethereal supervision over e-mail just about got me the rest of the way.

I would also like to thank all the scientists and staff at the CDU. In particular: Margaret, Joanne and Jane in the babylab and Geoff Hall, with whom I survived testing at the wrong end of the Northern Line. I am grateful too to all the staff and children at the many nursery schools in which the studies were carried out. At UCL: Kate Bradford and John Draper, without whom I would still be in some terrible scrape, Celia Heyes, for forcing me to question what I thought I believed and making me welcome among the other animals in the animal lab, Jonck, who made statistics amusing and finally all my fellow PhD colleagues, especially Tristan Nokes and James Blair, whose (albeit conflicting) advice ensured that the thesis ended up being in English rather than some other language.

Finally I would like to thank Elaine, who helped beyond the call of duty, and my parents for support which was more than just financial.
Abstract

This thesis investigated children's theory of mind explanation. In Chapter 1 the theory of mind literature, which centres on the false belief task (Wimmer & Perner, 1983), is reviewed in the context of two theoretical approaches to theory of mind. The first approach accounts for development in terms of a shift in children's theory of mind competence (CS accounts). The second views development as involving an early (perhaps innate) and stable theory of mind competence, coupled with changes in the performance resources hypothesized to be responsible for deploying this competence (CC accounts). A further approach to commonsense psychology, which challenges the idea that the child has access to a "theory" of mind at all, was dealt with in Chapter 2.

In Chapter 3 a test of a recent CC theory was presented. Two experiments provided evidence compatible with the CC approach. However, problems in interpretation were identified that suggested the investigation of children's explanations would be profitable.

The next four chapters were concerned with explanation. The explanation literature was reviewed, with special attention paid to methodology. A scheme for the categorization of explanations in terms of their hypothesized resource requirements was constructed. Children's explanations for simple search actions under various conditions were then investigated.

The results of the first investigations (Chapter 5) showed that children's explanations of successful actions, that is, those that satisfied an agent's expressed desire, were invariably desire-based. Unsuccessful actions were explained in terms of the agent's (false) beliefs, but only when there was no simpler desire explanation. A simple model of explanation was proposed to account for these results. In the next investigation (Chapter 6), the relationship between belief-based action explanation and belief-based action prediction was shown to depend on local task structure rather than the child's presumed competence stage. Such an effect was argued to be more easily accommodated within a CC rather than a CS approach to theory of mind. Finally, the action explanation paradigm was applied to the study of Self-Other differences in theory of mind (Chapter 7). Children were shown to be able to use video evidence to explain their own false belief-based actions better than the actions of another agent. Because no conceptual difference is hypothesized to exist between reasoning about oneself and reasoning about another, the results were once again interpreted as favouring CC theory approaches over CS theory approaches.

In the last chapter, some conclusions are drawn and future directions are identified.
Table of Contents

CHAPTER 1
Theory of mind: evidence and theory

1.1 Introduction .......................................................... 17
   1.1.1 A terminological dispute ........................................ 17
   1.1.2 The false belief task ........................................... 18
1.2 Review overview ..................................................... 19
1.3 Types of theory of mind task ..................................... 19
   1.3.1 Belief attribution tasks ......................................... 19
   1.3.2 Deception tasks .................................................. 21
   1.3.3 Misrepresentation tasks ....................................... 22
   1.3.4 Tasks involving other mental states ....................... 23
   1.3.5 Significant omissions .......................................... 23
1.4 "Theory-theories": competence shift versus competence continuity .......... 23
1.5 Competence shift theory theories ............................... 24
1.6 The child-as-scientist theory (Wellman-Gopnik) ............. 25
   1.6.1 Desires to beliefs to representation ....................... 26
   1.6.2 Beliefs as copies ................................................ 27
   1.6.3 How stage-like is the competence shift? ................. 29
   1.6.4 A mechanism for theory change? .......................... 30
1.7 The Representational theory of mind theory (Perner) .... 32
   1.7.1 Situation theorists and the prebelief hypothesis ........ 32
   1.7.2 Belief and Representation .................................... 35
   1.7.3 Theory change in the RTM theory .......................... 36
1.8 Competence continuity theory theories ....................... 37
   1.8.1 Informational access .......................................... 38
   1.8.2 Reality masking (Mitchell; Robinson & Mitchell) .... 38
1.9 The Theory of Mind Mechanism (ToMM) theory (Leslie) ... 39
   1.9.1. Pretend play and the metarepresentation .............. 39
CHAPTER 2

Simulation and theory:
Knowledge, ability and information processing

2.1 Introduction ................................................................. 46
2.2 Radical simulation: “The big issue” .............................. 47
2.3 Is simulation simple? .................................................... 51
2.4 Conclusions ................................................................. 53

CHAPTER 3

Testing Fodor’s theory of the child’s theory of mind

3.1 Introduction ................................................................. 55
3.2 Fodor’s theory of the child’s theory of mind .................. 55
  3.2.1 An outline of Fodor’s theory ................................. 55
  3.2.2 Explaining the data with the heuristic theory .......... 57
  3.2.3 Experiment 3.1: introduction ............................... 58
3.3 Experiment 1 method ..................................................... 59
  3.3.1 Design ................................................................. 59
  3.3.2 Subjects ............................................................... 60
  3.3.3 Tasks ................................................................. 60
  3.3.4 Procedure .......................................................... 61
3.4 Results ................................................................. 61
  3.4.1 Effect of splitting the bait ................................... 63
  3.4.2 Interim discussion .............................................. 63
CHAPTER 4

Action explanation:
A converging measure of belief understanding?

4.1 Introduction ........................................................................................................... 83
4.2 General overview: backwards reasoning tasks ....................................................... 83
   4.2.1 Bartsch & Wellman (1989) .................................................................... 84
   4.2.2 Moses & Flavell (1990) .......................................................................... 87
   4.2.3 Robinson & Mitchell (1992) .................................................................... 89
4.3 Action explanation .................................................................................................. 90
   4.3.1 Procedural differences ............................................................................... 90
   4.3.2 Categorization of explanation types ......................................................... 92
4.4. A revised scoring methodology .......................................................................... 95
CHAPTER 5
Actions speak louder than words:
Children's explanations of simple search action

5.1 Introduction ................................................................. 104
5.2 Actions and reasons: a framework for explanation .............. 104
5.3 Incompleteness of explanation: theory ............................... 105
5.4 The primacy of desire: theory ......................................... 106
   5.4.1 Davidson's argument for the primacy of desire explanations .... 106
   5.4.2 Wellman's argument for a simple desire psychology .......... 107
   5.4.3 Fodor's argument for action prediction based on a desire heuristic ... 109
   5.4.4 Changing one's mind: do actions speak louder than words? ... 110
5.5 Reality explanations as justifications ............................... 111
5.6 Existing evidence ......................................................... 112
   5.6.1 Incompleteness of explanation: evidence ......................... 112
   5.6.2 The primacy of desire: evidence .................................. 114
   5.6.3 Reality explanations .................................................. 115
5.7 Experiment 5.1: Introduction and summary of predictions ...... 116
5.8 Experiment 5.1: Method ................................................. 117
   5.8.1 Design ........................................................................ 117
   5.8.2 Subjects .................................................................... 118
   5.8.3 Stories ..................................................................... 119
   5.8.4 Materials ................................................................... 119
   5.8.5 Procedure ................................................................. 119
5.16 Conclusions ........................................................................................................................155
  5.16.1 When do actions speak louder than words? .................................................... 155
  5.16.2 What of "strong" primacy of desire ................................................................. 156

CHAPTER 6
Factors affecting the explanation of search action by older children

6.1 Introduction .................................................................................. 157
  6.1.1 General goal .................................................................................. 157
  6.1.2 Competence limitations on the calculation of enabling conditions ......... 158
  6.1.3 Routine versus forced calculation of the belief/condition ................. 159
  6.1.4 Performance resources in explanation ........................................ 160

6.2 Interim summary: predictions .......................................................... 161

6.3 Evidence: justification and prediction ........................................................ 162

6.4 Experiment 6.1: introduction and predictions .................................................. 165
  6.4.1 Design .............................................................................................. 166
  6.4.2 Subjects .......................................................................................... 167
  6.4.3 Stories ............................................................................................ 167
  6.4.4 Materials ........................................................................................ 168
  6.4.5 Procedure ........................................................................................ 168
  6.4.6 Scoring of explanations ................................................................. 169

6.5 Results ............................................................................................... 170
  6.5.1 Correct Predictors ........................................................................... 171
  6.5.2 Incorrect predictors .......................................................................... 174
  6.5.3 Not Explanation question .............................................................. 174

6.6 Discussion ......................................................................................... 175
  6.6.1 CS theory compatible accounts ...................................................... 176
  6.6.2 CC theory compatible accounts ...................................................... 177
  6.6.3 Age considerations in experiment 6.1 ............................................ 180
  6.6.4 Conclusions .................................................................................. 181
CHAPTER 7
From Actions to Attitudes:
Children's use of their own actions to compute belief

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Introduction</td>
<td>183</td>
</tr>
<tr>
<td>7.1.1 Self versus Other: brief history</td>
<td>184</td>
</tr>
<tr>
<td>7.1.2 Perspective taking, informational causation or general representation?</td>
<td>184</td>
</tr>
<tr>
<td>7.1.3 The deceptive box task: some concerns</td>
<td>187</td>
</tr>
<tr>
<td>7.1.4 Do actions speak louder than words II?</td>
<td>189</td>
</tr>
<tr>
<td>7.1.5 Experiment 7.1 introduction</td>
<td>189</td>
</tr>
<tr>
<td>7.2 Experiment 1 Method</td>
<td>190</td>
</tr>
<tr>
<td>7.2.1 Design</td>
<td>190</td>
</tr>
<tr>
<td>7.2.2 Subjects</td>
<td>191</td>
</tr>
<tr>
<td>7.2.3 Materials</td>
<td>191</td>
</tr>
<tr>
<td>7.2.4 Procedure</td>
<td>191</td>
</tr>
<tr>
<td>7.2.5 Explanation question scoring</td>
<td>194</td>
</tr>
<tr>
<td>7.3 Results</td>
<td>195</td>
</tr>
<tr>
<td>7.4 Experiment 7.1 discussion</td>
<td>200</td>
</tr>
<tr>
<td>7.4.1 Explanations</td>
<td>200</td>
</tr>
<tr>
<td>7.4.2 Belief attribution</td>
<td>200</td>
</tr>
<tr>
<td>7.4.3 Order effects</td>
<td>201</td>
</tr>
<tr>
<td>7.4.4 Problems and improvements</td>
<td>201</td>
</tr>
<tr>
<td>7.5 Experiment 2 Method</td>
<td>202</td>
</tr>
<tr>
<td>7.5.1 Design</td>
<td>202</td>
</tr>
<tr>
<td>7.5.2 Subjects</td>
<td>203</td>
</tr>
<tr>
<td>7.5.3 Materials</td>
<td>203</td>
</tr>
<tr>
<td>7.5.4 Procedure</td>
<td>203</td>
</tr>
<tr>
<td>7.6 Results</td>
<td>205</td>
</tr>
<tr>
<td>7.7 Experiment 7.2 discussion</td>
<td>210</td>
</tr>
<tr>
<td>7.8 General Discussion</td>
<td>211</td>
</tr>
<tr>
<td>7.8.1 Relation to other findings</td>
<td>211</td>
</tr>
</tbody>
</table>
CHAPTER 8
Summary and conclusions

8.1 Introduction ................................................................. 217
8.2 CS versus CC theory ...................................................... 217
  8.2.1 Prediction tasks ....................................................... 218
  8.2.2 Explanation tasks .................................................... 219
8.3 Further research questions ............................................. 222
  8.3.1 Improvements and extensions in theory of mind explanation .... 222
  8.3.2 Domain specificity in causal explanation ......................... 223
8.4 Concluding comments .................................................. 224

References ....................................................................... 226
List of Tables

Table 1.1 Tasks and results of Wellman & Bartsch 1988 (Experiments 2 & 3) ......... 28
Table 1.2 Situation versus Representation theorists (Perner, in press) ................. 32
Table 1.3 Distinguishing pretence from mistaken actions based on false belief ...... 34

Table 3.1 Children's consistency in passing/failing Standard and Split bait tasks in Experiment ........................................ 63
Table 3.2 Children's consistency in passing/failing Standard and Eat bait tasks in Experiment 3.1 ..................................................... 64
Table 3.3 Children's consistency in passing/failing Ambiguous and Specific Immediate tasks in Experiment 3.2 .................. 74
Table 3.4 Children's consistency in passing/failing Ambiguous and Specific Intervening tasks in Experiment 3.2 .................... 75
Table 3.5 Children's consistency in passing/failing Immediate and Intervening Ambiguous tasks in Experiment 3.2 ............. 75
Table 3.6 Children's consistency in passing/failing Immediate and Intervening Specific tasks in Experiment 3.2 ...................... 76
Table 3.7 Children's errors on each task in Experiment 3.2 .................. 77

Table 4.1 Examples of each story type used by Bartsch & Wellman (1989. Experiment 1) ........................................ 86
Table 4.2 Summary of Action Explanation procedures used by Bartsch & Wellman (1989. Experiment 1) and Moses & Flavell (1990) ......... 91
Table 4.3 Categories of explanation and exemplars Bartsch & Wellman (1989) .......... 92
Table 4.4 Categories of explanation and exemplars for Moses & Flavell (1990) .......... 93
Table 4.5 Revised categorization scheme ........................................ 96
Table 5.1 Summary of stories used in Experiment 5.1 ........................................................121
Table 5.2 Children's consistency in the use of desire explanation in 1 Location condition in Experiment 5.1 .................................................................127
Table 5.3 Children's consistency in the use of belief explanation in the 1 Location condition in Experiment 5.1 .................................................................128
Table 5.4 Children's consistency in the use of desire explanation in 2 Location condition in Experiment 5.1 .................................................................129
Table 5.5 Children's consistency with respect to the content of desire in 2 Location condition in Experiment 5.1 .................................................................129
Table 5.6 Means and standard deviations of scores (out of 2) on memory for initial utterances in Experiment 5.1 .................................................................134
Table 5.7 Summary of stories used in Experiment 5.2 .......................................................139
Table 5.9 Children's consistency in the use of desire explanation in 1 Location condition (Experiment 5.2) .................................................................142
Table 5.10 Children's consistency in the use of belief explanation in the 1 Location condition (Experiment 5.2) .................................................................142
Table 5.11 Children's consistency in the use of desire explanation in 2 Location condition (Experiment 5.2) .................................................................143
Table 5.12 Children's consistency with respect to the content of desire in 2 Location condition, Experiment 5.2 .................................................................144
Table 5.13 Means and standard deviations of scores on memory for initial utterances in Experiment 5.2 .................................................................146
Table 6.1 Children's consistency in desire/cause or belief/condition explanation across conditions in Experiment 6.1 .................................................................172
Table 7.1 Number of children offering belief/condition explanations for Self and Other conditions in Experiment 7.1 .................................................................196
Table 7.2 Number of children correct on belief attribution question for Self and Other conditions in Experiment 7.1 .................................................................198
List of Figures

Figure 1.1 Types of belief attribution task ................................................................. 21

Figure 3.1 Number of children in Experiment 3.1 passing Standard, Split Bait and Eat Bait versions of the false belief task ................................................................. 62

Figure 3.2. Specific Immediate Story ................................................................. 69

Figure 3.3. Ambiguous Immediate Story ................................................................. 69

Figure 3.4. Ambiguous Intervening Story ................................................................. 70

Figure 3.5. Specific Intervening story ................................................................. 70

Figure 3.6 Performance by youngest, middle and oldest age groups in each condition in Experiment 3.2 ................................................................. 72

Figure 4.1 Belief-desire reasoning scheme of Bartsch & Wellman (1988) ................. 85

Figure 5.1 Why did the chicken cross the road? (Reprinted from Larson, 1991) .......... 105

Figure 5.2 Frequencies of each explanation category in Consistent and Inconsistent stories in 1 and 2 Location conditions, Experiment 5.1 ......................................................... 125

Figure 5.3 Children’s answers to the desire prompt question for the Inconsistent action in each condition in Experiment 5.1 ................................................................. 131

Figure 5.4 Children’s answers to the belief prompt question for the Inconsistent action in each condition in Experiment 5.1 ................................................................. 133

Figure 5.5 Number of children in each explanation category for each action type in each condition of Experiment 5.2 ................................................................. 140

Figure 5.6 Children’s answers to the desire prompt for the Inconsistent action in 1 and 2 Location conditions in Experiment 5.2 ................................................................. 144

Figure 5.7 Children’s answers to the belief prompt for the Inconsistent action in 1 and 2 Location conditions of Experiment 5.2 ................................................................. 145
Figure 5.8 The explanation model in the situation where the action is consistent with the available desire information ....................................................... 152

Figure 5.9 The explanation model in the case where the action is inconsistent with desire information, and the location searched contains no object ....................... 153

Figure 5.10 The explanation model in the case where the action is not consistent with the desire information and the location searched contains another object ...... 154

Figure 6.1 Number of children offering cause and condition explanations in Prediction and No prediction conditions in Experiment 6.1 ............................................ 171

Figure 6.2 Number of "Belief" and "Informational access" condition explanations in Prediction and No Prediction conditions in Experiment 6.1 ......................... 173

Figure 6.3 Number of Children offering desire/cause and belief/condition explanations to for the "Not" explanation question in Experiment 6.1 ......................... 174

Figure 6.4 Possible explanations in the No Prediction condition ......................... 178

Figure 6.5 Possible explanations in the Prediction condition ......................... 179

Figure 7.1 Set-up of experimental room for Experiment 7.1 .............................................. 192

Figure 7.2 Number of children offering Desire/cause, Belief/condition and Nil explanations in Self and Other conditions in Experiment 7.1 .............................................. 195

Figure 7.3 Number of children correct on the false belief attribution question in each condition in Experiment 7.1 ................................................................. 197

Figure 7.4 Number of children correct on the belief attribution question under Self-Other and Other-Self presentation orders in Experiment 7.1 ......................... 199

Figure 7.5 Number of children correct on Recall/Prediction task in Self and Other conditions in Experiment 7.2 ................................................................. 205

Figure 7.6 Number of children offering Desire/cause, Belief/condition and Nil explanations in each condition in Experiment 7.2 .............................................. 206

Figure 7.7 Number children correct on the belief attribution question in Self and Other conditions in Experiment 7.2 ................................................................. 207
Figure 7.7 Number of children offering Desire/cause, Belief/condition and Nil explanations in each condition in Experiments 7.1 and 7.2 combined ...................... 208

Figure 7.8 Number children correct on the belief attribution question in Self and Other conditions in Experiments 7.1 and 7.2 combined ...................... 209
CHAPTER 1

Theory of mind: evidence and theory

1.1 Introduction

"Theory of mind" refers to the capacity to predict and explain the behaviour of others by the attribution of mental states such as beliefs and desires. This thesis examines the development of children's theory of mind, paying special attention to their explanation of action, an area that has been relatively neglected. In this first chapter I present a review of the experiments that have been performed, and the theories that have been devised, in charting the child's acquisition of a theory of mind during the preschool period. The review, like much of the literature, is concerned largely with action prediction and belief attribution. I deal specifically with explanation experiments in Chapter 4.

1.1.1 A terminological dispute

The term "theory-of-mind" was coined by Premack and Woodruff (1978):

An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of such a kind is properly viewed as a theory because such states are not directly observable, and the system can be used to make predictions about the behaviour of others. (Premack & Woodruff, 1978, p515).

The term "theory-of-mind" is theory laden in more than just the obvious sense, because within both philosophy of mind and psychology, the notion that these capacities are actually subserved by some theory-like structure does not go unchallenged. Perhaps for this reason, this package of capabilities has travelled under many other names - folk psychology, commonsense psychology, mentalizing and mindreading are just a few. The theories I will discuss all see themselves as brands of "theory-theory" (Morton, 1980), though there are a variety of views about what are the implications for cognitive structures of attributing a "theory" to the child. Despite some misgivings voiced in the present chapter, I will explain in Chapter 2 why I persist in talking of a "theory of mind".

17
1.1.2 The false belief task

The "central" task around which much of the theoretical work pivots is the "false belief" task. The form of this task was suggested independently by Bennett (1978), Dennett (1978) and Harman (1978), in their commentaries on Premack & Woodruff (1978). Premack & Woodruff had attributed a theory of mind to chimpanzees on the basis of what they saw as evidence that the chimpanzees could impute intention to human actors. According to Bennett, Dennett and Harman however, the chimpanzees might have responded on the basis of their own intention. The type of evidence required for the claim, as Harman puts it, is as follows:

There is no reason to require more of the chimp than of an adult human being. Suppose that a subject chimpanzee sees a second chimpanzee watch a banana being placed into one of two opaque pots. The second chimpanzee is then distracted while the banana is removed from the first pot and placed into the second. If the subject chimpanzee expects the second chimpanzee to reach into the pot which originally contained the banana, that would seem to show it has a conception of mere belief. (1978, p576).

The diagnostic test, as it were, for an individual to be said to possess a theory of mind is thus for that individual to be able to attribute a belief to another that conflicts with the first individual's belief.

Wimmer & Perner (1983) seized on this specification, and carried out a similar task with children aged between three and nine years. In the task the child watched a character who at $t_0$ placed a "bait" object in Location A and then departed. While she was away another person moved the bait object to Location B (creating a new situation at $t_1$). The child was asked to predict where the first character would look for the bait item on her return. Wimmer & Perner found that children younger than five years of age performed very poorly on this task. Baron-Cohen, Leslie & Frith (1985) slightly simplified the task content, in order to present it to autistic children. On this simplified version (the Sally/Anne version), a majority of 4-year-olds gave the correct answer, that the character would look in A, while most 3-year-olds predicted that the character would look in B, the new location. The present consensus is that a striking shift in performance on this task occurs between age three and four (see for example Astington & Gopnik, 1991), and the Sally/Anne version of the false belief task has hence become the standard version.
1.2 Review overview

There are now hundreds of experimental papers that report the results of tests in which children were presented with false belief tasks. I will be concentrating on the attribution of false beliefs in the review that follows. I have described above the paradigm that was adopted first, but before presenting any more specific experimental details, I will outline a framework that identifies the different forms that these tasks can come in: a classification of belief attribution and other theory of mind tasks. I will then consider the theory of theory of mind, and having drawn a distinction between two main classes of "theory-theory", I will outline some selected positions within each class. The theoretical accounts will be supplemented with the results from experiments or groups of experiments as required. By the end of this, I will have come down in favour of one class of theory-theory, and it is in testing a specific instance within this class that the empirical work begins in Chapter 3.

1.3 Types of theory of mind task

As one would expect, the largest group of tasks concerns the attribution of beliefs, and this group of tasks will be subdivided. I propose to introduce an additional three types of task, although arguably some of these types are actually belief attribution tasks of one sort or another.

1.3.1 Belief attribution tasks

Belief attribution tasks require the child to attribute a belief to an agent, as outlined in the original false belief task of Wimmer & Perner. There are two main factors which I will use to classify the procedures of belief attribution tasks. The first factor concerns the circumstances/situation in which the belief is created. The second factor is whether the question to the child is a direct question about the agent's belief or alternatively a question that gets at the agent's belief indirectly, such as asking where they might look or what they

---

1 Note that in the first instance this will be a classification of types of task procedure, and will consider features such as the questioning and situation under which the belief attribution takes place. The experiments can also of course be classified in terms of the results that they yield and which theoretical positions they, i) were designed to support and ii) actually do support. I will endeavour to provide analysis at this level after having presented the "procedural" classification.
might say.

To illustrate, in Wimmer & Perner (1983) the false belief is created in an action based scenario (where an agent has a desire for a bait object) through an object changing location. False beliefs created in these sorts of circumstances I will describe as action-based. The test question asked in Wimmer & Perner is an *indirect* assessment of the character's belief, because it actually requires a prediction of an action that will follow as a consequence of the false belief.

Consider now a different false belief task, first used by Hogrefe, Wimmer & Perner (1986), the use of which has since become extremely widespread. In this task a false belief was implanted by the use of a deceptive container: a child was confronted with a chocolate box (usually Smarties, hence the task is known as the "Smarties" task), and the child's belief about the contents was solicited. The box was opened and revealed to contain a different object (e.g. a pencil). The child was asked to state *what she thought was in the box before it was opened* (or alternatively what another child waiting outside would think was inside the box). Once again, children below the age of about four had difficulty with this, stating that they thought the box contained pencils. Above this age, children were able to correctly identify their past false belief. This type of task I will describe as the deceptive box paradigm.2

Note that this second task asks the child *directly* about the agent's belief, but could easily have used an indirect question, such as asking the child what they said was in the box. Likewise, in the Wimmer & Perner action prediction task, a direct question about the character's belief when she came back into the room could have been adopted instead of asking children to predict where she would search. Any combination of these factors is in principle possible, and the breakdown is shown in Figure 1.1.

This coarse-grained categorization can be refined. There are for example many different indirect measures that have been used to assess belief understanding. These

---

2 Many belief attribution tasks contain an element of deception in the "cover story", in choosing the terms I have tried to emphasize the situations in which the beliefs are implanted. In the "Smarties" task, however, an additional feature is that an utterance is elicited as an indication of the child's commitment to the belief. This feature is conceptually important, and in Chapter 7, I pick up on the "utterance-based" nature of this group of tasks, when I look at Self-Other differences in theory of mind. Deceptive box tasks are distinct from tasks that employ active deception (i.e. implanting a false belief in another) as an indirect measure of belief understanding (see Section 1.3.2).
include; what an agent would say (Wimmer & Hartl, 1991), which item an agent's utterance refers to (Robinson & Mitchell, 1992) and even where the child looks in anticipation when an agent is about to reappear (Clements & Perner, in press). These details will be made clear in the course of the review.

<table>
<thead>
<tr>
<th>ACTION BASED</th>
<th>UTTERANCE BASED</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
<td></td>
</tr>
<tr>
<td>INDIRECT</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.1 Types of belief attribution task. An example of each type is given.

1.3.2 Deception tasks

Many of the false belief attribution tasks described above use "deception", be it as a "cover story" (e.g. playing a trick on the protagonist), or to create a false belief in the subject (which must then be identified). The tasks I categorize in this section however, require children to deceive another agent by implanting a false belief. These tasks were originally assumed to be an indirect measure of belief understanding; after all, if one can deceive an agent by implanting a false belief, then one can understand false belief, or so the logic goes. Chandler Fritz & Hala (1989) made this sort of argument in presenting a deception task to very young children. Children of age two and a half were shown to be capable of laying a false trail to "trick" a protagonist out of treasure. More recently however, it was shown that the 2-year-olds did not discriminate between helping and
deceiving a protagonist, laying false trails under both conditions (Sodian, Taylor, Harris & Perner, 1991).

The assumption that these tasks are a form of belief attribution has since been questioned, and a distinction has been drawn between their deceptive and strategic demands (e.g. Sodian & Frith, 1992, Hughes & Russell, 1993). I will return to this issue in Section 1.9.4.

1.3.3 Misrepresentation tasks

There are two main types of misrepresentation paradigm. The first, which dates back almost as far as the original false belief task, involves an object that appears to be one thing but is really another (e.g. a sponge that looks like a rock). This "appearance-reality" task, as it is known, was introduced by Flavell & his colleagues (Flavell, Flavell & Green, 1983). After having been shown that the "rock" was really a sponge, children were required to state both what the item looked like and what it really really was. Children made realist errors (looks like sponge, really is sponge) before age four, but were accurate thereafter.3 There is obvious similarity between this task and the "Smarties" task, which has led some authors (e.g. Perner, Leekham & Wimmer, 1987) to argue that asking what the object "looked like" is really an indirect way of asking for a belief attribution: what would a given person think this item was if they saw it?

The second type of misrepresentation task is closely tied to theoretical positions that stress understanding of the role of "representation" in theory of mind. These tasks involve children assigning values to various types of public representation (e.g. photographs, maps, signs) under conditions similar to the false belief task. For example, in the false photo task (Zaitchik, 1990) a photo is taken at t₀, the target object is replaced by another object at t₁ and the child is asked to say what object is in the photo. Once again, the results showed a shift in performance at around age four. The important issues that arise from misrepresentation tasks are the relationship that they have with belief attribution tasks (Zaitchik, 1990; Perner. Leekham, Myers, Davis & Ogders, 1994), and the nature of this relationship in normal and autistic populations (e.g. Leekham & Perner, 1991; Leslie &

---

3 In different versions of the task, the errors were phenomenalist (e.g. looks like appearance, really is appearance).
1.3.4 Tasks involving other mental states

From the discussion so far it might appear that the ability of young children to attribute beliefs is the only important topic in "theory of mind". It is true that this is where most emphasis is placed, in part because of its perceived philosophical importance as an "acid test" of theory of mind (Bennett, 1978; Dennett, 1978; Harman, 1978; and also Perner, 1988, 1991, in press). However, the ability of children to understand other mental states, specifically those of pretend and desire, is also important in theoretical accounts of theory of mind acquisition. I will include descriptions of tasks involving the attribution of these other mental states in this review. The mental state of desire turns out to be especially important when one considers children's explanation of action, and more detailed review of the relationship between desire reasoning and belief reasoning appears in subsequent chapters.

1.3.5 Significant omissions

Before moving to the theoretical accounts, I will note briefly the omission from this review of experiments concerned with the development of theory of mind before age two. The theories do not all assume that there is no story to be told before two, and there is considerable diversity in the abilities that have been characterized as "precursors" to theory of mind (e.g. imitation - Meltzoff & Gopnik, 1993; shared affect - Hobson, 1990; and shared attention - Baron-Cohen, 1994). I will not be dealing in detail here with any of these proposals, since I am concerned with the characterization of the later theory of mind. For a clear and ambitious "full story", see Baron-Cohen (1994) and associated commentaries (e.g. Happé; Heyes & German; Mitchell & Saltmarsh; Segal; all 1994).

1.4 "Theory-theories": competence shift versus competence continuity

The theory-theories that I will present in the course of the rest of this review can be placed into one of two broad categories. I draw the distinction in terms of the general approach to accounting for what is a striking shift in performance on the false belief task (Wimmer & Perner, 1983). The first class of theories attempts to account for this shift in
terms of a shift in competence. Roughly, this amounts to attempting to characterize what a child "knows about" or "does not know about" at each age. The younger child has an absent or deficient concept of "belief", "false-belief" or "representation", and so fails the false-belief task. The older child has acquired the concept, and so passes. The child's theory of mind undergoes a number of shifts on this view, and gradually becomes closer to the adult theory. Theories of this type are differentiated in terms of i) precisely what theories/concepts the younger child is credited with and ii) how change in the child's theory is supposed to come about; they are similar in that they consider the end-point of this process to be a representational understanding of the mind. I will use the term Competence Shift theory-theories (CS theories) to refer to this class. Many authors subscribe to a theory-theory of this type (Gopnik & Wellman, 1992; Perner, 1991; Wellman, 1990) and they have hence been argued to represent the "consensus view" among developmental psychologists (Gopnik, 1993).

The second class of theory-theory rejects the notion that stage-like shifts in performance need be explained in terms of stage-like conceptual development. As far as competence is concerned, children are assumed to have essentially the same theory of mind concepts throughout the period of the shift. The tasks used to "diagnose" theory of mind, however, also require certain other general processing abilities, or performance demands, and it is the development of some general resource that allows older children to pass tasks that younger children fail. On this view then, it is not the theory that changes, but rather the efficiency with which it can be deployed. Theories within this class are differentiated in terms of what the processing resource(s) might be. I will use the term Competence Continuity theory-theory (CC theories) to refer to this class.

1.5 Competence shift theory theories

There are two grains at which the distinctions between CS theories can be cast. Firstly, there are differences in the nature of the theory that the child is posited to move from to get to a representational understanding of the mind. Gopnik (1993a) mentions how difficult it is to characterize a conception of the mind that is profoundly different from our

---

4 In fact, there are also differences in precisely how seriously the notion of representation is implicated. This is discussed in more detail in the sections dealing with specific theories (e.g. Sections 1.6 and 1.7).
own. The three-year old has been described by different authors as having "cognitive connections" to the world (Flavell, 1988), having a "copy theory of belief" (Wellman, 1990), having a view analogous to a "Gibsonian or Dretskean (Dretske, 1981) view of the mind" (Astington & Gopnik, 1991), and as being a "situation theorist", (Perner, 1991). These turn out to be remarkably similar proposals: indeed, in the words of Gopnik herself:

[The theory-theory accounts advanced by myself and others (such as Flavell [1988], Wellman [1990], and Perner [1991]) have shown a striking degree of convergence (to the dismay, I sometimes suspect, of their independent-minded advocates). (1993b, pp 90-91).

There are, nevertheless, points of difference worthy of discussion, which I will deal with in the next sections.

Differences can also be cast at a more interesting grain. The two specific issues that will come up in the review concern i) how much importance is attached to the notion of representation in the theories and ii) how much emphasis is placed on the phenomenon of transition between theories. I will outline two CS theories, which have complementary mixes of the two qualities expressed above: the Wellman-Gopnik theory, which stresses the "mechanism" of theory change, and the Perner theory, which is clearest in advocating a full "representational theory of mind" (RTM) in the young child.

1.6 The child-as-scientist theory (Wellman-Gopnik)

As noted above, the initial empirical findings in the theory of mind domain revealed a sharp divide between 3- and 4-year-olds on a number of tasks: notably false belief tasks, both action-based and deceptive box versions (Wimmer & Perner 1983; Hogrefe et al., 1986; Perner et al., 1987), and appearance-reality tasks (Flavell et al., 1983). In addition, investigators reported high correlations in performance between all these tasks (Astington & Gopnik, 1988). More recent findings, such as the "false" photograph task (Zaitchik, 1990), and scepticism about demonstrations of deception in young children (Chandler et al., 1989; Sodian et al., 1991: see Section 1.3.2), buoyed the general outlook: that some (severe) conceptual deficit in the 3-year-old, coupled with some conceptual development subsequently, could account for all these results. The deficit in question is held in this theory to be the absence of a "representational model of mind" (Forguson & Gopnik, 1988; Gopnik & Wellman, 1992). For example in Gopnik & Wellman (1992):
The child's view of the mind becomes fully intentional. To use Dretske's terminology, perceiving becomes perceiving that, and desiring becomes desiring that, we might even add, that believing becomes believing that (Dretske, 1981). (1992, p153).

But what of the younger child? What sort of conceptual competence gives way to a representational understanding of mind?

1.6.1 Desires to beliefs to representation

For the Wellman-Gopnik theory, the initial conception of the mind is mentalistic, and two basic categories of mental state cover each of Searle's (1983) "directions of fit" between mind and world. Desire covers the world-to-mind direction (in that the state of the world is fitted to meet the desire), while perception covers the mind-to-world direction (in that the perception is made to fit the contents of the world).

Desires and perceptions are held to be understood in non-representational terms at this stage: desires as simple drives toward objects and perceptions as simple awareness of objects. In neither case need the child conceive of a complex relationship to a proposition. The evidence for desire understanding comes from Wellman & Woolley (1990), who showed that (older) 2-year-olds could predict the cessation of an agent's first action (searching for a pet) and the start of a second action (taking the pet to school), when their desire was satisfied (they found the pet). They could equally predict the continuance of the search action when the desire had not been satisfied (they found nothing or a different pet). A critical discussion of this experiment appears in Chapter 5, but it is interesting to note that the simple drive notion of desire has attracted considerable criticism. Firstly, it is not complicated enough to do the work that Wellman wants it to. As Leslie (in press) points out, denying the child the ability to represent a propositional content as the object of the desire makes the putative desire notion almost useless for predicting behaviour:

When Wellman's child thinks of Mary as wanting an apple, he is incapable of representing what Mary wants to do with the apple. When Wellman's child thinks that Billy wants a swing, he cannot represent whether Billy wants to sit on the swing, swing on the swing, just sidle up and be close to the swing, or anything else specific. Billy simply wants the swing, full stop. (Leslie, in press).

Wellman himself has demonstrated that children are able to predict specific actions from desire. Wellman's desire notion cannot explain how the 2-year-old children in the Wellman
& Woolley (1990) task, described above, can predict that an agent will visit school once
(s)he has found the desired object. A drive notion of desire cannot allow a child to predict
the agent will visit school once the pet is found, because “to take the pet to school” is a
proposition, not simply an object.

Secondly, children must take at least some aspect of the belief component into
account by age two because they are able to understand situations in which agents do not
act in way that is objectively conducive to satisfying their desires, for example, when agents
are pretending (Leslie, 1987, 1988, 1994). Leslie’s own position with respect to desires is
detailed in Section 1.9. Now, I turn to the next stage in the Wellman-Gopnik theory.

1.6.2 Beliefs as copies

The next stage in competence in the Wellman-Gopnik theory-theory is based on an
elegant set of experiments done in the late 1980’s (Wellman & Bartsch, 1988). These
experiments were action-based, indirect belief attribution tasks, using a question about
where an agent would look for a bait object. The authors manipulate the story character’s
beliefs, whether these beliefs correspond to reality and whether reality is known to the child.
A summary of the tasks and results appears in Table 1.1 below.

In seven out of eight tasks that the children’s performance was good; the only task
at which 3-year-old children performed poorly was the task that required them to
understand an explicit false belief. This pattern was interpreted as indicating that 3-year-olds
understand belief, but not false belief. For 3-year-olds, beliefs are copies of reality.

Perner (1989) criticized this claim, arguing that the data do not support it. He
outlined three strategies that the children might have adopted in solving the tasks. None of
these strategies required the child to understand belief, and two of them apply to all the
seven tasks on which Wellman & Bartsch showed better performance in the younger
children. He concentrated most on the third of these three strategies, the claim that
"thinking" is not necessarily "belief".
Table 1.1 Tasks and results of Wellman & Bartsch 1988 (Experiments 2 & 3)

<table>
<thead>
<tr>
<th>Task</th>
<th>Features</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Belief</td>
<td>Agent's belief explicitly mentioned, actual location unknown.</td>
<td>77%</td>
</tr>
<tr>
<td>Not Belief</td>
<td>Agent's belief that object not in location mentioned, actual location unknown.</td>
<td>90%</td>
</tr>
<tr>
<td>Not Own Belief</td>
<td>Subject's belief solicited, agent given opposite belief, actual location unknown.</td>
<td>83%</td>
</tr>
<tr>
<td>Changed Belief</td>
<td>Agent has one belief, then changes to another, actual location unknown.</td>
<td>90%</td>
</tr>
<tr>
<td>Inferred Belief Control</td>
<td>No belief mentioned, agent said to have seen object in one location, actual location not known.</td>
<td>88%</td>
</tr>
<tr>
<td>Discrepant Belief</td>
<td>Agent believes object in one location but not other, objects seen in both locations.</td>
<td>63%</td>
</tr>
<tr>
<td>Explicit False-Belief</td>
<td>Object said to be in one location, agent believes object is in other (wrong) location.</td>
<td>16%</td>
</tr>
</tbody>
</table>

This claim is part of Perner's own theoretical position, and it will be outlined in greater detail in Section 1.7.1. Briefly however, the claim was that thinking might have been understood as \textit{thinking of} rather than \textit{thinking that} by the children in these experiments. For example, the Discrepant Belief experiment becomes:

\textit{Look there are bananas in the cupboard and bananas in the fridge. Jane wants a banana. Jane only thinks of the bananas in the cupboard; she}
doesn't think (much)\(^5\) of the bananas in the fridge.

The children passed those seven tasks where interpreting "see" and "think" as indicating preference lead to the correct response. The other task, the one that required an understanding of "think" as belief, was not passed.

Wellman & Bartsch (1989) replied to these criticisms. They claimed that Perner's strategies did not cover the results in the way that he claimed. Although two of Perner's strategies applied to each of the tasks, it was not always the same two, implying that children must have been able to switch selectively between them. In addition, the strategies give a clear prediction on the Explicit False-Belief task - that children should have passed.\(^6\)

There are, nevertheless, problems with the copy theory of belief. For example, Perner (1991) argued that the copy theory of belief cannot in fact account for performance on the Not Own task (see Table 1.1). Perner's argument was that if children understand belief as a direct copy of reality, then this commits them to assuming that an object is in two locations at the same time: in the location of which the agent's belief is a copy and in the (other) location of which their own belief is a copy. Fodor (1992), has made a similar point.

1.6.3 How stage-like is the competence shift?

I present in this section some of the evidence that proves problematic for the Wellman-Gopnik theory. This evidence is in fact problematic for CS theories in general, but I present parts of it here because it was instrumental in provoking Wellman & Gopnik's subsequent emphasis on the "mechanisms" of theory change (Wellman, 1990; Gopnik & Wellman, 1992; Gopnik, 1993). I will outline two important experiments that undermine the notion of a sharp competence shift during children's development of theory of mind, a deceptive box example and an action-based example.


\(^5\) Perner includes this as a reminder that thinking can carry information about preference.

\(^6\) Perner has recently introduced the notion of "preface" as an elaboration of his "3-year-old" theory (Perner, in press). I will deal with it in Section 1.7.1.
The children posted a picture of Smarties into an opaque container (a "postbox") at the time of their first belief-based utterance (i.e. before the box had been opened). When these children were asked the direct belief attribution question, the great majority were able to answer correctly. In a control condition, where children posted an irrelevant picture (of Mickey Mouse), there was no improvement. The authors suggest that providing a "marker" for the child's belief, can overcome the child's tendency to respond by referring to reality. Mitchell's "reality masking hypothesis" is a version of competence continuity theory-theory, and I will postpone more detailed discussion of the proposal until later. For the moment, note that false belief acknowledgement below age four is not in principle compatible with CS theories.

The action-based experiment was performed by Siegal & Beattie (1991), but received little attention in the literature. The authors used an indirect (where will x look..?) attribution task based on the Explicit False Belief task of Wellman & Bartsch (1988; see also Table 1.1). In the experimental condition, the question carried extra information that specified that it was where they would look first that was important. This manipulation improved performance dramatically. Moreover, a second experiment ruled out the possibility that asking about the first look would demonstrate that the first look would definitely fail. The second experiment used the same manipulation, but required the children to predict action from a true belief. Children were not directed into picking the empty location, as would be expected if some simple strategy was operating. Like the Mitchell & Lacohée (1991) study, this result embarrasses CS theories.

These two experiments, among others, have blunted what was originally seen as quite a sharp shift in performance, and thereby undermined the neat mapping between stage-like performance and stages of competence. It is this counter-evidence that has provoked a newer slant to the Gopnik-Wellman position, one that emphasizes an analogy between the scientist-as-theorist and the child-as-theorist.

1.6.4 A mechanism for theory change?

Some theorists have drawn an analogy between theory change in science and theory change in the development of false belief theory海峡.
change in cognitive development (Carey, 1985, 1988; Karmiloff-Smith, 1988). Gopnik & Wellman (1992) appeal to the analogy to account for the transition between a child's successive theories of mind. Their point is summarized as follows.

In science, there are periods of time where theorists find themselves between two theories. For example, the 50 years between the publication of the ideas of Copernicus, and Kepler's theory of planetary movement. In these periods there is an accumulation of counter-evidence which gradually forces scientists to accept a new theoretical perspective:

In a transitional period the crucial idea of the new theory can appear as an auxiliary hypothesis couched in the vocabulary of the original theory, or be used to deal with particularly salient counter-evidence (Gopnik & Wellman, 1992, pp 156-157).

Gopnik & Wellman are now in a position to explain the salient counter-evidence described above to the competence shift theories: the children in these situations are positing *ad hoc* auxiliary hypotheses to pass these tasks, but have not yet learned to apply the new (representational) theory more generally. The new representational view, when it is acquired after this transitional period, is one that totally replaces the old theory:

*The representational model of the mind* is an essential part of the commonsense adult notion of intentionality. (Gopnik, 1993a, p6).

There are, however, problems with the child-as-scientist metaphor when it is used as explanation. The most serious is the underspecification of the "mechanism of change". In the case of the auxiliary hypotheses posited by the child, for example, it would be desirable to have some indication of at least i) where these theories might come from, ii) what constrains the admissibility of the hypotheses tested and iii) why the same specific range of hypotheses is tested by all children given the large number of possible hypotheses (see also Russell, 1992; Leslie & German, in press).⁸

---

⁸There is also a touch of irony about this strong emphasis on the "child as scientist" analogy. Another way of looking at the situation is to view Gopnik & Wellman's appeal to children positing "auxiliary hypotheses" as an auxiliary hypothesis itself. After all, it is used to deal with salient counter-evidence to the CS theories of theory of mind. Not only is this "auxiliary hypothesis" auxiliary hypothesis *ad hoc*, but a theory "limping along under the weight of its auxiliary hypotheses" in Gopnik & Wellman's own words, is "ugly and messy instead of beautiful and simple" (1992, p149).
1.7 The Representational theory of mind theory (Perner)

Perner (1988, 1991, in press), offers a CS theory that also reserves a special place for the notion of representation, but makes far more specific claims about the child's "representational theory of mind" (RTM) than the Gopnik-Wellman position. Perner's account also differs from Gopnik-Wellman with respect to the notion of theory change.

Perner's criticisms of Wellman's copy theory of belief (see Section 1.6.2) were motivated by his different views on the theory of mind that should be attributed to the 3-year-old. According to Perner, the theory that precedes the RTM is best characterized as a "situation theory" (Perner, 1988, 1991). More recently, he has introduced the notion of "prebelief" (Perner, in press), which is a notion that fits into the earlier theory, rather than replacing it (Perner, personal communication). I will focus on Perner's most recent exposition of his ideas (Perner, in press) in the outline that follows.

1.7.1 Situation theorists and the prebelief hypothesis

Perner argues that 3-year-olds are "Situation theorists" because they understand mental states as relations to semantically evaluable situations. They do not however understand that these relations are semantically evaluable. This distinction is shown in Table 1.2.

<table>
<thead>
<tr>
<th>Understanding</th>
<th>Equivalent to...</th>
<th>Situation theorist?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental state as relation to semantically evaluable situation</td>
<td>O --- attitude → (P), P is false</td>
<td>yes</td>
</tr>
<tr>
<td>Mental state as evaluating relations to semantically evaluable situations</td>
<td>O --- attitude → (P is true), P is false.</td>
<td>no</td>
</tr>
</tbody>
</table>
Semantic evaluation of propositions, in the situation theorist, takes place only outside the mental state term. By contrast, the representation theorist understands mental states as \textit{relations to semantically evaluable situations] that are themselves evaluated by the holder of the mental state}. This ability amounts to understanding i) that propositions are being evaluated and ii) that different people might evaluate one and the same proposition differently. Perner argues that situation theory, as represented in Table 1.2, is used to understand desire, and can account for performance on simple desire reasoning tasks (e.g. Wellman & Woolley, 1990; see also Section 1.6.1 and Chapter 5 [Section 5.4.2]; Yuill, 1984). Perner illustrates this as follows:

Unfulfilled desire: \hspace{1cm} O \rightarrow \text{wants} \rightarrow (P), \text{ and } P = \text{false} \\
Fulfilled desire: \hspace{1cm} O \rightarrow \text{wants} \rightarrow (P), \text{ and } P = \text{true}

Moving to belief understanding, he disagrees with the idea that the young child has \textit{no notion of belief} (a "simple desire psychology", Wellman & Woolley), on the strength of the evidence from pretend play (Leslie, 1987; see also Section 1.9). Children are capable of understanding that agents can act in ways that do not objectively satisfy their desires; that agents can act "as if" false propositions were true. Although Perner concedes that situation theorists can appreciate something about the role belief has in determining action, he attributes less significance to the ability to understand pretence than does Leslie. Specifically, he denies that these children distinguish the \textit{defining difference between pretend and belief}. In terms of the scheme above, the children cannot understand that actions according to a false proposition (\(P\)) based on \textit{believing that} \(P\) are different to actions according to the same false proposition based on \textit{pretending that} \(P\). This distinction requires that the child understand that holders of mental states evaluate those mental states and is represented in Table 1.3 below.

Perner supports this claim with evidence that, before the age of four, children do indeed find it hard to attribute correctly pretence or false belief to an agent acting as-if (\(P\)) (Perner, Baker & Hutton, 1994). In their experiment, two conditions were compared. The action was the same in each condition: feeding a rabbit that was no longer in a hutch. In one condition the agent knew that the rabbit had been moved and in the other he did not. In the
former situation, pretence should be attributed to the agent, in the latter, the agent should be described as really thinking that (P). Children younger than four years of age showed no sign of understanding this distinction, attributing pretending indiscriminately, while older children performed significantly better.

Table 1.3 Distinguishing pretence from mistaken actions based on false belief

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A --- believes → (This object (banana) is a telephone) implies:</td>
<td></td>
</tr>
<tr>
<td>A --- prelieves → (This object (banana) is a telephone),</td>
<td></td>
</tr>
<tr>
<td>A --- evaluates as true → (This object (banana) is a telephone)</td>
<td></td>
</tr>
<tr>
<td>but &quot;this object is a telephone&quot; is false.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>whereas...</td>
</tr>
<tr>
<td>2. A --- pretends → (This object (banana) is a telephone) implies:</td>
<td></td>
</tr>
<tr>
<td>A --- prelieves → (This object (banana) is a telephone),</td>
<td></td>
</tr>
<tr>
<td>A --- evaluates as false → (This object (banana) is a telephone)</td>
<td></td>
</tr>
<tr>
<td>and &quot;this object is a telephone&quot; is false.</td>
<td></td>
</tr>
</tbody>
</table>

Perner argues then, that children have an undifferentiated understanding of belief and pretence, and it is this notion that the belief tasks of Wellman (see Section 1.6.2) are tapping. Children are hypothesized to predict search in the real location of an item unless there is strong inducement to assume that the agent would act as if it were somewhere else. So for example with the Not Own versus Explicit False Belief:

Since the mere saying "but he thinks it's in the garage" is not strong enough [for children to assume the agent will act as if in the Explicit False Belief task] most 3-year olds opted for "porch". In the Not Own belief condition this obvious strategy is not available. There, only assumptions (prelieves) are available: the child prelieves it's under the porch, Sam prelieves it's in the garage. Forced to say where Sam will look, the most plausible option is where he prelieves it is, and most children indeed opted for this plausible option. (Perner, in press).
Thus, children will use the "prelief" strategy when they are induced into assuming the agent will (or is) act(ing) as if. In Chapter 4.1 I will outline Perner's use of this strategy to explain backwards reasoning tasks, where the agent is shown to be acting as if, and the children are asked to attribute the belief (Bartsch & Wellman, 1989, Experiment 2; see also Chapter 4, Section 4.2.1).

I now turn to Perner's views on representation within his RTM view of the child's theory of mind.

1.7.2 Belief and Representation

Children of about 4 years understand the need for semantic evaluation and can represent conflicting truth values and this understanding does not develop in isolation but is tied to children's acquisition of the concept of representation. (Perner, in press).

Perner's claim is that children who understand false belief have mastered the defining feature of representation: the distinction between representing (representing something) and representing as (representing something as being a certain way). This claim ties Perner to specific predictions about children's conceptual abilities derivable from other features of representation. For example:

To understand something as a representation is to assume that there is a concrete representational medium (vehicle) that bears the propositional content. [This] should lead to various changes in children's understanding of mental states and non-mental representations concurrent with their mastery of the false belief task. (Perner, in press, italics added).

For this reason Perner has recently been concerned with tasks of non-mental representation. The best known of these is the photos task (Zaitchik, 1990) described in Section 1.3.3. Perner and some colleagues have refined their analysis of this task. Originally, the task was considered to be good evidence for the general nature of the competence shift. Perner et al. (in press) however, reported low correlations between performance on photo and belief tasks, despite a similar level of overall performance, indicating that the tasks were not passed and failed by the same children. Perner suggests that photos are crucially different from beliefs because they cannot be false in the same way. He argues that when the situation (S) is changed at t, the picture taken at t, does not misrepresent St, it simply becomes an accurate but out-of-date representation of St. False photos do not require understanding of misrepresentation then, and Perner et al. (1994) went on to demonstrate
that manipulations that improve performance on the photo task do not help in the belief task (e.g. drawing the child's attention to the back of the photo versus drawing the child's attention to the agent's head). Photos are understood differently to beliefs then, according to Perner.

Are there tasks then, that can truly be described as misrepresentation tasks? Parkin & Perner (1993) have assessed children's performance on a "misleading signs" task. They argued that a sign is different from a photo because when the situation changes, the sign now misrepresents the new situation. Their task was similar to the photo task; a sign pointing to an object at \( t_0 \), which then moved to a new location at \( t_1 \). Children were asked where the sign showed the object to be. Performance on this task, a photo task (Zaitchik, 1990) and a false belief task (Wimmer & Perner, 1983 [direct belief attribution question]) were compared. The results showed that children performed poorly at the signs task, and that performance correlated strongly with performance on the a false belief task. Performance on the photo task, though at a similar level as in the other two tasks, did not correlate with performance on either.

Perner makes two other "representation-based" predictions; specifically, children should recognize that representations have i) *etiological* and ii) *formal* properties that are not shared with what they represent. As far as etiology is concerned, Perner cites evidence concerning children's understanding of the causal origins of mental states (Hogrefe *et al.*, 1986), which emerges at around age four (but see also Pratt & Bryant, 1990). These factors have been stressed in Wimmer's informational access theory (Wimmer, Hogrefe & Sodian, 1988; Wimmer & Weichbold, 1994) which I deal with in the section on CC theories, and in Chapters 3 and 6. For the formal properties, Doherty (1993) reported a relationship between children's false belief understanding and their awareness that words can express the same meaning, but have different formal properties (e.g. synonyms or different phrases that can express the same meaning).

1.7.3 Theory change in the RTM theory

Perner's position on theory change differs somewhat from that of Wellman-Gopnik. In his 1991 book (Chapter 10), he outlined his views on theory change in the child's developing commonsense psychology. He drew on an analogy between the theory status of
knowledge about tuberculosis possessed by a medical practitioner as compared to the status of that possessed by a bacteriologist. He argued that the pattern of development described in his theory can be viewed as "theory change": The child moves from a (mentalistic) theory of behaviour to a "theory of mind". This step is likened to crossing the gap from having a theory of symptoms to a theory of disease.

Perner explicitly ruled out the possibility that the change is one of replacement (e.g. Kepler's laws of planetary motion replacing Ptolemaic astronomical calculations (Hanson, 1958), which contrasts sharply with Gopnik & Wellman (1992; see Section 1.6.4). In fact, he argued that the representational theory of mind is only ever necessary in some limited circumstances. For the most part, situation theory will suffice (Perner, 1991, Chapter 5).

Perner introduced a different "history of science" analogy: that of theory extension (e.g. Kitcher, 1984). The representational view does not supplant the situation theory but rather extends it, allowing amendments for particular problems. He cited the problems of mutual knowledge (Lewis, 1969) and understanding word meanings (Grice, 1957) as examples of cases in which being a situation theorist might be preferable to taking a representational view.

1.8 Competence continuity theory theories

So much for the "consensus" view of development of commonsense psychological knowledge. I now move on to outline four CC theories, and also one theory that looks a little like a CC theory but does not fall neatly into the category. The majority of these outlines will be brief, and I shall concentrate on one specific CC theory: that of Leslie & colleagues (e.g. Leslie, 1987, 1988, 1994, in press; Leslie & Frith, 1988, 1990; Leslie & Roth 1993; Leslie & German, in press; Baron-Cohen, Leslie & Frith, 1985, 1986; Roth & Leslie. 1991, 1994; Surian & Leslie. 1994). This is because other CC theories, namely those of Wimmer, Mitchell and Fodor, in fact have much in common with Leslie's account; in fact Fodor's theory is left entirely until Chapter 3. Russell's most recent theory (Hughes & Russell, 1993) raises the notion of rejecting "theory of mind competence" itself. Despite this, however, there are also close similarities between Russell's position and that of Leslie, which I shall outline in Section 1.9.3, as well as critically discussing the notion of rejecting theory of mind competence.
1.8.1 Informational access

Wimmer (Wimmer et al., 1988; Wimmer & Weichbold, 1994) proposed a theory in terms of the child's increasing understanding of the informational access conditions that cause beliefs. Wimmer viewed this idea as a form of competence shift, but I classify Wimmer's theory as a CC theory (rather than a CS theory) for the following reason: having specific knowledge about the conditions that give rise to false beliefs is not the same thing as a concept of belief the logic of which allows beliefs to be false. For Wimmer, the child is learning about belief, but this must presuppose that the child has some way of knowing what (s)he is learning about. Learning specific things about beliefs over time may eventuate in a "concept" (or perhaps a category; see Rosch, 1978) of false belief type situations (e.g. a "trick" category), but before this learning can begin, some way of attending to the relevant materials must be in place. Learning about the role that information plays in causing beliefs is plausibly something that goes on continuously rather than something that occurs suddenly, hence conceptual continuity along with performance.

1.8.2 Reality masking (Mitchell; Robinson & Mitchell)

Mitchell (1994; Mitchell & Lacohée, 1991; Robinson & Mitchell, 1992) suggests that children's early theory of mind competence is attenuated, or masked by their focus on the real world. He ascribes early competence to the child, and argues that:

"Development takes the form of children gradually coming to attach more weight to a representational criterion rather than a realist on when judging about belief. (1994, p19)."

The root of this bias, according to the reality masking view, lies in the adaptive value that might be attached to being preoccupied with reality in early childhood. The "dormant" theory of mind competence, coupled with the gradual change, clearly makes this a CC theory-theory. In general terms, the view that (at least one of) the child's problem(s) in false belief task is to do with disengaging from reality is central to many other CC theories (e.g. Russell; Leslie). Russell et al., 1991 are in fact credited in Mitchell (1994), as the authors of one of the "primitive" versions of the reality-masking hypothesis. The fact that the CC theories have so much in common makes-their treatment here relatively simple; the differences are largely a matter of emphasis. In Chapter 3 for example, I present an experiment that addresses the "salience of reality", in the specific context of Fodor's theory
of the child's theory of mind. However, the results of the experiment are also relevant to the
reality masking hypothesis, as well as to Leslie's ToMM theory to be detailed in the next
section. Mitchell stresses the adaptive advantage that an early reality-bias might provide,
while Fodor & Leslie, as we shall soon see, emphasize how the fact that beliefs are usually
ture affects children's belief reasoning, each in a different way. In the course of this thesis,
I will be paying special attention to the CC theories of Leslie and Fodor (see e.g. Chapters
3, 5 and 6). I will return briefly to reality masking, however, in the introduction to
backwards reasoning and explanation, in Chapter 4.

1.9 The Theory of Mind Mechanism (ToMM) theory (Leslie)
This theory is definitely the best specified among CC theories, and arguably among
theory-theories as a whole. Leslie traces theory of mind competence back to the emergence
of the joint capacities to produce and understand pretend play in 18-24 month old children
(Leslie. 1987, 1988, 1994: Harris & Kavanagh. in press). Pretending is highly significant
for Leslie, because it demonstrates the computational competence required for theory of
mind. The "competence" is therefore in place no later than age 24 months, and development
consists in the ability to deploy this competence, rather than in the competence itself. I will
proceed here in a similar manner to that adopted for the CS theories. I start with the earliest
phenomena, and describe in detail the nature of the competence itself: a domain specific
processing mechanism called the Theory of Mind Mechanism (ToMM). From this I move
to performance concerns, and introduce a second mechanism, the Selection Processor (SP),
which deals with the general problem solving demands that theory of mind tasks present.
Leslie's position also relies heavily on experiments conducted with autistic individuals, and
I will spend some time describing these data.

1.9.1. Pretend play and the metarepresentation
Leslie (1987, 1988, 1994) argued that a cognitive account of pretence must address
four critical features of pretend play - that there are three fundamental forms, that these
forms emerge as a package, that solitary pretence is yoked to understanding pretence in
others, and that pretend content is anchored in reality (I will not go into great detail here,
see Leslie, 1987 for a detailed account).
Leslie proposed that an innate, domain-specific modular mechanism (ToMM), that employs a special type of representation provides just such an account. The representation that this mechanism computes is what Leslie referred to as the "metarepresentation". The metarepresentation specifies four types of information to arrive at agent centred descriptions of events: an agent, an attitude (informational relation), an anchor to reality and a proposition expressing a state of affairs. For example:

Billy PRETENDS (that) "it is a lobster" (is true of) this piece of Plasticine.

Leslie went on to point out that this same structure allows different attitudes to fill the attitude slot. The attitude of believing, for example, might be substituted for pretending in the above example.

ToMM deploys basic attitude concepts such as believing, desiring and pretending within the metarepresentation. The ToMM module is functional from late in the second year of life, as evidenced by the emergence of pretend play. This then is the competence in the ToMM theory of theory of mind.

1.9.2 Performance considerations: the Selection Processor (SP)

An immediate question for the ToMM account is the relation between the concepts of pretending and believing. Given that the same representational competence is implicated in deploying both attitude concepts, why is there a two year delay in the ability to understand one of them? Roth (1993; Leslie & Roth, 1993; see also Leslie & Thaiss, 1992 and Leslie 1994) suggested that tasks which make general processing demands (i.e. not specific to theory of mind tasks) can reveal limits in the modular ToMM. The standard false belief task (Wimmer & Perner, 1983) for example, despite having been proposed as a diagnostic test of theory of mind (see Section 1.1.2), makes other demands on the child. In order to calculate the correct content for the agent's belief, the child has to select the correct situation from memory (St, -object in A) to use as the inference in the face of salient

Leslie borrowed this term from Pylyshyn (1978), who used it to mean representation of a representational relation. Pemer criticizes Leslie's use of the term (e.g. Pemer, 1991). The issue turns on how the representational relation is represented, but Leslie has since adopted the term "M-representation" to distinguish the output of ToMM from metarepresentation as construed by Pemer (see Leslie, 1994).
competing information (St, object in B). This process is akin to "executive-type" functions, and Leslie & Thaiss (1992) posited a mechanism that dealt with the selection and inhibition demands of the type found in false belief tasks. This mechanism was termed the "Selection Processor" (SP). SP gradually increases in function during the preschool period, and its increasing efficiency accounts for the various findings on false belief and other theory of mind tasks.

Many of the theory of mind tasks described in this review share the same structure identified earlier (e.g. false-belief, deceptive box and action versions, appearance-reality, photos etc.). The ToMM-SP account explains failure on all of these in terms of the problem solving demands that this structure places on SP.

Children pass versions of false belief tasks that place less stress on SP. Tasks do this either by drawing attention to the relevant basis for selection (e.g. Mitchell & Lacohée, 1991), or else attenuating the prepotent response (Wellman & Bartsch, 1988; Roth & Leslie, 1991). The results of Zaitchik (1991) are a nice example of this latter possibility. She compared a "seen" condition where the subject sees the object's true location, with an "unseen" condition, where the subject is only told where the object really is. Children are significantly better in the "unseen" condition, which in terms of SP, is the condition where the situation at t, is less salient, and therefore more easily inhibited as the basis for the belief calculation.

Roth & Leslie (1994, Experiment 1) extend this analysis. They compared a standard version of the false belief task with a "partial true belief" task (Leslie & Frith, 1988). The structure of these two tasks was held constant, which allowed the assessment of falseness of the belief itself, a conceptual factor. In the partial true belief task the actor hides a bait object as usual (Location A), but when he leaves the room another bait object is placed in Location B. The child is asked where the actor will look for an object on returning. The results showed no difference in difficulty between the two tasks, indicating that conceptual factors alone cannot be the source of difficulty for the 3-year-old. Note that the logic of this experiment is exactly that of Wellman & Bartsch's "Inferred Belief Control" task (1988). Wellman & Bartsch found 3-year-olds could pass this task. The conceptual content is the

---

10 The notion that "executive function" is important in these tasks is similar to Russell's theory. As will become clear in the next section however, Russell wishes to make executive functioning where "all the action" is.
same in the two tasks, but in the Wellman Task, the children are only told what the actor saw and did not see. The selection and inhibition problems are reduced relative to Roth & Leslie's study, and this is consistent with less load on SP.

One crucial requirement one might place on Roth & Leslie's performance model is to substantiate the claim that the processing resources are general. In a second experiment, Roth & Leslie (1994, Experiment 2) do just this. In their "Screen" task, they attempted to stress SP outside the context of mental states. The child was presented with a toy basket and a toy box, and a marble is placed in the basket. A screen was placed between the child and these items, and an identical box and basket and marble were placed on the child's side of the screen. The (new) marble was placed again in the (new) basket, and then moved to the (new) box. The children were asked where the marble was, behind the screen. Here there are the same structural elements of theory of mind tasks (inference to be based on St^ in the face of a confusing alternative St^), but without the conceptual content (no belief, no representation).

They compared this task with both a standard false belief task and a further "non-standard" false belief task: the "Search" task. The Search task was designed to attenuate the stress on SP by changing certain structural features. Two characters had the same desire (for biscuits). One searched in one place and found biscuits, while the other searched in another place and failed to find biscuits. The contrast between the behaviours, and the desire-frustrating search cue were predicted to help the calculation of the (false) content of the belief. The results showed that 4-year-olds performed well at all three tasks (about 70%). For 3-year-olds, the performance was significantly better on the Search task than it was on the Screen task.

By way of interim summary: I have outlined Leslie's arguments that 3-year-olds' underlying competence with belief is not necessarily tapped by the tasks designated as

---

11 The notion that reasoning backwards from the action to the belief might provide support for young children is not new (e.g. Bartsch & Wellman, 1989; Moses & Flavell, 1990; Robinson & Mitchell, in press). A similar argument has been made for utterances (Roth & Leslie, 1991). Since this reasoning is also essentially the rationale behind explanation studies, I will leave discussion of this evidence until later chapters (e.g. Chapters 4 & 7).

12 The 3-year-old children performed unusually badly on the false belief task (less than 10%). Performance on the Screen task was better than this (about 35%), but the difference was not significant.
diagnostic of theory of mind. Instead, these diagnostic tasks tap general resources, for which there is independent evidence. Leslie (1994) also notes:

This performance squeeze gradually relaxes over the course of the fourth year, and probably beyond (1994, p235).

The notion that performance factors play a role after the period critical to competence shift theorists is very important. I will develop this idea in the chapters concerned with the explanation of action (e.g. Chapters 5 and 6). I turn now to the domain-specificity of ToMM, and the evidence that can be brought to bear on this issue.

1.9.3 The domain specificity of ToMM

Baron-Cohen et al., (1985), showed that children and adolescents with autism show a marked impairment in their ability to pass the false belief prediction task (Wimmer & Perner, 1983). For example, autistic adolescents with verbal mental ages in excess of eight years typically perform like normal 3-year-olds. This finding has proved to be of enormous significance in the theory of mind literature (see also Baron-Cohen, 1991a; Baron-Cohen et al., 1986; Leslie & Frith, 1988; Leslie & Roth, 1993; Perner, Frith, Leslie & Leekham, 1991).

The issue raised by this is whether the surface similarity between the performance of autistic children and that of normal 3-year-olds suggests that autistic children are "stuck" with a 3-year-old theory of mind. Leslie & colleagues argue that the behavioural similarity between 3-year-olds and autistic children is misleading, and that the 3-year-old child knows more about belief than the typical able autistic. Leslie's metarepresentational theory (1987, 1988) argued that autistic children fail theory of mind tasks because they lack the conceptual basis for understanding mental states: the ability to compute metarepresentations. Strong evidence in favour of this view is that autistic children show less spontaneous pretend play than normal children (Baron-Cohen, 1987; Sigman & Ungerer, 1984). The autistic child's theory of mind is not "stuck" according to this view, but has been abnormal all along. The autistic child is hypothesized to have a damaged or absent ToMM, and therefore lack the capacity to acquire a theory of mind.

The most striking evidence in favour of this view has been provided recently (Charman & Baron-Cohen, 1992; Leekham & Perner, 1991; Leslie & Thaiss, 1992). Leslie
& Thaiss, for example, compared the performance of 4-year-olds and autistics on a standard action-based false belief task (Baron-Cohen et al., 1985) and the "false" photographs task (Zaitchik, 1990). The results showed a remarkable group by task interaction: 4-year-old children were better at understanding beliefs than photos, while autistics were at near ceiling on the photos, but performed very poorly on the belief task.

The photo task, as argued above, shares the same structure as belief tasks, and consequently places similar demands for any problem solving mechanism. The fact that autistics pass the photo task with such ease suggests that the problem solving demands cannot explain their failure on false belief tasks. The ToMM-SP model sees autistic children as the "mirror-images" of 3-year olds. SP is intact, as evidenced by their success with photos, while ToMM is impaired. For 3-year-olds an immature SP limits the deployment of an intact ToMM.

This evidence is extended in Roth & Leslie (1994, Experiment 2; see Section 1.9.2). A group of autistic children were also included in their experiment, and showed a pattern of performance unlike that of either the 3- or 4-year-old group. They were virtually at ceiling on the Screen task (testing task structure demands), but considerably worse on the Search and False belief tasks (with theory of mind conceptual load).

1.9.4 Executive functioning or ToMM?

As noted previously, an executive-type mechanism is implicated in false belief tasks sharing certain features (see Section 1.9.2). Some theorists have viewed this type of finding as the basis for speculation that theory of mind as a whole might be some special species of executive function, or perhaps depend on intact executive functioning and that autism results from a deficit in executive functioning (e.g. Hughes & Russell, 1993; Ozonoff, Pennington & Rogers, 1991).

Russell and his colleagues have argued that young children and autistics show perseverative behaviour on the "windows" deception task (Russell, Mauthner, Sharpe & Tidswell, 1991). In the windows task, children sit opposite an "opponent" and are presented with two opaque boxes. While the child is not looking, a smartie is placed in one of the

---

13 The main effect was not significant, but overall, five experiments have found performance on photos to be lower than on beliefs in normal children.
boxes, and the child is invited to select (by pointing for example). The children are trained in the following way; whichever box they choose is opened by their opponent, and they keep the contents of the other box, that is they learn the rule; to win the smartie point to the empty box. Once they have learned this rule they are presented with boxes with windows in. They can see the contents but their opponent cannot. They are required to "deceive" their opponent into opening the empty box. Russell et al. showed that 3-year-olds and autistics perseverate in choosing the wrong box. Not only do they fail on the first trial, but continue to do so for up to 20 trials, losing each smartie. This failure to "disengage" from the prepotent "true" location is the executive type facility that Russell et al. argue is impaired in autism, a speculation they support with a further experiment that shows that the presence of an opponent is not necessary to get the effect (Russell, Jarrold & Potel, in press).

Despite these results that show there to be executive problems in autistic children, and the finding that autistic children also seem to be impaired in certain other kinds of "executive functioning" (Ozonoff, Pennington & Rogers, 1991), the photograph and screen tasks described in Section 1.9.3 provide an extremely close problem solving control for false belief tasks. The fact that autistic individuals were at near ceiling on these tasks suggests that executive problems cannot be the reason they fail false belief tasks. To collapse theory of mind into executive functioning would be premature.

1.10 Overview of next chapters

This completes the overview of theories and evidence in theory of mind. During the course of the next chapters I will expand certain parts of the review and cover them in more detail. In the next chapter I deal with the suggestion that theory of mind is not in fact a theory. Having done that I return to CC theories of mind, and present an empirical investigation of a CC theory I have not outlined here: that of Fodor (1992). This investigation leads me into the study of explanation, which is the topic of the remaining chapters. In Chapter 4, I present detailed methodology and results from explanation experiments. In Chapters 5, 6 and 7, I present my own investigations into the explanation of action. In Chapter 8, I round everything off with some conclusions.
CHAPTER 2

Simulation and theory: Knowledge, ability and information processing

2.1 Introduction

Chapter 1 presented an overview of the theory of mind literature, and an appraisal of one class of theories of the acquisition of commonsense psychology. This chapter concerns the challenge to theory-theory from philosophers and psychologists who suggest that our ability to predict and explain the behaviour of others is subserved, not by a theory, but by a process of "simulation" (e.g. Goldman, 1992, 1993a; Gordon, 1986, 1992; Harris, 1991, 1992). In the course of the chapter I express scepticism about what I identify as radical simulation-theory, and endorse theory-theory in general terms. Parts of this argument appear in Leslie & German (in press).

The challenge that simulation-theory poses to theory-theory has been stressed as one of great philosophical importance. A number of fundamental assumptions in cognitive science, as well as certain questions in the philosophy of mind, are held to be at stake. Stich & Nichols (1992) outlined the wide range of cognitive capacities for which the dominant explanatory strategy has been to posit an internal "knowledge structure", or theory, which guides the agent's abilities in the domain in question. For example, there is a rich literature on the child's capacity to acquire a natural language (Chomsky, 1965; Pinker, 1989), ability to recognize objects visually (Gregory, 1970), and the prediction of the behaviour of objects (McClosky, 1983), along with a host of others. Stich & Nichols pointed out that if this dominant explanatory strategy were to turn out to be wrong in one of the most popular areas of its application (commonsense psychology), it would amount to an important challenge to the employment of the strategy elsewhere.

A further consequence identified by Stich & Nichols concerns a central debate in the philosophy of mind, that between eliminativism and its opponents. They argued that the emergence of simulation theory threatens the elimination of eliminativism itself. To illustrate what they mean I will describe briefly the issue at stake. Ramsey, Stich and Garon (1991)
describe eliminativism as the thesis that certain entities or properties postulated by a common-sense or scientific account of the world do not exist. Examples of the victims of eliminativism include witches in the domain of folk explanation, and calorific and phlogiston in the domain of science. The successor theories had no role for these entities and so they were eliminated.

Here I refer to eliminativism in the domain of commonsense psychology. Commonsense psychology is held, by eliminativists, to be a radically false theory. It is held to be false in the same way that explaining phenomena such as illness and abnormal behaviour by means of a framework that recognizes entities such as demons and witches is now held to be radically false. At stake is the existence of the propositional attitudes themselves. If folk psychology is radically false then beliefs and desires, wishes and intentions do not exist. Stich & Nichols argued that if simulation theory turns out to be right then it provides a problem for eliminativists. Their argument will have been effectively trumped by proponents of simulation theory, who state that commonsense psychology is not a theory at all. If something is not a member of a class then it cannot be a radically false member of that class. According to Stich & Nichols then, versions of simulation theory that deny the existence of theory of mind knowledge have this extra philosophical consequence.

The concise argument to be presented here has two parts. The first part identifies the fundamental issue that separates radical simulation theory from theory-theory. Radical simulation is argued to be rendered implausible by analogy with language. This section also suggests that not all proponents of simulation theory seem to be defending such a radical version, and that weaker versions of simulation theory can be formulated which, although interesting, are actually just versions of theory-theory. The second part of the argument looks in more detail at these weaker versions of simulation theory and proposes that some of their apparent advantages over versions of theory-theory are misleading. The importance of specifying theories in terms of their information processing requirements is stressed.

2.2 Radical simulation: "The big issue"

According to Stich & Nichols (1992, in press), the fundamental issue separating theory-theory and simulation theory is the issue of cognitive penetrability (Pylyshyn, 1984). A process being cognitively penetrable amounts to its being influenced by knowledge (or
representations) in a sequence of inference. If a process is so influenced then it can be said to be cognitively penetrable and if not, then it is cognitively impenetrable. Theory-theory accounts of commonsense psychology assume that there is a system of knowledge implicated in the processes of understanding the minds and behaviour of others, that is they assume that commonsense psychology is cognitively penetrable. Simulation accounts, by contrast, assume that commonsense psychology is cognitively impenetrable. Leslie & German (1994) suggest that another way of characterizing this divide is in terms of "knowledge and ability". The debate, according to this view, amounts to whether both knowledge and ability or simply ability alone can account for the acquisition and deployment of commonsense psychology.

The simulationists' challenge amounts to denying that commonsense psychology needs knowledge or representation. Their argument is that people can understand the minds of others without the use of a knowledge database by exploiting the fact that they are people who emit behaviour themselves. For example, Person A, by running the system that generates their own behaviour "off-line" (so that real actions do not result) and using "pretend" inputs for the Person B, can use the off-line output of their action decision system (ADS) as an action prediction for Person B. In this way, simulationists can dispense with the idea that commonsense psychology needs knowledge or representation, and assume instead that ability only is required: the ability to run the ADS with off-line inputs. If the acquisition and deployment of commonsense psychology can be shown to require only "simulating", using "pretend" inputs to the ADS, then theory of mind knowledge would be shown to be just ability. This is a stronger, and very different claim than that made by theory-theory, because theory-theory assumes that abilities are required to deploy theory of mind knowledge. Simulation could be one of the abilities associated with theory of mind knowledge.

Leslie & German (1994) noted that the strategy of trying to reduce knowledge to ability is not new, and that previous attempts to do so in the case of language were unsuccessful (Chomsky, 1965, 1975, 1988). Chomsky (1988), for example, pointed out that language knowledge and language ability cannot simply be equated. It is possible both for ability to be impaired with no loss of knowledge and vice versa. In Chomsky's example, Juan suffers a head injury and loses the ability to speak and understand Spanish. However, this
does not mean he has lost all knowledge of Spanish, because he may recover his ability later without having to repeat the initial process by which he first acquired the language. In the converse case, Juan's ability might improve with no gain in knowledge; Juan might access his existing knowledge more efficiently and express it in a more polished performance.

One of the things that makes it difficult to reduce language to ability is the evidence that languages involve structured knowledge (grammar), as well as the ability to deploy this knowledge. In a similar vein, commonsense psychology may have a system of representation that is specialized for representing an ontology, principles, rules and relations of the domain. To claim that commonsense psychology is just ability would be to deny a representational system of this kind.

It is revealing that many versions of simulation theory appear not to defend such a radical position. For example, Goldman (1993b), in a response to commentaries on his defence of simulation theory by both Perner (1993) and Leslie, German & Happé (1993), appealed to a simulation process that recognized the attribution of mental states. Consequently, his position assumes a knowledge and ability mix. In arguing that the action planning system might be able to simulate recursively (see Section 2.3.1 for a detailed account of this issue), Goldman (1993b) makes use of representations of propositional attitudes, which provide inputs to, and act as outputs from, a simulation process:

[T]o simulate Mary [who believes that John believes that \( p \), one will] generate some initial beliefs she would have about John. I put myself in Mary's shoes of agreeing with John that he will put away the chocolate. I feed an awareness of this agreement into my Mary simulation and allow an inferential process to operate on it. This inferential process outputs the conclusion that John will put the chocolate in some spot \( x \) and remember which spot it is. So I ascribe this belief to Mary. (1993, p107).

Similarly, the position adopted by Harris (1992) is also a retreat from an extreme simulation position, suggesting that the notion that commonsense psychology proceeds via simulation need not result in denial that propositional attitudes are attributed. In Harris' words:

In the version of Simulation Theory that I try to develop below, I shall argue that the child does attribute propositional attitudes but not by virtue of subscribing to a theory. Of course, one might wish to claim that a simulation process allows (the subject) to dispense not just with a theory of the attitudes but with the attitudes themselves, but in my view such a position is untenable. (1992, p121. italics added).
Thus, Harris' position, with its mixture of knowledge and ability, is a version of theory-theory. Although now not a radical challenge to theory-theory, this weaker version of simulation still claims that commonsense psychology is mostly ability, and that the ability is simulation. Furthermore, Harris' position incorporates the claim that where propositional attitudes are attributed within simulation theory, their understanding is based on simpler understanding of simulation. I will return to this latter point in the next section.

It is important to note here that the term simulation can be applied liberally or conservatively, and versions of theory-theory that incorporate some "simulation" ability are common or rare according to how widely one wants to apply the term. For example, Leslie (1987) in his treatment of children's capacity to pretend, described the way that the use of one set of inferences can be applied to other situations. Another possible instance of a simulation ability within theory-of-mind is described and discussed in Chapter 7. Briefly, the suggestion there is that children might sometimes make use of their own (recent) experience in order to help calculate the content of the mental states of other people. It does not seem unreasonable to suppose that one way that I might go about calculating what move my opponent will make in a game, is to imagine myself in his or her position, with his or her goals, and decide what I would do. This simulation process depends on the attribution of theoretical constructs, it does not replace theory in this case, but can be regarded as an ability that facilitates accurate deployment of the theory. If children could be shown to have made use of their own recent experience in calculating a prediction of another's action, then this might support the idea that some type of simulation ability was available to them. This is the suggestion that is discussed in Chapter 7. The important point for now, however, is that none of these instances are examples of knowledge-free simulation, which would be required for simulation to pose a radical challenge to theory-theory.

One further consideration relevant here is Fodor's claim that the ADS itself exploits a theory, or access to ToM knowledge. Stich & Nichols (1992) address this possibility. They state that if the simulation claims about commonsense psychology being reducible to off-line operation of the ADS and Fodor's claims about the ADS itself employing a theory both turn out to be correct, then they will concede simulation theory to be true, although they recognize that "it is a bit odd to draw the battle lines this way" (1992, footnote 7). Stich & Nichols go on to suggest that they can defeat simulation claims that off-line

50
operation of the ADS is used for commonsense psychology, *however the ADS works*. One problem with Stich & Nichols' concession is that a world where commonsense psychology is a matter of off-line simulating with an ADS that is knowledge-based, is a world where knowledge *and* ability are required for commonsense psychology. The concession turns the debate into an issue of *where* ToM knowledge might be located in the information processing system, rather than *whether* there is ToM knowledge stored at all. Leslie & German (in press) argue, in contrast to Stich & Nichols, that if Fodor's ADS claim and simulation theory's ADS claim both turn out to be right, then the result will be a version of theory-theory.

I turn now to the less than radical (theory-theory compatible) versions of simulation theory.

2.3 Is simulation simple?

The less than radical versions of theory-theory which view commonsense psychology as mostly simulation ability, all place the "action decision system" (ADS) in a central role (e.g. Harris 1992). Despite this, none of the theories really specify this system in enough detail to support their case. One of the main advantages that simulation-theory has been claimed to have over theory-theory is based on considerations of simplicity or parsimony, and it is this claim that I focus on here.

Theory-theories posit an extensive knowledge database that provides the representational wherewithal and the principles and rules of its application. Given that the individual itself engages in behaviour, arguing that the individual exploits *the system that plans its own behaviour* to understand the behaviour of others, allows the simulationists to claim a huge reduction in complexity. The ADS needs to be posited in any account of behavioural control, so if it is used for understanding the behaviour of others as well, then in simulation-theory, the database that would otherwise be required by theory-theory *comes for free* (or, to put it another way, need not come at all).

This "simplicity" advantage claimed for simulation-theory over theory-theory by adherents of the former is misleading. Exploiting the existence of pre-required architecture in place of positing a cumbersome knowledge database looks attractive at first glance, but just how simple does "running the action planning system off-line" prove to be? I will outline
a number of hidden complexities in the capabilities required by an ADS in order to support an enterprise like this. It will be seen that the hidden complexities bring "simulation ability based" versions of theory-theory in line with other theory-theories in terms of their complexity. The simplicity of simulation versions of theory-theory is derived from the lack of specificity furnished in the theory in the first place.

The system is described as running "off-line" in order to capture the notion that real actions/behaviour are not produced. However, it is conceivable that people themselves have goals which, though real, are not to be acted upon immediately (or indeed ever), and these too are presumably "off-line". For example, my ADS might come up with the goal of learning to fly a plane, but this goal is certainly not issuing real behaviour now, and it may never do so. The system therefore needs to be able to keep track of, not only another persons real goals (off-line), but also my not-to-be-acted-upon (off-line) goals. With just one degree of freedom (on/off-line), there is insufficient detail to distinguish one's own off-line goals from the off-line goals of any person one might simulate.

Further complexity is added when one considers the case of pretence. According to simulation theory, the way to understand someone acting seriously is to "pretend" to be in their shoes. However, the implication of this is that in order to understand someone else pretending one needs to pretend to be pretending. In pretend play, one can pretend to pursue real goals or really pursue pretend goals: for example, I could pretend to turn over a cup that is really full of water, or really turn over a cup that I pretend is full of water. Once again, the on/off-line distinction alone cannot support this level of complexity. The natural solution for dealing with this complexity, the use of recursive representations, is blocked by the need for simulation to be ability only, or in the case of the mostly-ability versions of simulation theory (e.g. Harris 1992: see above section), the need to derive propositional attitude understanding from simpler simulation ability.

The implication of this is that the additional complexity has to be built into the mechanism that supports the ability. If the system needs to be employed off-line to simulate another's real goals, then it needs a further embedded ADS to handle another's pretend goals (off-off-line). This problem reemerges each time a level of recursion is required in commonsense psychological reasoning. Adults rarely have trouble with understanding at least triply embedded mental states (e.g. Arthur wanted to persuade Mary that the new
teacher wasn't expecting things to go swimmingly), which suggests that the ADS would need to come with double or even triple embedding at minimum. Moreover, this embedding would be required only for theory of mind work, not for the usual action planning work for which the system would be primarily responsible, which makes the claim that the theory of mind database posited by theory-theory comes for free in simulation theory less than credible.

Theory of mind calculations have been shown to require several degrees of freedom, which poses problems for any simple or parsimonious mechanism posited to deal with them. All the complexity that theory of mind knowledge provides must be handled in the mechanisms in simulation theory. Unfortunately, specific details of how the action planning system delivers this complexity are not provided by current formulations of simulation theory, and thus the claim that the simulation theory has a "simplicity" advantage over theory-theory is clearly one to be evaluated in the context of the relative specification of detail in the two classes of theory.

2.4 Conclusions

Two strands of argument have been presented in an attempt to rebut the challenge posed to theory-theories of commonsense psychology by the notion of simulation. The first strand stressed the fundamental issue at stake between radical simulation theory and theory-theory: the possibility that theory of mind knowledge might be reducible to ability only. Adopting this line of attack would constitute a major challenge to theory-theory, but it turns out that there are few proponents who do so, and that such an attack is rendered implausible when one considers the analogous case in language. Weaker versions of simulation theory that propose a mostly-ability mix were demonstrated to be lacking considerable detail in their specification. On close examination, many of the advantages claimed for weak simulation over theory-theory were seen to stem from under-specification. The information-processing demands of theory of mind that are made explicit in (at least some) versions of theory-theory are largely ignored in simulation theory.

The foregoing discussion emphasizes the desirability of outlining all models in terms of their information processing demands. As noted in Chapter 1, information processing considerations are also overlooked in some versions of theory-theory. Indeed, information
processing specification in a theory-theory is considered by some proponents of CS theory-theory to be a challenge to theory-theory (Gopnik & Wellman, 1992; Gopnik, 1993a; but see also Leslie, German & Happé, 1993; Leslie & German, in press). The theory-theories that come closest to specifying adequately information processing demands are those that consider conceptual development to be a process involving both competence and performance.

In Chapters 5, 6 and 7, the effect of increasing performance resources on the deployment of theory of mind in young children's explanations is examined. Before that, in Chapter 3, Fodor's (1992) competence continuity theory is tested empirically.
CHAPTER 3

Testing Fodor's theory of the child's theory of mind

3.1 Introduction

Chapter 1 introduced a number of theories of the development of folk psychology and classified them according to whether they were competence-shift (CS) theory theories or competence-continuity (CC) theory theories. Theories were divided on the basis of how they explained the shift in performance on certain belief reasoning tasks, particularly belief attribution tasks such as that used by Wimmer & Perner (1983). In that task, recall, an agent hid a bait item in a location (say container A) and left the room. While the agent was away, another character moved the item to a new location (container B). The task was for the children to predict where the first agent would look for the item on her return: an indirect action-based belief attribution. It is a task of this form that will be used in the experiments to be presented here.

The present chapter takes a closer look at one of the CC theories, that of Fodor (1992), and reports two experiments that test the predictions of this theory. To recap briefly, the CC theorists endorse the view that the shift seen at age four in children's capacity to solve belief reasoning problems is a result, not of a shift in the theory of mind competence itself, but rather of increases in resources available to the child to deploy the theory efficiently.

3.2 Fodor's theory of the child's theory of mind

3.2.1 An outline of Fodor's theory

Fodor's (1992) theory proposes an ontology and covering laws of a primitive intentional psychology (a very simple theory of mind - VSTM) that is seen as being continuous with adult folk psychology. 14 The ontology of VSTM is simple in that it

---

14 This is true in that the ontology acknowledges no entities not also recognised by the adult folk psychology, and that the empirical generalizations are all true according to the adult intentional psychology.
recognizes the existence of only beliefs and desires while the fully fledged adult version can appreciate many other attitudes as well (e.g. hopes, lusts, fears, hunches,...etc). VSTM is different to the adult theory then in quantitative rather than qualitative terms. The covering laws of VSTM are:

i) In normal circumstances people act in a way that will satisfy their desires if their beliefs are true.

ii) In normal circumstances what people believe is true.

The 3-year-old child's theory of mind is thus proposed to be simpler than, although continuous with, the adult's commonsense psychology. According to Fodor, the 4-year-old also subscribes to VSTM. This account then, does not appeal to a competence difference in order to explain the difference in performance shown on false belief tasks by 3- and 4-year-olds. Fodor offers an account of this shift in terms of a performance mechanism that children exploit in order to solve action prediction problems using their VSTM. Fodor's suggestion is that the children solve such problems by applying the following heuristics.

**H1** *Predict that the agent will act in a way that will satisfy his desires.*

**H2** *Predict that the agent will act in a way that would satisfy his desires if his beliefs were true.*

Both 3- and 4-year-olds rely on the first heuristic for their action predictions as long as certain conditions hold. What explains the difference between the older and the younger children is the conditions that they will accept before using H2:

3-year-olds use H1 *whenever it affords a unique behavioural prediction*; they use H2 only when this uniqueness condition is not satisfied.

4-year-olds, by contrast, use H1 *whenever it yields a unique behavioural prediction that is consistent with what they know about the agents beliefs.*

Given the covering laws of VSTM, it is of note that the predictions made using H1
will generally be reliable. H2 covers a situation that is possible according to VSTM but that cannot be dealt with by using H1: the case where an agent acts in a way that would have satisfied her desires if only her beliefs were true. The move from using H1 to H2 is a move from a less complicated (but also less reliable) procedure for making action predictions, to one that is both more complicated and more reliable. If what is developing between the ages of three and four is something like "access to computational resources", then trading reliability for computational simplicity at the younger age is exactly the pattern of development that might be expected.

3.2.2 Explaining the data with the heuristic theory

Fodor stressed that the false belief task is peculiar in that it is the only belief reasoning task in use where the child following H1 (assuming that the agent will act in such a way as to satisfy his desires) results in a prediction from desires that is both unique and incorrect. The 3-year-old child need not consider what Maxi believes at all because the information from Maxi's desires yields the prediction "look in container B". The older child, by contrast, uses H1 only when the prediction it generates is compatible with what is known about the agent's beliefs, and in the case of the false belief task, it is not. The older child takes into account the fact that Maxi believes the bait is in A and derives her prediction by following H2. The resulting prediction (look in container A) is correct.

To support his contention, Fodor cited belief reasoning tasks that younger children do not appear to have difficulty with (e.g. Wellman & Bartsch 1988). In one of the tasks reported in that paper (Not Own Belief; see also Chapter 1, Table 1.1), the child was asked to guess where she thought the bait item was. The experimenter manipulates the story such that the agent was said to think that the bait was in the other location. Under these conditions, the child had no difficulty in predicting that the agent would act in accordance with his beliefs and look in a different place to where the child would look (see Chapter 1, Section 1.6.2, for more detail). The crucial difference in this task, according to the heuristic theory, was that the child did not know where the bait item was, and as a consequence the prediction derived from the agent's desire was not unique. This ambiguity was enough to force the child to use what (s)he knew about the agent's beliefs, which (s)he does successfully.
Another strand of evidence cited by Fodor was the data from Bartsch & Wellman (1989). Bartsch & Wellman presented children with the task of reasoning backward from an agent's mistaken action (search in A) to infer the agent's false belief (agent falsely believes bait is in A). This experiment has more recently been the subject of some criticism, and a more detailed appraisal of explanation methodology and the results of "backwards reasoning" manipulations appears in Chapter 4. Nevertheless, the results of Bartsch & Wellman (1989) can in the meantime be taken to indicate that children younger than four find reasoning backwards to a false belief easier than a similarly structured belief prediction task (1989, Experiment 2). Fodor argued that his heuristic theory could account for this; in the action explanation situation the child was told that the prediction from $H1$ fails. In such circumstances, $H2$ should be adopted. The action that satisfied this heuristic was the case where the agent believed (wrongly) that the bait was in A, so the child ascribed that belief to the agent.

In the following experiment I will present two tests of the central prediction made by Fodor's theory. I will postpone a more general appraisal of the theory until the results of these tests have been gathered (Section 3.10).

3.2.3 Experiment 3.1: introduction

The first experiment reported in this chapter tests Fodor's theory of the child's theory of mind. Two manipulations that Fodor (1992) suggested at the end of his paper are tested against a "standard" action-based false belief task.

Fodor's first suggestion was that in order to create a situation where the behavioural prediction that is generated by $H1$ is not unique, the bait item in the false belief scenario should be moved from the initial hiding place (Location A) and split between two new locations (Locations B and C). In these circumstances, the prediction from $H1$ would not be unique and this should result in $H2$ being consulted. $H2$ takes account of what the child knows about the agent's beliefs, and so the children should be able to make a belief-based prediction in this version of the task. The first manipulation in the current experiment is therefore to split the bait between two new locations.

Fodor's second suggestion was that if the bait item were to be destroyed after having been hidden in Location A, rather than moved to another location, then again there would
be no unique prediction from desires and children should consult H2. This experiment was in fact one of the experiments conducted by Wimmer & Perner in their original 1983 package. In their "Disappear" condition however, Maxi's chocolate, which in the story was grated into a cake, was in reality simply placed "behind the scenes". The performance of 3-to 4-year-old children was not considerably improved, but there was some benefit for 4 to 5-year-olds. Importantly however, the errors in the 3 to 4-year-old group in the disappear condition consisted almost entirely of children predicting that the agent would look "behind the scenes" for the chocolate. It would appear that Wimmer & Perner did not succeed in convincing the children that the chocolate had been destroyed. There is some possibility therefore, that destroying the chocolate more thoroughly might set up the non-uniqueness that Fodor predicts will be of help to the children. In the current experiment, the chocolate is indeed destroyed more thoroughly.

Fodor's heuristic hypothesis predicts that performance under the conditions of the two manipulations suggested above will be better than that on a standard task where the behavioural prediction from the agent's desires is both unique and wrong.

3.3. Experiment 1 method
3.3.1 Design

Two new versions of the false belief task were designed to test the hypothesis that a belief reasoning task with no unique behavioural prediction derivable from the agent's desire would result in better performance by three-year old children than that obtained using a standard task. In one task the ambiguity was created by the bait being transferred to two locations instead of one. Search at either of these locations would satisfy the agent's desire, and thus no unique prediction can be derived. In the other task, the bait was destroyed so that no action would satisfy the agent's desire. Again there is no clear prediction from the agent's desire in this case. Both new versions of the task were given to a single group of children along with a version of the standard false belief task.

The design was within-subjects, with each child being assessed on each task. The prediction made by the heuristic theory of performance on the false belief task was that both

---

15 Comparison with later experiments is not easy; in Wimmer & Perner (1983) the shift in performance occurs slightly later than in many more recent experiments (see Astington & Gopnik, 1991).
the new versions of the false belief task would result in better performance than the standard task. The order of presentation of tasks was counterbalanced across the group, with the same number of children receiving each of the six possible orders of the three tasks.

### 3.3.2 Subjects

Children were required to pass the control questions of at least two of the three tasks before being included in the study (see section on tasks below). Children were tested until 24 had met this criterion, with an equal number of children encountering each presentation order. The resulting group were aged between 39 months and 44 months (mean = 41.8, sd=1.7). Subjects were recruited from the subject pool at the Cognitive Development Unit in London, which is drawn from a predominantly white middle class population. All subjects received all three tasks, so no allocation to condition was required. Subjects were allocated randomly to one of the six task orders.

### 3.3.3 Tasks

Three tasks were used. Each task involved enaction of a story in front of the children using props. Each story involved a bait object, a main protagonist, a second protagonist and 3 containers (A, B and C). The stories were based loosely on the Sally/Anne task (Baron-Cohen et al., 1985).

The Standard false belief task in this experiment had the following content: a small boy (Billy) leaves a ball in Location A (a box) while he helps his father in the garden. While he is absent, his mother enters and moves the ball to Location B (a bag). Children were asked two control questions at this point: "Where was the ball right at the beginning?" and "Where is the ball now?". Finally, the test question was asked: "When Billy comes back from the garden, where will he look for the ball?".

There were two new tasks. The "Split Bait" task and the "Eat Bait" task. The structure of both these tasks was essentially the same as that described for the standard task. In the Split Bait task the bait object was a pile of beads. The protagonist (Sally) hid the beads in Location A (a basket) and her friend (Anne) moved some of them to Location B (a box) and some of them to Location C (a tin). The control and test questions were the same as for the standard task, and were asked in the same order.
In the "Eat Bait" task, the bait object was a small piece of chocolate. Again the bait was hidden in location A (a box), but for this story it was moved from Location A and given to the experimenter as a present. The experimenter said that he liked to eat chocolate and proceeded to eat the bait. In this way the bait was truly destroyed, rather than simply being placed "behind the scenes". After the chocolate had been consumed, the test questions were asked in the same order as before.

3.3.4 Procedure

Each child was tested individually in a quiet room in the babylab at the Cognitive Development Unit. The children were seated opposite the experimenter at a low table. The experimenter presented the child with the three tasks, one after another in one session. Each task took approximately three minutes to complete and the whole session lasted about twelve minutes. Each child received one of the six possible orders of the three tasks, with four children receiving each order. Each child was videotaped and the videotapes were scored later. During each task, when the two control questions were asked, the following procedure was employed. If on a given task a child failed to answer either control question accurately, then the story was repeated and the questions asked again. If a child still failed either control question in such circumstances then he or she was dropped from the study and replaced. Children who failed control questions (after repetition) for just one story were included in the study, but their answers to the test question for that story were not included in the analysis. For the "Eat Bait" task it is not really clear that there is a correct answer for the "where now" question. The children were not penalised for responding that they didn't know, but if they referred to any of the locations included in the story (A, B or C) then they were scored as incorrect.

3.4 Results

Three children were excluded from the analysis of the "Split Bait" task after having failed control questions. All children were included in the analysis of the other tasks. Children were scored correct if they predicted, by pointing or with a verbal response, that the character would look in the original location (A). Figure 3.1 shows the number of children in Experiment 3.1 who passed Standard. Split Bait and Eat Bait versions of the
false belief task.

Initial inspection of these results shows that children did not find any of these tasks particularly easy. The best performance appears to be the "Eat Bait" task, with performance of the other two tasks nearly at floor. Six children passed just one of the three tasks, and all passed the same one (Eat Bait). This pattern of performance was analysed using Cochran’s Q, for repeated measures on related samples, and was found to be significantly different from the pattern expected by chance \( Q = 12, p<0.002 \).\textsuperscript{16} This indicates that the Eat Bait

\textsuperscript{16} Siegel & Castellan (1988) note that when \( N<24 \) (\( N \) = the "effective" sample size: the number of cases where the values are not all passes or fails; \( k \) = number of measurements), the exact sampling distribution fails to approximate \( \chi^2 \). The exact distribution of \( Q \) for small \( N \) appears in Patil (1975), but for very small \( N \) it can be calculated from the permutation arguments. Here the probability can be calculated by analogy with a 3 sided coin
3.4.1 Effect of splitting the bait

Fodor's hypotheses relate to the performance of both new tasks as against performance on the Standard task. In the next analyses, the specific effect of each manipulation is assessed against the performance on the standard task. Table 3.1 shows the consistency of performance across the Split Bait and the Standard tasks. Children are scored into one of four categories: passing both tasks, failing both tasks, passing one but not the other, and passing the other but not the one. It is clear from the table that all the children are consistent, mostly failing both tasks. There is therefore no evidence that the children found the Split bait task any easier than the Standard task.

<table>
<thead>
<tr>
<th></th>
<th>Pass Split Bait task</th>
<th>Fail Split Bait task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass Standard task</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Fail Standard task</td>
<td>0</td>
<td>18</td>
</tr>
</tbody>
</table>

**N.b. N=21 owing to children being eliminated from analysis of split bait task.

For the Split Bait task, all errors were predictions of search in Locations B and C. Children responded by saying: "There and there..." and either pointed with the same hand to B then C (C then B) or pointed simultaneously at B and C, using both hands to do so.

3.4.2 Interim discussion

The Split Bait task does not reveal an improvement in children's performance relative to tossing 6 times. The probability of 6 heads is 1 divided by 6^6 = 0.0014.
to the standard task. This is contrary to the prediction made by Fodor (1992).

Wimmer & Weichbold (1994) reported results that are similar in many respects to those presented here, but that are also crucially different. Wimmer & Weichbold presented children with the Split Bait task, and also found that children were no better than on a standard task. However, they did not report exactly what errors children made, though they did note that children seldomly referred to both locations of the chocolate. By contrast, the error pattern revealed in Experiment 3.2 above suggests that the Split Bait manipulation may not in fact be a good test of Fodor's theory. Fodor makes the assumption that children individuate actions in the same way as he does. That is, the child may have considered \[ \text{agent will look in B\'n\'C} \] as one unique behavioural prediction, and according to Fodor, H1 should still be invoked and H2 will not be activated. The errors that were observed for the Split Bait task in this experiment provide some support for exactly this interpretation. There is a problem then, with how the child goes about individuating actions. What is to count as a unique behavioural prediction? The heuristic theory clearly needs to provide an account of how uniqueness is to be characterized. Otherwise, it is not clear that the tasks that are devised to test the predictions of the theory will be adequate tests. For these reasons I now set aside the Split Bait task and concentrate for the remainder of the chapter on the Eat Bait task.

### 3.4.3 Effect of eating the bait

In Table 3.2 the children's consistency with respect to performance on the Eat Bait and Standard tasks is presented. In this table, as for the Split Bait task (presented in Table 3.1), many children are consistent in failing both tasks.

<table>
<thead>
<tr>
<th>Pass Standard task</th>
<th>Pass Eat Bait task</th>
<th>Fail Eat Bait task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass Standard task</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Fail Standard task</td>
<td>6</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 3.2 Children's consistency in passing/failing Standard and Eat bait tasks in Experiment 3.1
However, there are six children who passed the Eat Bait task while failing the Standard task, and no children who show the opposite pattern. This is a significant difference (McNemar test of change, binomial, p=0.016).

For the Eat Bait task, the errors were more diverse than those observed for the Split Bait task. Some children (56%) were unconvinced even now that the Bait item had been destroyed. These children predicted the protagonist would search in a specific location (e.g. "in your tummy...", "in your mouth..." or "in there" [pointing to experimenter's tummy]). Just like in Wimmer & Perner (1983), the story did not successfully convey the notion that the bait object had been destroyed to many of the children. The remaining children responded with "I don't know...", or made no response.

3.5 Discussion: Experiment 1

Performance on the Eat Bait task was more encouraging for the heuristic hypothesis. However, children still performed very poorly, even on this task. The errors that children made on the Eat Bait task show that there is undoubtedly a problem here in convincing children that the Bait Item has been destroyed. The "pull" of the final resting place of the item, first demonstrated in Wimmer & Perner (1983), was even more striking here. Removing the influence of reality more thoroughly still is one of the aims of Experiment 3.2. Despite this, the results did indicate that it was possible to create a task that lessened the salience of the real object, and that performance on such a task was improved relative to a standard task. However, it is crucial for Fodor that the reason that performance was improved was that the children were consulting their belief based heuristic: H2.

There are in fact some other possible factors that might help explain the improved performance observed on the Eat Bait task. For example, Russell (1992) has noted that the influence exerted by the real location of the bait object is important. In Chapter 1 we saw that his theoretical account stresses the role of the "executive inhibition" in solving belief tasks. He argued that children may be doing better in situations where they don't know where the bait item is because they follow a guessing strategy.

3.5.1 Guessing

Many children in Experiment 3.1 were not convinced that the chocolate had
disappeared. Of those that were, however, it is quite possible that their prediction of search was not belief-based, but simply a guess. In the Eat Bait task, the correct location (A), as well as being the location that the agent should search on the basis of their false belief, is the only location that has any connection to the story at all. If children have no information about the agent's beliefs, but also do not know where the agent should go to satisfy their desire, then it does not seem unreasonable that they might, as a guess, pick the only other location with which the Bait was associated. On this view, the improved performance is not a result of forcing children into considering what they know about the agent's beliefs, à la Fodor, but rather results from children having no other plausible alternative guess that would be wrong. Assessing this "guessing" strategy account is another aim of Experiment 3.2.

3.5.2 Summary Experiment 3.1

Thus, Experiment 3.1 was not completely successful in creating situations where there is no unique behavioural prediction based on the agent's desires. Firstly, splitting the bait simply begs the question of how children individuate behaviours, and therefore cannot be regarded as an ideal test of Fodor's theory. Secondly, destroying the bait leads to both practical problems, in that a real item will by necessity have some final location associated with it, and interpretation problems, because once the Bait has disappeared, one needs to be sure that better performance is not a result of guessing.

The next experiment to be reported here represents another attempt to test Fodor's heuristic hypothesis. The question of the individuation of actions raised by the Split Bait task is not dealt with here, and Experiment 3.2 is based instead around the promising result in the Eat Bait condition of Experiment 3.1. In addition to creating a situation where the final location of the bait object is not known to the children, Experiment 3.2 attempts to tackle the issue of guessing strategies that children might employ.

3.6 Experiment 3.2 introduction

This experiment tested children on four new versions of the false belief task. These versions were designed to test Fodor's heuristic theory for the child's performance on action prediction problems, by comparing situations where the bait item disappears with situations
where the bait item is transferred to a specific location known to the child.

All the stories were presented using pictures rather than props. This was in order to avoid the practical problems of having to destroy items in real space. The new versions of the task varied with respect to two factors. Half of the tasks presented a situation where the action prediction from HI was not unique: the bait item was moved from the hiding place and disappeared. The other tasks involved the bait item being moved to a specific final location. As before, the prediction was that if Fodor is right, children should perform better on the tasks with a non-unique prediction from desires than on those with a specific prediction from desires.

The second factor was designed such that the experiment would be sensitive to the guessing strategy that confounded the results of Experiment 3.1. Half the stories involved the bait item moving to an intervening location before the final transfer. If the ambiguous tasks do succeed in inhibiting the response from HI, then a guessing strategy could result in responses being directed at either the correct or the intervening location. In addition, the intervening location is the most recent resting place of the bait item before its location becomes ambiguous. The prediction of the guessing hypothesis was that although responses in this task should be less likely to be associated with a final location (there is none) they would not be universally directed at the correct location.

It was hoped that these new tasks would overcome the problems associated with the attempt to test the theory presented in Experiment 3.1.

3.7 Experiment 3.2 method

3.7.1 Design

A 2 x 2 factorial within-subjects design was employed. Each child received a story with each possible combination of the two factors: four stories in all. The first factor, as described above, was whether the final location of the bait item was known to the child. The bait would either i) end up in a location known to the child (Specific) or ii) end up in a location unknown to the child (Ambiguous). The second factor was the number of locations that the bait rested in during the story. The bait either i) travelled directly from the hiding place to the final location (Immediate stories), or ii) travelled first to an intervening location, and then on to the final location (Intervening stories).
Each child received the stories in one of four counterbalanced orders. Each order alternated ambiguous and specific stories, starting with an ambiguous story. This was to ensure that general effects of repetition would be distributed across the conditions. Equal numbers of children received each order.

3.7.2 Subjects

48 children were used in this experiment. The children were taken from three age ranges, with 16 children in each range. The youngest group comprised eight girls and eight boys aged between 38 and 45 months (mean = 42.1, sd = 1.8). The middle group comprised eight girls and eight boys aged between 46 and 52 months (mean = 48.9, sd = 1.8). Finally, the oldest group comprised seven girls and nine boys aged between 52 and 55 months (mean = 53.6, sd = 1.0). The majority of children were recruited from a range of nursery classes and year one classes in North London. The schools supported a wide range of socio-economic backgrounds, with roughly equivalent numbers of children in each age group drawn from each school. The remaining children were recruited through the subject pool at the Cognitive Development Unit, and it was ensured that there were equal numbers of these children in each age group. None of these children had taken part in Experiment 3.1. All children received all four tasks, so no allocation to condition was required. The children were assigned randomly to one of the four orders.

3.7.3 Tasks and materials

All the tasks were presented using pictures drawn on cards in order to help solve the practical problem of destroying a bait item that affected the results of Experiment 3.1. An object in a picture can be made to disappear very much more readily than a real object. The stories in the tasks all involved action based indirect belief attribution ("Where will <agent> look...?") based loosely on the Sally/Anne task of Baron-Cohen et al. (1985). A protagonist hid an object in a hiding place and then left the scene. The object was moved elsewhere and the children were asked to predict where the character would search for the object on their return to the scene. The four versions used in this experiment differed on two dimensions, as described in the design. For both of the Immediate tasks, the story was presented using four pictures. For the Intervening tasks, six pictures were used. In each story there were
the same Locations: a box, a desk with a tray on top and a drawer, a bin on the floor and
an open window. The stories were all set in the same room, with different characters in each
one. Schemes depicting the form of the stories appear in Figures 3.2, 3.3, 3.4 and 3.5, with
the protocols alongside.

Experiment 3.2 tasks: Immediate stories (Figures 3.2, 3.3)

1. Here is Jane. Can you see Jane has some chocolate? Jane is going to eat her chocolate
but before that she is going for a walk with her friends.

2. Next, the protagonist is shown placing the bait item in the hiding
place. It was a box for all tasks. Jane puts her chocolate over here in the
box. Can you see that? The control question is asked after the second
picture has been turned over: Where did Jane leave her chocolate..?

3. In the third picture a second character removes the item from its hiding place. Look what
happens when she isn't here. Her friend Cleo comes along and he takes the chocolate out
of the box. She eats the chocolate all up. The chocolate is completely gone. (She puts the
chocolate into this drawer - Specific version).

4. The final picture shows the empty room. The test question was asked with this picture in front of
the children: When Jane comes back, where will she look for her chocolate..? Where is her
chocolate really..?
Experiment 3.2 tasks: Intervening stories (Figures 3.4, 3.5)

1. Here is Jane. Can you see Jane has a picture? Jane is going to colour her picture in. But before that she is going for a walk with her friends.

2. Next, the protagonist is shown placing the bait item in the hiding place. It was a box for all tasks. Jane puts her picture over here in the box. Can you see that? The control question is asked after the second picture has been turned over: Where did Jane leave her picture?..?

3. In the third picture a second character is shown. The character removes the item from its hiding place. Look what happens when she isn't here. Her friend Cleo comes along and she takes the picture out of the box. She puts the picture in this tray. Did you see that..?

4. Cleo goes out. The second control questions are asked at this point, once the fourth picture has been turned over: Where did Jane hide the picture right at the beginning...? Where did Cleo move the picture to...?

5. The wind comes along and it blows Jane’s picture into the bin! (…blows the picture right out of the window! The picture is completely gone - Ambiguous version).

6. The final picture shows the empty room. The test questions are asked while this picture is in front of the children: When Jane comes back, where will she look for her chocolate..? Where is her chocolate really...?
3.7.4 Procedure

Each child was tested individually. In the schools, the child was tested in a quiet corner of the main area. In the lab, the child was tested in a small quiet room. The child sat opposite the experimenter, who presented the stories one at a time in one of the four orders. Each story lasted between two and four minutes, and the replies to the control and test questions were recorded at the time of the experiment. The whole session lasted about 15 minutes. For each story the cards were held by the experimenter and laid down one at a time in front of the children. After the first two pictures had been laid, both were turned over and the first control question was asked. If a child failed to answer this question correctly then the story up to this point was repeated using the pictures again. If the child still failed the control question then he or she was excluded from the study and replaced. For the Intervening stories the pictures were again turned over for the second control question, and the same criterion was used when children were wrong. The final picture, showing the empty room, was left face up in front of the children whilst they were asked the test question. The response could be thus verbal or by pointing, and was recorded on a sheet of paper by the experimenter.

3.8 Experiment 3.2 - results

Nine children were eliminated for failing one or more of the control questions. Performance on the test question in each condition, for each age group, appears in Figure 3.6 below. Children were scored correct if they predicted that the agent would search the original location (Location A). This Figure shows the number of children from each of the three age groups who passed each of the four tasks. On inspection of Figure 3.6, the first main impression is that the oldest group of children appear to be doing better than the other two groups. However, the youngest age group does not appear to be very much worse than the middle group, and the difference between all the groups is smaller for both Ambiguous tasks.

These speculations about the effect of age were tested by performing separate Chi-squares for each condition. For the Ambiguous Immediate and Intervening conditions, there was no significant effect of age ($\chi^2|N=48, df=2| = 4.67$, two-tailed $p=0.09$, and $\chi^2|N=48, df=2| = 3.73$, two-tailed $p>0.1$ respectively).
Figure 3.6 Performance by youngest, middle and oldest age groups in each condition in Experiment 3.2

For the Specific Immediate task, there was a significant effect of age ($X^2[N=48, df=2] = 7.18$, two-tailed $p<0.05$), which further partitioning of the Chi-squared table revealed to be from the youngest and middle versus the oldest age group (for youngest versus middle: $X^2_{1,\text{corr}}[N=48, df=1] = 0.13$, two-tailed $p=N.S.$, for youngest + middle versus, oldest: $X^2_{2,\text{corr}}[N=48, df=1] = 7.05$, two-tailed $p<0.05$). For the Specific Intervening task there was also a significant effect of age ($X^2[N=48, df=2] = 12.19$, two-tailed $p<0.01$). Again, partitioning revealed this effect to be between the oldest and the other two groups ($X^2_{1,\text{corr}}[N=48, df=1] = 0.13$, two-tailed $p=N.S.$, and $X^2_{2,\text{corr}}[N=48, df=1] = 12.06$, two-tailed $p<0.001$).

Overall, this pattern is consistent with Fodor's Hypothesis, with older children
significantly better at only the standard versions of the tasks. On the Ambiguous tasks, children from both the youngest and middle groups performed well enough to be statistically indistinguishable from the oldest children.

Since the oldest children are performing nearly at ceiling on all tasks, they will not be considered further. The remaining analyses concentrate on the specific effects of the ambiguity and extra location manipulations within the youngest and middle age groups. Because the youngest and middle groups were not significantly different in any of the conditions, the groups will be collapsed for these remaining analyses.

The effects of ambiguity and extra location, and their possible interaction, were tested as follows. Children were scored out of 1 for each of the four tasks. For ambiguity, each child's score for each of the two Specific tasks was added together (giving a score out of 2). This score was subtracted from the child's score out of 2 for the Ambiguous tasks. This gives each child a score that ranges from -2 to +2, which represents his or her tendency to perform better under conditions where there is no unique behavioural prediction from the agent's desire. A child scoring +2 would have done very well on the ambiguous tasks and poorly on the tasks where a unique prediction can be derived from the agent's desires; the pattern predicted by Fodor. A child scoring -2 would represent the opposite pattern. The null hypothesis predicts that children will perform no better on non-unique tasks than on unique tasks. Under H₀, then, the scores from the children should be normally distributed with a mean of 0, indicating that the children are no more likely to pass the Ambiguous tasks than the Specific tasks. Fodor's heuristic hypothesis predicts that the score will be significantly better than 0, indicating improved performance for the ambiguous tasks, relative to that observed on the specific tasks. The score for each child was calculated, and the distribution of these scores was found to have a mean significantly above 0 (t[31] = 3.1, p<0.002).

This procedure was also used to calculate the effect of the extra location factor. Here the scores that a child received for both Intervening tasks are added together and subtracted from the sum of the child's scores for the Immediate tasks. Here, a child scoring above 0 shows a tendency to perform better when there are fewer locations and a child scoring below 0 performs better when there are more locations. These scores were also calculated and the resulting main effect for extra location was not significant (t[31] = 0.57,
Finally, the procedure was also used to test whether the two factors interacted. The children's scores for the Immediate Ambiguous and Intervening Specific tasks were added and this score (out of 2) was subtracted from the sum of their scores from the Intervening Ambiguous and Immediate Specific tasks. This measure represents the degree to which each child's performance is indicative of an interaction between the two factors. The resulting distribution of scores was not significantly different from 0 (t[31] = 0.7, N.S.).

This analysis shows that, as Fodor predicted, children perform better at versions of the false belief task that do not involve a unique behavioural prediction from the Agent's desires. There is no evidence that the number of locations through which the bait item passes affects the advantage enjoyed in the ambiguity conditions.

These effects were also assessed non-parametrically. Children were classified according to their consistency across the Ambiguous and Specific tasks they received for each level of the extra location factor. Table 3.3 shows the number of children passing and failing the Ambiguous and Specific versions of the Immediate tasks. Table 3.4 shows the same classification for the Intervening tasks.

<table>
<thead>
<tr>
<th>Table 3.3 Children's consistency in passing/failing Ambiguous and Specific Immediate tasks in Experiment 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass Specific</td>
</tr>
<tr>
<td>Pass Ambig</td>
</tr>
<tr>
<td>Fail Ambig</td>
</tr>
</tbody>
</table>

It can be seen in each of these tables that although there were many children who were consistent in their performance across the two tasks, there were some children who passed only one of the two tasks. In both cases, the vast majority of children passed only the Ambiguous task (McNemar test of change, binomials, p=0.003 and p=0.035 respectively).
Table 3.4 Children's consistency in passing/failing Ambiguous and Specific Intervening tasks in Experiment 3.2

<table>
<thead>
<tr>
<th></th>
<th>Pass Specific</th>
<th>Fail Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass Ambig</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Fail Ambig</td>
<td>1</td>
<td>14</td>
</tr>
</tbody>
</table>

This same method was used to confirm the analysis of the effect of extra location described earlier. This time children were classified according to their consistency across the Immediate and Intervening tasks they received for each level of the Ambiguity factor. Table 3.5 shows the number of children passing and failing the Immediate and Intervening versions of the Ambiguous tasks.

Table 3.5 Children's consistency in passing/failing Immediate and Intervening Ambiguous tasks in Experiment 3.2

<table>
<thead>
<tr>
<th></th>
<th>Pass Intervening</th>
<th>Fail Intervening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass Immediate</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>Fail Immediate</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

The guessing hypothesis outlined in the introduction suggests that a child's better performance under conditions where there is no unique prediction from desires may result not from the child now being able to predict on the basis of what she knows about belief,
but rather from guessing on the basis of where else the object has been. If this is true then it predicts that performance on the Ambiguous Immediate task will be better than that on the Ambiguous Intervening task. From the table above it can be seen that there are more children passing the Immediate task and failing the Intervening task than vice versa. However, this difference is not significant. (McNemar test of change, binomial, p = 0.145).

Table 3.6 shows the number of children passing and failing the Immediate and Intervening versions of the Specific tasks in Experiment 3.2.

Table 3.6 Children's consistency in passing/failing Immediate and Intervening Specific tasks in Experiment 3.2

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immediate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Fail</td>
<td>2</td>
<td>19</td>
</tr>
</tbody>
</table>

Looking at the consistency over the Specific tasks also shows no indication that Immediate tasks were performed better than Intervening tasks, with the same number of children passing only the Intervening Specific task as passing only the Immediate Specific task.

The children's wrong answers were also recorded. For the Specific tasks the children's errors were mostly directed at the final location of the object (17 out of 21 for Immediate task, 19 out of 21 for Intervening task). The remaining errors were children responding that they didn't know. For the Ambiguous tasks, some children were still inclined to offer the objects presumed final resting place. Five children in the Immediate task predicted search in the "mouth" or "tummy" of the protagonist who had eaten the chocolate, while eleven children in the Intervening task predicted search "outside" or "through the window" in the Intervening task. The remaining errors in the ambiguous tasks were split between "don't know" and search at a further location (five and one respectively for the
Immediate task, two and two for the Intervening task). These results are presented in Table 3.7 below.

Table 3.7 Children's errors on each task in Experiment 3.2

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Final Location</th>
<th>Distracter</th>
<th>Don't know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immediate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguous</td>
<td>21</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Intervening</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguous</td>
<td>17</td>
<td>11</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Immediate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific</td>
<td>11</td>
<td>17</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Intervening</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific</td>
<td>11</td>
<td>19</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

It is interesting to note that no responses were directed toward the "intended" distracter location (the tray) in the Intervening Ambiguous task. All were in fact directed toward the Bin, which was present in all stories. This might have been a result of intrusion from other tasks, but cannot have been in every case since one such error occurred on the first task presented to one child.

Overall the numbers of errors to distracter locations was too small for there to be any hope of substantiating the hypothesis that children were guessing on the basis of where the item had been. It seems that the greater number of errors on the Ambiguous Intervening task as compared with the Ambiguous Immediate task was most likely a result of the materials used; the item disappearing through a window was a less convincing disappearance than the item being destroyed. Performance on the conditions was not significantly different however (see Table 3.5). This possibility suggests that the materials should have been counterbalanced across conditions.

3.9 Experiment 3.2 discussion

The results of Experiment 3.2 indicate stronger support for Fodor's heuristic theory.
Children in the younger and middle age groups were significantly better at correctly predicting an agent's action under conditions where there was no unique behavioural prediction available from the agent's desire. Presenting the tasks via pictures was more successful in creating ambiguity about the items final location than the attempt using real objects in Experiment 3.1.

Including an additional Location with which the bait item had been associated did not produce a significant decrement in performance. However, this was most likely a result of the specific story used in the Intervening task. Overall, the children's improved performance on the Ambiguous tasks did not appear to be a result of their having been induced into guessing: the errors that children made were no more diverse in the Intervening condition than in the Immediate condition.

3.10 General discussion

It is possible, then, to conclude that there is some support for the heuristic hypothesis of Fodor in the results of these experiments.

Experiment 3.1 compared performance on two versions of false belief tasks that created ambiguity in the prediction derived from considering the agent's desires (Fodor's H1) with performance on a standard task. Fodor's hypothesis predicted better performance on both of these two new tasks relative to performance on the standard task. The results indicated that splitting the bait among two new locations does not improve children's performance. This result was also found by Wimmer & Weichbold (1994), but, as argued earlier, there are reasons to suspect that the split bait experiment is not a good test of Fodor's theory.

The results from the manipulation involving the destruction of the bait item were more encouraging. However, a large number of children still appeared unconvinced that the item had been destroyed and thus this test too may well have been flawed as a test of ambiguity in prediction from the agent's desires. In addition, the children who did succeed may have done so as a result of a guessing strategy. The correct Location in the Eat Bait task was also the only Location mentioned during the story. Children may have predicted search here for this reason alone.

Experiment 3.2 succeeded in overcoming the reservations expressed about
Experiment 3.1. Although children were not all convinced the bait item had disappeared (and indeed the materials in one condition were more convincing than those in the other), there were fewer final location responses in Experiment 3.2 than there had been in Experiment 3.1. There is still an interpretation problem, however, with the results of Experiment 3.2.

3.10.1 Guessing revisited

The guessing hypothesis that was presented and tested in Experiment 3.2 can be made slightly more elaborate. For example, in Experiment 3.1, the correct (i.e. belief-based) location was argued to be the only one that had been mentioned in the story, and hence the obvious "guess" for a child who had no other idea where to predict. In Experiment 3.2 there are two locations with which the bait item has been associated, but only one of these two had been associated with the bait item and agent together: the correct Location. Although the bait item is associated with an intervening location, this does not happen when the agent is in the room, and so this association may be less relevant than the location that is associated with the item and the agent at the same time.

If children's guessing is as complicated as this, then performance on the tasks of Experiment 3.2 can still be explained with no need to accept that it is information about the agent's beliefs that is responsible for their improved performance.

3.10.2 The "Seeing" control for dumb strategies

The solution to this problem is not initially clear. It would be possible to design an experiment that included a control condition for this new guessing hypothesis. The agent would have to leave the item in one location and then move it to another location before leaving the room. This would associate two locations with the agent and item together before the item was moved to the ambiguous location. If guessing, the children would predict some searches to the wrong but still associated location. However, they could still in principle perform this task correctly on the basis of some factor other than an appreciation of belief (e.g. the correct location is always now the last of the 2 locations associated in this

\[\text{On reflection it would have been desirable to manipulate the materials between the two Ambiguous tasks. In fact however, there is a better control procedure against "guessing" strategies that I will introduce shortly.}\]
way). In addition, the overall strain on the children's memory becomes even higher. Although such an experiment decreases the likelihood of a false positive, it also increases the likelihood of a false negative. Recall that the assumption of Fodor's and other CC theories is that children's problems are a result of an overall strain on resources. For example the ToMM-SP model assumes that there are both inhibition and selection problems involved in false belief tasks. Making the stories more complex in the manner suggested above may well remove certain inhibition problems but at the same time may stress the selection problems, and performance might end up no better.

Recently, a better solution to this type of problem has been suggested by Surian & Leslie (1994). Surian & Leslie were also interested in the role of ambiguity of desire in false belief problems. They present children with a modification of a task introduced by Robinson & Mitchell (1992), a task that has also been open to a "false-positive" interpretation. In the new task the child's problem is to identify the referent of a speaker's message. Sally has four pencils, three of which are sharp and the other broken. Sally leaves the pencils and then enters the next room. While she is gone Ann breaks the sharp pencils and then sharpens the broken pencil. Sally then calls to Ann: "bring my favourite pencil - you know, the broken one...". The children are asked which pencil Sally really means.

Here there is ambiguity in Sally's desire which, according to Fodor, should result in the child consulting her belief. Sally's belief is of course that the sharp pencil is the broken one, which is the correct answer. Surian & Leslie found that 48% of children passed this task, which was better than on a standard task. However, it is essential to rule out false positives. For example the word favourite might pick out a "unique individual", which would be the sharp pencil. The appropriate control, according to Surian & Leslie, is to include a "See" condition, where sally sees what Ann does to the pencils. If the children show the same pattern of responding when Sally sees as when she doesn't see, then it is reasonable to assume that some "dumb" strategy, such as the unique individual idea suggested above, is at work. Surian & Leslie find however, that only 15% of children pick the sharp pencil in the See control condition, a proportion significantly different than that observed in the No-see experimental condition.

The seeing control has proved useful in resolving other issues. Recall the experiment performed by Siegal & Beattie (1991; see Chapter 1, Section 1.6.3, also footnote 7). These
authors found that children performed well if asked to predict where Sally would look "first" for the bait object. One dumb strategy suggested for this result is that children might assume that "look first" requires a look that will fail. The control reported by Siegal & Beattie is not compelling however, because as pointed out by Surian & Leslie (1994), the "inferred belief" control has items in both locations, so there is nowhere for Sally to direct a "failing look". Surian & Leslie ran a "look first" version of the standard false belief task, and included a Seeing control, where Sally sees Ann move the bait. Under these conditions the "failing look" dumb strategy does make a prediction: that Sally will search the empty location. However, children are sensitive to whether Sally saw or not in their choice of the empty Location, and only predict search there in the Not See experimental condition.

Clements & Perner (in press) also use a Seeing control for their "implicit" eye-direction measure of false belief understanding. They found that children are reliably better at false belief tasks if one measures where they look in anticipation of where an agent with a false belief will return to retrieve an object. Children pass the task on this measure six months before they pass when required to make an explicit verbal or pointing response. The result is not due to the children simply following the sequential location of the bait item, however; the children do not look to the original location in conditions where the agent is present, and sees the transfer event.

The appropriate control for "dumb" strategies then, would appear to be a Seeing control. It is not clear that such a control would work for the stories employed in Experiment 3.2, however, because it is not clear that a character who saw chocolate disappear should be inclined to search anywhere for it. The procedure would have to be more extensively modified.

3.10.3 Explanation as a converging measure of belief understanding

Another way of determining whether children are consulting belief in making their predictions is to ask them to justify their prediction. This can be done either by asking them why they think the agent will search where they predict he will, or alternatively simply by asking them to explain the action itself. Wimmer & Weichbold (1994) made use of the former requirement. Such a measure might be considered conservative: providing an explanation for an action (or justifying a prediction of an action) may simply be harder than
predicting an action. In fact, comparisons of prediction and explanation have appeared to indicate that explanation is easier (e.g. Bartsch & Wellman, 1989). Nevertheless, in general, the process of explanation in theory of mind has received far less attention than the processes of action prediction and belief attribution. I turn in the next four chapters to the topic of "backwards reasoning" in theory of mind, and concentrate on the issue of action explanation in particular.

3.10.4 Fodor’s heuristics in the context of other CC theories

Before moving to consider explanation it is worth placing Fodor’s heuristic theory in the context of other CC theories, notably the ToMM-SP theory. These theories share the following assumptions: that performance limitations are responsible for the 3-year-old’s failure at the false belief task, and that the concept of belief held by preschoolers is normative (see Chapter 1). One principal difference between the accounts is the role of the normativity of belief. For Fodor, the 3-year-old child only calculates belief when certain conditions hold, conditions that depend on the action prediction derivable from the agent’s desires. In ToMM-SP, by contrast, the child routinely calculates belief. The problem is that though the child routinely calculates the belief content, she has to overcome a prepotent response created by the normativity of belief.

Leslie & German (in press), have argued that the assumption that the 3-year-old does not routinely calculate belief is incorrect. For example, 3-year-olds fail direct belief attribution tasks, where they are simply asked to calculate the belief contents. This finding is troubling for the proposal that the problem is simply an error of omission, and reveals that Fodor’s theory is limited by its consideration only of action based belief. In Chapters 5 and 6, I return to the issue of whether or not children routinely calculate belief, and supply evidence from the explanations that they offer for incorrect search actions.
CHAPTER 4

Action explanation:
A converging measure of belief understanding?

4.1 Introduction

In this chapter I present an overview of "backwards reasoning" experiments in theory of mind. This sort of task involves presenting children with a false belief situation, and giving them more evidence with which to calculate belief than is typically offered in the standard false belief task. For example, children might be shown a protagonist searching in the wrong place for an item, or expressing surprise that an item is absent, and asked to reason backwards from this evidence to the protagonist's false belief. This group of tasks includes studies of the explanation of action, where the indirect measure of belief attribution is an open-ended "why..?" question (e.g. "why did <agent> look there..?").

I will deal here with the general rationale behind backwards reasoning, introduced in the last chapter, and present the picture that emerges from three significant investigations: Bartsch & Wellman (1989), Moses & Flavell (1990) and Robinson & Mitchell (in press). I will then outline and address a number of methodological issues that arise from the action explanation procedures contained in two of these papers (Bartsch & Wellman and Moses & Flavell), focusing on the issue of how the explanations of young children are to be categorized. I present a scheme for categorization that is based on considerations of what a child's explanation means in terms of how (s)he might have processed the task. The scheme is based on those used in both Bartsch & Wellman (1989) and Moses & Flavell (1990), but has some modifications.

4.2 General overview: backwards reasoning tasks

The reasoning behind standard action prediction and belief attribution tasks, which were discussed in Chapter 1, is based on the notion that commonsense psychology holds
that\textsuperscript{18} actions are caused by an interaction of mental states (e.g. Davidson, 1963). For example, in the false belief action prediction task (Wimmer & Perner, 1983; see also Chapter 1), the child must infer that a character's belief will cause him to search the empty location. In backwards reasoning tasks the logic is the same, only now the task is to reason backwards from a character's mistaken search and infer the false belief that caused it. In the next sections I will outline three papers where the authors apply this logic (Bartsch & Wellman, 1989; Moses & Flavell, 1990; Robinson & Mitchell, in press). The authors are interested at times in what I refer to as (backwards) belief attribution, where the children are asked to infer beliefs directly, and at other times with what I refer to as action explanation, where the children are asked simply to give an open ended answer to a "why...?" question.

This turns out to be an important difference, for reasons that I will expand on in the introduction to my first explanation investigations in Chapter 5 (see sections 5.1 and 5.2). Briefly, however, it is because open-ended action explanation questions have "correct" answers that do not refer to belief (despite the logic outlined above). Deciding whether a particular explanation is based on backwards reasoning to belief may be quite complicated. For example, a child who explains a protagonist's incorrect search in a false belief task with reference to the protagonist's desire to find the bait object has not indicated that (s)he understands the false belief. However, the desire is part of the cause for the action, and it is not possible rule out the possibility that the child understands this, without further prompting. The backwards reasoning studies that I outline adopt a mixture of these procedures, with greater or lesser emphasis on action explanation. In Chapter 5 I look in more detail at the relationship between desire, belief and action, to derive some specific predictions in an action explanation procedure.

\textbf{4.2.1 Bartsch & Wellman (1989)}

Bartsch & Wellman (1989) present a belief-desire reasoning scheme that was devised

\textsuperscript{18} Theorists studying the acquisition of commonsense psychology might, but need not necessarily believe that commonsense psychology is true. That is to say that, although actions might really be caused by an interaction of a number of real indivisible mental entities (such as beliefs and desires), it is also possible that commonsense psychology is in fact false (see for example; Ramsey, Stich & Garon, 1991; see also Chapter 2). This possibility notwithstanding, in commonsense psychology we do in fact attribute desires and beliefs to agents in order to explain their actions. What is at stake here is not whether the theory is true but rather what form this knowledge base or theory takes, and how its acquisition proceeds.
on the basis of their simple belief reasoning tasks (Wellman & Bartsch, 1988; see Chapter 1). This figure is reproduced in Figure 4.1 below. The authors state their purpose as being: "to examine young children's ability and inclination to reason backwards with respect to this scheme..." (1989, p947).

![Figure 4.1 Belief-desire reasoning scheme of Bartsch & Wellman (1989).](image)

In the first of their experiments they employ an open-ended explanation question methodology, and supplement this with some prompt questions. Bartsch and Wellman presented adults, 4- and 3-year-old children with simple descriptions of characters engaging in actions, along with a picture of the event. The children were then asked to explain the character's action. Three different descriptions of action were used: neutral descriptions of single simple actions; and two sorts of descriptions of actions that were contrary to the Agent's apparent desires. These could be anomalous belief actions, where the action was further specified so as to create the contradiction (a false belief situation), or anomalous desire actions where the desire of the character was further specified to create the contradiction. An example of each story type appears in Table 4.1. Three examples of each
story type were presented to each subject in one of two random orders.

Table 4.1 Examples of each story type used by Bartsch & Wellman (1989, Experiment 1)

<table>
<thead>
<tr>
<th>Story Type</th>
<th>Story Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral Story</td>
<td>Here's Jane. Jane is looking for her kitten under the piano.</td>
</tr>
<tr>
<td>Anomalous Desire Story</td>
<td>Here's Beth. Beth hates apples. But she's taking a bite out of this apple. Why do you think Beth is doing that?</td>
</tr>
<tr>
<td>Anomalous Belief Story</td>
<td>Here's Jane. Jane is looking for her kitten. The kitten is hiding under the chair. But Jane is looking under the piano. Why do you think Jane is doing that?</td>
</tr>
</tbody>
</table>

Immediately following the presentation of the pictures and story, the children were asked the open ended explanation question: "why do you think Jane is doing that..?". If in answer to this question the subject did not use a belief or desire term, then they asked a further prompt question. The prompt question was either a desire prompt: "what does Jane want..?" (for the anomalous desire stories and one of the neutral stories), or a belief prompt: "what does Jane think..?" (for the anomalous belief stories and the other two neutral stories).

This study is used as the basis for the explanation studies to be presented in the next chapter, and a detailed appraisal of the results appears there. Briefly, however, the experiment showed that children were capable of offering relevant belief and desire explanations for each of the story types. The patterns of explanation were not different across the three age groups, and explanations included references to false belief even among the 3-year-olds, especially when the prompt questions were taken into account.

Bartsch & Wellman were encouraged by the results of this open-ended action explanation experiment to further investigate what they describe as "an understanding of
belief not evidenced in previous studies" (Bartsch & Wellman, 1989, p946). This "belief-understanding" was inferred largely on the basis of the prompted answers to the explanation questions in the first experiment. Their second experiment went on to pit understanding of false belief in an explanation setting, against understanding false belief in a prediction setting. In this second experiment they also included an open ended explanation measure, but were most concerned with the children's answer to a belief prompt: "what does <agent> think...?"; a direct measure of belief attribution. They found that 3-year-olds were correct about a third of the time when predicting action, but about two-thirds of the time in the explanation task. They went on to argue that these results demonstrate that false belief attribution in explanation is easier than prediction of actions induced by false beliefs. They suggest that this is because in prediction, but not explanation, there is a conflict between the protagonists desire and their false belief. In the next Chapter, I will argue that desire is in fact just as important (if not more important) in explanation as it is in prediction, and provide evidence to substantiate this claim.

These data, and the subsequent conclusions, have attracted much criticism (e.g. Perner, 1991; Moses & Flavell, 1990). For example, Perner (1989, 1991) has argued that children might understand thinking as "thinking of" rather than "thinking that". This argument was used by Perner to undermine Wellman's belief based prediction tasks (Wellman & Bartsch, 1988: see Chapter 1 Section 1.6.2). In addition, Wellman (1990) noted that the false belief explanations were largely from the older 3-year-olds, and suggested that only children who are at a transitional phase in development (see Chapter 1, Section 1.6.4) find explanation easier. I return to Wellman's position on explanation in the next chapter, where the results of his first explanation experiment are dealt with more thoroughly.

4.2.2 Moses & Flavell (1990)

Moses & Flavell (1990) also criticized the conclusions drawn in Bartsch & Wellman (1989). Moses & Flavell also presented two experiments, both of which were concerned with young children's ability to "reason backwards to the belief from its effects (e.g. from a protagonist's actions and reactions)" (1990, p929). The measure of primary interest for Moses & Flavell was a direct belief attribution question, just like Bartsch & Wellman's second experiment. Children were assigned to one of three conditions in which they were
shown two short films of a false belief scenario. The scenario was a standard one: an agent (Cathy) placed an item (crayons) in a location (bag) and left the room to search for some paper. A clown came in and replaced the crayons with rocks, putting the crayons in a drawer. A direct measure of belief attribution was employed: children were asked what the protagonist believed was in the bag.

In the first condition, this question was asked when the film had been stopped at the point that the protagonist came back into the room, requiring children to infer the false belief on the basis of the lack of perceptual access to the transfer event. In the second condition the film was stopped after the protagonist had approached the bag, and was about to open it, allowing the children information about the agent's actions as well as information about the lack of perceptual access. In the last condition, children were also shown the protagonist's surprise reaction on discovering rocks in the bag, and were allowed this extra information as well. Moses & Flavell hypothesized that performance would improve across the three conditions. In addition to the belief attribution questions, some of the subjects were also asked an open ended explanation question after the second of their two stories.

With respect to the belief attribution results, although Moses & Flavell demonstrated a monotonic trend across the conditions, with performance improving as a function of the amount of information available about the agent's belief, they were not convinced that the absolute performance was better than chance (one correct out of two stories). They replicated the surprise condition, where they had observed the best performance, with a further 19 children. This second experiment also included an open-ended explanation question. The conclusion from this replication was that children's performance, even in the surprise condition where there is much information about what the agent believes, is not reliably better than chance.

As far as explanation was concerned, the following procedure was carried out. On the second of the two films, after the belief attribution questions had been asked, the film was stopped at the point at which the agent was looking in the wrong location. Half the children (n=24, eight from each condition) were asked the open ended question: "why is Cathy looking in the bag?". Following this, these children, and most of the remaining children (who hadn't been asked the action explanation, total n=46), were shown the character's surprise reaction and asked the surprise explanation question: "why is Cathy
surprised..?

Analysis of children’s answers showed that for the action explanation children most frequently explained in terms of the agent’s desires, while for the surprise explanation children most frequently explained in terms of the outcome of the search. Moses & Flavell make the important observation that neither of these patterns reflects incorrect responding; the desire is one of the causes of the mistaken search (along with the false belief about the bait’s location); the outcome of the search is a reason for surprise (along with the expectation that a different item would be located). Moses & Flavell make the point that in both cases where there is a choice in explanation between a belief construct and some other construct, belief is rarely offered. This point is crucial in setting explanation aside from backwards belief reasoning; and will be raised again in the introduction to Chapter 5. Moses & Flavell interpret this and their other results as indicating a profound problem for 3-year-olds in understanding the representational nature of beliefs.

4.2.3 Robinson & Mitchell (in press)

Robinson & Mitchell (in press) present an outline of the backwards reasoning debate and argue that there is still good reason to expect that backwards explanation to be easier than belief prediction. Robinson & Mitchell use no open-ended explanation measures, however; instead their tasks use various forms of indirect and direct belief attribution questions. For this reason the experiments in this paper are less relevant to explanation than the two mentioned previously, but are nevertheless important from the point of view of backwards reasoning in general.

In a series of experiments, Robinson & Mitchell compared backwards explanation with forwards prediction, using an indirect measure of belief attribution. The child’s task was to identify which of two identical twins must have performed certain actions; the twins can only be tracked in terms of the different beliefs they entertain. In the first study for example, the twins were present when the bait item (ball) was hidden. One twin went outside and the ball was moved. In the prediction condition children were asked to predict where the twin who had gone outside would search for the ball. In the explanation condition the children

19 In their first investigation Robinson & Mitchell do use a "Why...?" question, but force the children into a choice: "Was it because of x or was it because of y...?".
were shown both twins searching in different locations for the ball, and were asked whether the twin who had gone to the wrong place had done so because he was outside or had stayed inside when the ball had been moved. The results showed the backwards explanation condition to be significantly easier than the prediction condition, and performance was significantly better than chance. Robinson & Mitchell went on in further experiments to show that this resulted from the *behaviour having happened* in the explanation case, rather than the direction of reasoning *per se*, and also that various association strategies (e.g. Perner, 1991) could not account for the results.

Robinson & Mitchell drew a number of important conclusions from these results. These conclusions are for the most part consistent with competence continuity (CC) versions of theory-theory. The role of reality is stressed, as well as the idea of giving children extra help in calculating the contents of beliefs (see Chapter 1, Section 1.8.2). The issue of extra cues to belief calculation within an explanation situation are readdressed in Chapter 7. I turn now, however, away from backwards reasoning in general, to the specific issues involved in the open ended action explanation measures used in Bartsch & Wellman (1989, Experiment 1) and Moses & Flavell (1990).

### 4.3 Action explanation

An action explanation task has been defined here as a task that employs an *open ended* explanation question. In the experiments carried out by Bartsch & Wellman and Moses & Flavell, the action explanations were similar in that they required the child to explain why a character was looking in a particular location. However, there were also differences between the two papers in terms of both the procedure adopted around the explanation task and also the methodology employed in categorizing the explanations themselves.

#### 4.3.1 Procedural differences

Procedural differences between Bartsch & Wellman (1989) and Moses & Flavell (1990) are summarized in Table 4.2. One of the purposes of the experiments reported in the next few chapters (5, 6 and 7), is to assess whether differences in task structure, like those mentioned in Table 4.2, can influence the pattern of explanations that children might offer.
For example, we have already seen in Chapter 3 that presenting children with picture stories rather than real objects was successful in inducing children to be more convinced by a disappearance event. Bartsch & Wellman presented their explanation situations via pictures in which the location of objects was unknown, and observed more belief responses than Moses & Flavell (1990), who presented children with videos in which the location of the items was known. In this vein, in Chapter 5 I compare the pattern of explanation offered for successful (desire satisfying) actions with that offered for unsuccessful (desire frustrating) actions. Bartsch & Wellman included different story outcomes in their experiment (e.g. see Table 4.1), but their assessment of that manipulation was limited. In Chapter 6, I address the question of what effect making an action prediction has on subsequent explanation of the same action, and in Chapter 7, I look in more detail at the issue of presenting children with video evidence of false belief.

Table 4.2 Summary of Action Explanation procedures used by Bartsch & Wellman (1989, Experiment 1) and Moses & Flavell (1990)

<table>
<thead>
<tr>
<th>Types of action to explain</th>
<th>Bartsch &amp; Wellman (1989, Expt 1)</th>
<th>Moses &amp; Flavell (1990 Expts 1 &amp; 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral stories</td>
<td>False belief stories</td>
<td></td>
</tr>
<tr>
<td>Anomalous desire stories</td>
<td>(anomalous belief)</td>
<td></td>
</tr>
<tr>
<td>Anomalous belief stories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reality status of objects/events</td>
<td>Presentation using pictures. Items of search not shown: locations not known to child.</td>
<td>Presentation using video films. Items of search shown: locations known to child.</td>
</tr>
<tr>
<td>Prompt questions after explanation?</td>
<td>Children asked belief and/or desire prompts after initial explanation.</td>
<td>Some children asked to explain Agent's surprise reaction after initial explanation.</td>
</tr>
<tr>
<td>Questions before explanation procedure?</td>
<td>No questions before explanation.</td>
<td>Child asked about Agent’s beliefs before explanation question.</td>
</tr>
</tbody>
</table>

I turn now to the issue of the scoring of children's explanations.
### 4.3.2 Categorization of explanation types

Bartsch & Wellman (1989) discussed the possibility that children might reason backwards with respect to a belief-desire reasoning scheme (see Figure 4.1), which includes many explanation types. Some of these types are classified as belief- or desire-based, others invoke the more "primitive" concepts of the scheme, such as drives and perceptions. Bartsch & Wellman thus expected that their open ended question ("why is <agent> doing that...?") would result in a wide variety of possible responses. It is important for the purpose of interpreting the results of this type of experiment, to be able to both i) identify the categories of response that are of interest and ii) reliably score responses into these categories. The present section deals with the scoring methods used in each of the action explanation procedures detailed, with respect to both these issues.

#### Table 4.3 Categories of explanation and exemplars

Bartsch & Wellman (1989)

<table>
<thead>
<tr>
<th>Categorisation</th>
<th>Exemplars</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Psychological</strong></td>
<td></td>
</tr>
<tr>
<td>Beliefs &amp; Desires</td>
<td>Beliefs</td>
</tr>
<tr>
<td>Desires</td>
<td>think</td>
</tr>
<tr>
<td></td>
<td>know</td>
</tr>
<tr>
<td></td>
<td>guess</td>
</tr>
<tr>
<td></td>
<td>expect</td>
</tr>
<tr>
<td></td>
<td>believe</td>
</tr>
<tr>
<td></td>
<td>want</td>
</tr>
<tr>
<td></td>
<td>hope</td>
</tr>
<tr>
<td></td>
<td>wish</td>
</tr>
<tr>
<td></td>
<td>desire</td>
</tr>
<tr>
<td>Other psychological states</td>
<td>physiology</td>
</tr>
<tr>
<td></td>
<td>perception</td>
</tr>
<tr>
<td></td>
<td>emotion</td>
</tr>
<tr>
<td></td>
<td>changed-mind</td>
</tr>
<tr>
<td></td>
<td>pretence</td>
</tr>
<tr>
<td></td>
<td>preference</td>
</tr>
<tr>
<td></td>
<td>oughts</td>
</tr>
<tr>
<td></td>
<td>traits</td>
</tr>
<tr>
<td><strong>Non-psychological</strong></td>
<td></td>
</tr>
<tr>
<td><strong>No explanation</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>don't know</td>
</tr>
<tr>
<td></td>
<td>no response</td>
</tr>
<tr>
<td></td>
<td>repetition</td>
</tr>
</tbody>
</table>
The complete schemes are presented in Tables 4.3 and 4.4. It can be seen that there is relatively high correspondence between the two systems of categorization. With respect to point i) above, both schemes identify the same specific categories of interest: belief-based and desire-based explanations. A notable difference between the schemes is that Bartsch & Wellman’s has a structure that identifies a hierarchical organisation in the explanations. In addition it is more detailed, especially in terms of psychological explanations other than beliefs and desires.

Table 4.4 Categories of explanation and exemplars for Moses & Flavell (1990)

<table>
<thead>
<tr>
<th>Categorisation</th>
<th>Exemplars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desire related</td>
<td>wants</td>
</tr>
<tr>
<td></td>
<td>goals</td>
</tr>
<tr>
<td></td>
<td>intentions</td>
</tr>
<tr>
<td></td>
<td>likes</td>
</tr>
<tr>
<td></td>
<td>needs</td>
</tr>
<tr>
<td>Belief related</td>
<td>false-belief</td>
</tr>
<tr>
<td></td>
<td>ignorance</td>
</tr>
<tr>
<td></td>
<td>lack of perceptual access</td>
</tr>
<tr>
<td></td>
<td>deception</td>
</tr>
<tr>
<td>Outcome</td>
<td>current contents</td>
</tr>
<tr>
<td></td>
<td>absence of original content</td>
</tr>
<tr>
<td></td>
<td>how contents came to be there</td>
</tr>
<tr>
<td>No Explanation</td>
<td>no answer</td>
</tr>
<tr>
<td></td>
<td>did not know</td>
</tr>
<tr>
<td></td>
<td>irrelevant answer</td>
</tr>
<tr>
<td></td>
<td>uninterpretable answer</td>
</tr>
</tbody>
</table>

There are three reasons why there are a greater number of tokens in the Bartsch & Wellman scheme. Firstly, as indicated above, Bartsch & Wellman expected that children’s explanations would reflect backwards reasoning with respect to the scheme reproduced in Figure 4.1. The scheme includes concepts and connections beyond simply beliefs and desires, hence the inclusion of additional concepts in the Bartsch & Wellman scheme. Moses
& Flavell do not mention this possibility. Secondly, in Bartsch & Wellman the tasks were given to adults as well as children. The adults were anticipated (and indeed were observed)\textsuperscript{20} to have a wider repertoire of terms at their disposal. Moses & Flavell tested only children on their task. Finally, there were several different story types in the Bartsch & Wellman experiment. The anomalous desire stories for example (see Table 4.1), where a character states a desire and then violates it, seem likely to invoke a wide range of explanations than do the anomalous belief type stories. Recall that Moses & Flavell's explanation situations were all standard false belief scenarios.

The above three points notwithstanding, Bartsch & Wellman do report that not all of the categories included in their scheme were necessary. 87\% of the belief-based explanations used the term "think", with the rest using the term "know" (excepting one use each of "expect" and "believe"). For desire-based answers, all except one used the term "want" (see also footnote 20).

As noted, both schemes classified responses to explanation questions into roughly the same categories. This classification is, on the surface, quite uncontroversial; it seems obvious that mental explanations should be distinguished from non-mental, and that mental explanations that are belief-based are to be differentiated from those that are desire-based. It also seems obvious that knowing and guessing are tokens of the belief type, while wants and wishes are tokens of the desire type. I intend to argue, however, that a slightly more rigorous system is desirable. It is necessary to know, for example, not only that the belief type and its tokens are different from the desire type and its, but also what it is about these different types and tokens that makes them different. Similarly it is important to know whether there are explanation tokens that, though they indicate that the child has understood certain similar aspects of the situation, fall in quite different intuitive categories. By designing the coding scheme in terms of \textit{how the child might process the task}, rather than on the basis of what intuitions one has about how explanation types cluster, problematic tokens will be more easily classified.

\textsuperscript{20} It isn't entirely clear whether the Bartsch & Wellman scheme was devised before their experiment was carried out, or constructed on the basis of the answers that appeared when the data had been collected. It appears initially that the scheme was created after the explanations had been collected, as the scheme appears in the results section (see Bartsch & Wellman, 1989, p950). However, later this is not so clear. Some categories end up being superfluous, as there are no instances of answers of that type (e.g. Bartsch & Wellman, 1989, p951: "all the desire explanations involved want, except for one use of hope by an adult..." italics in original).
The following section presents a hierarchical coding scheme. This scheme will rely heavily on the two schemes presented above. In addition to specifying the rationale for differentiating explanation tokens, the scheme will consider what a particular explanation token tells us about how the child has processed the task.

4.4. A revised scoring methodology

The coding scheme to be presented here retains a hierarchical structure similar to that of Bartsch & Wellman's (1989) scheme. I present the hierarchy "top-down", considering all the possible answers that a child might give in response to an open ended explanation of a search action (e.g. "why did <agent> look there..?"): the scheme appears in Table 4.5.

4.4.1 Explanation versus no explanation

The initial categorisation to be made is relatively straightforward; whether or not the child offers an answer at all. Clearly a child who says something (anything) has at least processed the question as a request for an utterance. If the child stays silent then one can learn nothing about how the child has understood the situation. Note that this is not the same as saying that we have learned that the child knows nothing about the situation: we simply can't tell. The child could answer with an utterance, but the utterance could be "I don't know". Again, in this case, we have learned nothing.

The first level of the hierarchy then, is whether or not the child has given us any information about how (s)he has processed the task. This is not always straightforward to judge. For example, a child might offer an utterance that is difficult to interpret. Moses & Flavell's coding scheme included a "no explanation" category, which includes responses that are "uninterpretable" or "irrelevant" (1990, p934). Unfortunately, Moses & Flavell did not describe what these responses looked like. Bartsch & Wellman's scheme also included such a category, but here, only "no response", "don't know" and "repetition of the story" responses were included. Responses such as these are all of little use in determining what a child has understood about the situation, and so can be scored as "No Explanation".
### Table 4.5 Revised categorization scheme*

<table>
<thead>
<tr>
<th>Explanation</th>
<th>Mental</th>
<th>Intentional</th>
<th>DESIRE BASED (Object only)</th>
<th>BELIEF BASED (Object + status)</th>
<th>Non-Intentional</th>
<th>Non-Mental</th>
<th>Object only</th>
<th>Object + status</th>
<th>No Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>wants</td>
<td>thinks (that)</td>
<td>drives</td>
<td></td>
<td>item name</td>
<td>item location</td>
<td>no response</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>has goal</td>
<td>believes</td>
<td>emotions</td>
<td></td>
<td></td>
<td>how item got there</td>
<td>don't know</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>intends</td>
<td>knows</td>
<td>traits</td>
<td></td>
<td></td>
<td></td>
<td>repeat story</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tries</td>
<td>guesses</td>
<td>wrong</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>looks for</td>
<td>doesn't know</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>thinks (of)</td>
<td>wrong (about)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>likes/needs</td>
<td>sees/didn't see</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The specific numbered points will be dealt with in the main text

---

### 4.4.2 Mental state explanations

Once it has been established that an explanation has been offered, the next distinction I will draw is whether the explanation uses a mental state term or not. Bartsch & Wellman include the mental/non-mental distinction in their coding scheme, although they cast it as being between psychological and non-psychological explanations. Moses & Flavell also code some explanations as referring to a state of the world: hence their "outcome" category (see Table 4.4). Mental explanations differ from non-mental explanations in where the cause of
the action is located: mental state explanations locate the cause as internal to the agent, non-mental state explanations appear to locate the cause in the outside world.

This distinction needs to be made because it tells us whether there are any restrictions on where the cause of an action by an agent can be located relative to that agent. For example, if we were to find that children always explain actions using mental state terms, we might go on to infer from this some general relationship in the child's causal knowledge between actions and mental states. On the other hand, we might find that in certain circumstances children are happy to explain actions with reference to states or situations that exist in the outside world.

This latter possibility is worth dwelling on briefly. Bartsch & Wellman and Moses & Flavell both had categories of explanation that suggested that the child was locating the cause of the action in the real world. For example, Bartsch & Wellman had a "reality" category, and gave as an example "Because that's where her kitten is...". Moses & Flavell had an "outcome" category, which included the current contents of the location as well as how the contents came to be there. Now although these categories were named so as to suggest the child is locating the cause in the world rather than the mind, it is interesting to consider just how plausible this is. Certainly in the case of the example given by Bartsch & Wellman above, a plausible alternative interpretation might be that the child was offering a desire explanation that was simply not labelled as such: the desire term was left implicit. This highlights the danger of focusing on the presence or absence of certain linguistic tokens, rather than taking into account what the explanation might say about how the child has processed the task. It is important to decide just what it might mean for a child to locate the cause of an action in non-mental terms, as well as identify explanations that appear to be examples of the strategy. This issue is addressed empirically in Chapter 5.

4.4.3 Intentional versus non-intentional mental states

The next distinction I wish to draw is that between intentional mental states and non-

---

21 This is true not only for how the same children go about explanation in various different circumstances, but also for how different children might explain in the same set of circumstances. It is quite possible that studies on young children's explanation will reveal individual differences in the way children assign causes. There is an extensive social psychological literature on adults' style of explanation (see for example Hewstone, 1989), as well as a number of proposals that some psychological disorders are caused (or at least maintained) by "dysfunctional attributional style" (Abramson, Seligman & Teasdale 1978; Hamilton & Abramson, 1983).
intentional mental states. An intentional mental state is one that has "aboutness" (Dennett, 1978b; 1983). That is to say, it is a mental state that is in some sense "directed at" something in the world. Mental states such as emotions or drives can be used to explain actions (e.g. "She did it because she's sad"; "He ate them cos he's hungry..."), but they do so by referring to a state that is entirely internal to the agent: the mental states are not "about" anything. Beliefs and desires, by contrast, are intentionally directed; one has beliefs and desires "about" things (e.g. "Because he wants to keep the lawnmower..."; "Because he thinks the television's too loud...").

The scheme offered by Bartsch & Wellman also made a distinction within the mental state category. However, it was a distinction between "beliefs and desires" on the one hand, and "other psychological states" on the other. There is a problem with this distinction. Of the "other psychological states" in their scheme (see Table 4.3), some are intentional and some are non-intentional mental states. It isn't really clear why these states should be grouped together, other than on the basis of Bartsch & Wellman's commitment to their 1988 reasoning scheme (see Figure 4.1). For example, Bartsch & Wellman's scheme codes pretending among the "other psychological states". Although the precise significance of pretence is disputed (Leslie, 1987, 1994; Perner, 1991, in press; see also Chapter 1), what is clear is that it is an intentional state (pretends are "about" something), and is closely related to belief. In fact, as we saw in Chapter 1, Perner suggests that young children might in fact fail to distinguish between pretence and belief at all (Perner, Baker & Hutton, 1994). My proposal is to draw a distinction on the criterion of intentionality, which is characterizable as "aboutness". This maintains a distinction that classifies beliefs and desires as different from some other mental states, but differentiates these categories in a more systematic manner.

An explanation that uses an intentional mental state is giving more specific information about the cause of the action than one that uses a non-intentional state; namely,

---

Perner (1991, p136), gives a good outline of why emotions, though they can intuitively be "about" things, are not intentional states. This is on the grounds that one can rephrase emotions that are "about" things as emotions being "caused by" things. He goes on to cite Searle's claim that emotions are only ever intentional because they depend on other intentional mental states such as beliefs and desires (Searle, 1983, p29-36). There is another way in which to characterize the "aboutness" of intentionality. It corresponds to a certain logical feature of things that we wish to call intentional; known as referential opacity (see for example Dennett, 1983, p344-345). It is on this criterion that the mental state of perception is viewed as intentional.

98
that although the cause of the action is something internal to the agent, it also has some relationship with a state of the world. For example, Bartsch & Wellman's (1989) scheme placed "likes" in the non-intentional category (Table 4.5, Point 2). Here, however, likes are placed with dispositional intentional states because they are "about” an object. This treatment agrees with that of Moses & Flavell (1990). It might be argued that this misses the idea that "likes" specify a more general, "long-lasting” pro-attitude, but Yuill (1991a, 1991b) has shown that children below an age of about seven are unlikely to attribute preferences to indicate stable characteristics of agents, but more likely as indicating short term pro-attitudes.

4.4.4 Beliefs versus desires

Bartsch & Wellman went on to distinguish beliefs from desires:

Belief-desire reasoning depends on a crucial distinction between beliefs as mental states of conviction versus desires as mental states of disposition” (1989, p948).

Searle (1983) notes that beliefs and desires have in common a propositional content that stipulates their conditions of satisfaction. He goes on to emphasize however that the way in which the states are satisfied (vis à vis the world) is different. The difference is in the direction of causality between the mind and the world, an idea that cropped up briefly in Chapter 1 (see Section 1.6.1). Desire-type (dispositional) intentional states have a mind-to-world direction of causality, in that the mind has a causal influence on the world. Belief-type (epistemic) intentional states have the opposite direction of causality, where states of the world cause states in the mind.

As well as the direction of causality, there is another important difference between epistemic and dispositional mental states, which concerns what they tell us about how the subject has processed the question. An explanation that is concerned with disposition (desire-based explanation), offers less specific information about the cause of the action than does an explanation concerned with beliefs. Desire-based intentional states are concerned only with the agent's relation with a goal object: there is no further reference to anything about the goal object, or its status with respect to the world. Belief-type mental state explanations, by contrast, make reference to the status of the goal object with respect to the
world.  

To illustrate, a desire explanation such as "Because he wants the sweets...", makes reference to a mental state that makes explicit the goal of the action, but not any information about the world status of that goal object. A belief explanation, such as "because he thought the sweets were still in the box...", contains more specific information, relating to a specific characteristic of the goal object. Belief type explanations clearly contain more information than desire type explanations. This is important for deciding what the explanation offered tells us about what resources the child has brought to bear in processing the task. Belief-type explanations reveal a more detailed level of explanation than do desire-type explanations. In the next paragraphs I expand on the remaining belief versus desire classification points which appear in Table 4.5.

The distinction between "thinking of" and "thinking that" (Table 4.5, Point 1) first became an issue when Perner (1989) criticised the experiments presented in Wellman & Bartsch (1988), who then responded in a later issue (Wellman & Bartsch, 1989). The main body of the argument is outlined in Chapter 1 (Section 1.6.2). With respect to explanation, the distinction that Perner made needs to be retained. Although "thinking" is a belief based mental state, think explanations that mention only the object of the thought don't qualify as belief-based explanations because they make no reference to the status of the object in the world. It is on these grounds that in the present scheme, references to "think" that make reference only to the object (e.g. "because he thinks of apples..."), are to be scored as desire-based (dispositional) rather than belief-based explanations.

"Knowing" and "not knowing" (Table 4.5, Point 3) are, intuitively, belief-based mental states. However, there are different ways that knowledge or the lack thereof might be used in an explanation. Consider, for example, the following situation: person A doesn't know that an object is in location x, while person B doesn't know that the object was moved away from location y, person C doesn't know where the object is and person D doesn't know anything at all. All of these people might explain mistaken searches for the object by

---

23 This is linked to the "direction of causality" discussed above. A desire-type state is part of a reason that causes the agent to act toward a particular goal object (mind to world causality). A belief type state is also held to be part of the cause of the action (e.g. Davidson, 1963; see also Chapter 5). However, the belief state is itself caused by an object's status in the world (world to mind causality), and therefore makes reference to the object's status in the world.
referring to a knowledge-based mental state, and all but person D would be categorised, in the Table 4.5 scheme, as having used a belief-based explanation. Person D's knowledge based explanation is crucially different from the others. In terms of the hierarchical scheme the explanation does not even mention an object, and for this reason an explanation of this type cannot be scored as an intentional explanation. This last case is similar to an explanation such as "Because she was wrong", which simply refers to the agent having made an error; such an explanation might even be considered as a description rather than an explanation. Note that this is different to an explanation that states that "She was wrong about where the item was", which is belief-based.

Reference to "perception" (Table 4.5. Point 4) is a controversial category. It was coded by Bartsch & Wellman as a non-intentional psychological state, and while Searle (1983; see also Dennett, 1983) argued that seeing is intentional, Perner (1991) questioned whether young children attribute "seeing" to mean an intentional state. Wimmer & Weichbold (1994), discuss reference to "access to informational conditions" as an explanation type distinct from belief, as part of their CC theory of theory-of-mind development (Wimmer et al. 1988; see Chapter 1). In the Table 4.5 scheme, use of "seeing" or "not seeing" in conjunction with information about the status of the object (i.e. "not seeing that..." or "not seeing where..."). qualifies the explanation as belief-based. In Chapter 6, however, I assess the idea of "access to information" as a form of explanation in itself, following the proposals of Wimmer & Weichbold (1994).

4.4.5 Specificity of non-mental explanations

In Section 4.2.2, I introduced the idea that explanations of actions that referred to non-mental states were subject to mentalistic interpretations. This issue is dealt with more thoroughly in Chapter 5, but it is important to emphasize here that "non-mental state" explanations can vary with respect to how specific they are, just as was shown for mental state explanations in the section above. To illustrate this, I take the example of a "non-psychological" explanation given in Bartsch & Wellman (1989, p950). The explanation is given as "Because her kitty is under there...", and clearly does not refer explicitly to a mental state. However, references to current or past locations of objects (Table 4.5. Point 5), do indicate that the child has specified certain aspects of the cause. The explanation makes clear
an object, and also information about the status of that object with respect to the world, just like the belief explanation example given in the section above.

It is arguable that a non-mental explanation of this kind, and the belief-based explanation ("Because he thought the sweets were still in the box...") have more in common than do the same belief explanation and a desire-based mental state explanation (e.g. "Because he wants the sweets..."). This is in spite of the fact that the latter two share the characteristic of being intentional mental states (Section 4.4.3). Finding examples of explanations that, though they do not make a mental state explicit, seem to indicate that the child is offering the same level of specificity in the explanation, supports the idea that "non-mental state" explanations are in fact implicit references to mental states. This issue is addressed in more detail in Chapters 5 and 6, where the similarity between belief type explanations and certain other explanations is placed within a more theoretical framework.

4.4.6 Relevance of explanation

The final consideration in categorization concerns what Bartsch & Wellman identify as the "relevance" of explanations. Belief & Desire explanations were coded for how relevant they were, with regard to three aspects of the explanation: the protagonist, the topic and the proposition (Bartsch & Wellman, 1989, p951). Bartsch & Wellman assigned a score to an explanation on the basis of how many of the three topics were made explicit. For example, in assessing the topic of the explanation, a child who answers the explanation question with: "I think the kitten is under the chair" would be scored as having offered an irrelevant topic. A child answering an explanation question in such a manner has clearly answered a different question altogether. Interpretation of such "explanations" is extremely difficult. Similarly, irrelevance can turn up in what Bartsch & Wellman term the proposition of the explanation. For example, a child might explain an agent searching in an empty location by saying: "Because she thinks the sweets aren't there". Even though this specifies information about the status of the object in the world, it clearly specifies information that is irrelevant to the Agent's search at that location, and interpretation is therefore difficult.

In the scheme presented here (Table 4.5), explanations are also to be screened for irrelevances. Whereas Bartsch & Wellman scored children for relevance, any irrelevant explanations are to be scored by this scheme as "No explanation", because of the difficulty
in interpretation described above. This is conservative, but ensures that the numbers of belief, desire and other categories of explanation will not be inflated with instances where the route to the explanation might have been very different.

The Table 4.5 scheme will be used as the basis for scoring the explanation experiments that appear in the next chapters. In certain chapters, however, particular explanation categories are emphasized over others. For example, in Chapter 6 I look specifically at "informational access" explanations (Chapter 6, section 6.3). In this and the all other cases, however, the guidelines that have been presented here will inform how the categories used in each chapter are interpreted and discussed.
5.1 Introduction

In this chapter I will introduce a framework for the study of explanation, and present two experiments testing specific proposals within the framework. The framework is based on the philosophical work of Davidson (1963), and is illuminating with respect to the logic behind the existing explanation experiments described in the last chapter. This logic asserted that the study of explanation (e.g. Bartsch & Wellman, 1989; Moses & Flavell, 1990) would provide a measure of false belief understanding convergent with assessing action prediction on the basis of false belief (e.g. Wimmer & Perner, 1983). As well as enabling an appraisal of the explanation logic, Davidson made some quite specific observations about the explanation of action, observations on which the predictions for the current experiments are based. The chapter will begin by setting Davidson's observations in the context of some more recent ideas, from authors within both philosophy and developmental psychology.

5.2 Actions and reasons: a framework for explanation

In this section I present an appraisal of Davidson's (1963) work on explanation. The aim is to use Davidson's analysis as a framework within which to view explanation.

In 1963, Davidson revived the commonsense notion that actions are caused by reasons. A reason for an action, according to Davidson, consists in a belief and pro-attitude (desire) pair. On this view, every action must be caused by a reason, contrary to advice offered to the chicken in the cartoon presented in Figure 5.1 (reprinted from Larson, 1991). If this chicken were in fact to cross the road, then according to Davidson, this action would have been caused by the chicken having wanted to get to the other side, and having believed that by crossing the road she would satisfy this desire.
5.3 Incompleteness of explanation: theory

Davidson’s first observation is that explanations are typically not complete. In his own words:

"[A] primary reason consists of a belief and an attitude, but it is generally otiose to mention both. If you tell me that you are easing the nub because you think it will stop the main from backing, I don’t need to be told that you want to stop the main from backing...". (1963 [reprinted 1968, p82]).

Similarly, considering Figure 5.1, if we are told that the chicken wants to get to the other side, we don’t need to be told as well that she thinks that by crossing the road she will get to the other side. On the basis of this observation, the first prediction we can derive from Davidson’s action-reason framework is that explanations are likely to be incomplete.
5.4 The primacy of desire: theory

In this section I describe a number of positions that share the general view that the concept of desire is an important entity in the development of a fully fledged theory of mind. I begin by continuing with the observations of Davidson (1963).

5.4.1 Davidson's argument for the primacy of desire explanations

Davidson's key observation within his action-reason framework was that, of the two components of the reason, the desire component is primary. Paradoxically, he illustrates this point by referring to an example where a desire that causes an action is in fact omitted in explanation, on account of its primacy:

When we know some action is intentional, it is empty to add that the agent wanted to do it. In such cases, it is easy to answer the question "Why did you do it?" with "For no reason", meaning not that there is no reason but that there is no further reason, no reason that cannot be inferred from the fact that the action was done intentionally; no reason, in other words, beside wanting to do it. (1963 [reprinted 1968, p32] Italics in original).

A number of interesting predictions can be derived from this observation. Firstly, consider a situation where an agent performs an action that fulfils one of that agent's desires (e.g. I search in the fridge to get a beer). According to Davidson, such an action is caused by a reason (belief-desire pair). The reason in this case consists of the desire for the object (beer) and the (true) belief that the location (fridge) contains the object.

The first prediction derivable from Davidson's observation concerning the primacy of desire is this: when a child provides an incomplete explanation for an action (see Section 5.3), it is likely to refer to the desire component of the agent's reason, rather than the belief component.

Consider the situation where an agent searches the wrong location for the object of his or her desire (e.g. I search the cellar for beer, when it is really in the fridge). Just like the successful action described above, this action of searching the wrong location is in Davidson's framework also caused by a reason. The belief-desire pair in this case are the desire for the bait item (beer) and the (false) belief that the bait item is in a certain location (the cellar and not the fridge). Davidson does not mention the case of an action frustrating a desire explicitly, but in the context of the action-reason framework, I will take the strong prediction from the primacy of desire observation to be that desire explanations may still
eclipse belief explanations in such situations. This prediction is at odds with intuition, which
tells us that explaining an action by referring to a desire that has been frustrated by that very
action is inappropriate. However, even "wrong" actions are caused by beliefs *and* desires
and if desires are primary, then they might be preferred to beliefs even in situations where
it seems intuitively inappropriate. This is the important reason for having drawn the
distinction between action explanation measures and backwards belief attribution measures
in the last chapter. Desire explanations, such as those observed in Moses & Flavell (1990;
see Chapter 4. Section 4.2.2), are *correct* explanations of incorrect search, and if Davidson
is right, then one possible reason for a predilection for young children to explain using
desires is that desires are *preferred* as explanations. This is different from the conclusion
drawn by Moses & Flavell (and see next section), which was that desire explanation
reflected the fact the belief explanations were *unavailable*. The resolution of this difference
is one theme that is developed and investigated in the current chapter.

5.4.2 Wellman's argument for a simple desire psychology

Wellman (1990; Wellman & Woolley, 1990) viewed the child's intentional
psychology as developing from a simple desire psychology, with which he credited the 2-
year-old, through a more competent version that includes understanding of belief (but not
belief that conflicts with reality) at age three, to comprehension of false-belief in the 4-year-
old.

The younger child's understanding of desire was assessed in Wellman & Woolley
(1990). 2-year-old children's desire understanding was assessed in two ways: i) appreciation
that desire satisfaction should result in the cessation of action and that desire frustration
should provoke continued action and ii) appreciation that different outcomes of action
should result in different emotions: happiness from success, sadness from failure.

Children were presented with stories where a protagonist was searching for an item.
Protagonists either found the item they wanted, found a different item, or found nothing.
Children were required either to predict whether the character would continue searching or
predict whether the character would feel happy or sad. Older 2-year-olds (mean age = 34
months) showed good evidence of desire understanding on both measures. In a second experiment (1990, Experiment 2) Wellman went on to argue that 2-year-olds (slightly older than those in Experiment 1) could pass desire-based tasks that are extremely similar to the Wellman & Bartsch's belief-based tasks (1988), while failing the analogous belief tasks. For example, in the Not Own Desire task, children were presented with a character who could have had one of two desires (wanting to go swimming or wanting to play with the dog). The child's own desire was elicited ("which would you want to do...?"), and the character was said to have the opposite desire. The child's task was then to predict the character's action. Compare this with the Not Own Belief task (Wellman & Bartsch, 1988: see also Chapter 1. Table 1.1). A character was presented who had one of two beliefs (ball in the garden or ball in the shed). The child's own belief was elicited ("where do you think the ball is...?"), and the character was given the opposite belief. Again, the child's task is to predict where the character would look. Wellman & Woolley (1990) found that the older 2-year-olds were better at the desire version of the task (though better than chance at both).

Wellman's construal of desire primacy differs from that of Davidson (1963). Where Davidson might view a predilection on the part of young children to explain actions in terms of desire as indicating that it is their primary means of explanation (see previous section), Wellman might argue instead that desire is their only means of explanation (see also Moses & Flavell, 1990: this chapter, previous section).

Wellman's position does view desire explanations as emerging before explanations that involve the more complicated notions of belief and false-belief. It is not clear from Wellman & Woolley (1990), however, whether Wellman believed that 3-year-olds would also prefer desire based explanations to belief based explanations in a situation where both were available. In Bartsch & Wellman (1989, Experiment 1), however, it was clear that Wellman expected competence in belief reasoning to be reflected in explanation. He suggested that 3-year-old children may have been able to solve some of the belief-based action prediction tasks presented in Wellman & Bartsch (1988) on the basis of their

---

24 Perner (personal communication), has failed to replicate these findings. In Experiments 5.1 and 5.2 in this chapter, children are presented with a situation similar to the Wellman & Woolley task, which will help clarify the issue (see Section 5.4.4).
understanding of desire alone. For example, in the Discrepant Belief task, he argued that children might have understood "Fred thinks there are only magic markers in the desk, not on the shelf" as meaning something like "Fred likes the desk and not the shelf", consequently predicting Fred's action correctly on this basis. Bartsch & Wellman indicated that explanation would help to resolve this issue, and predicted that if 3-year-old children construe actions only in terms of desires, their explanations would reflect this limitation.

The implications of this are twofold. Firstly, as recommended by Bartsch & Wellman (1989), the child's appreciation of other possible components of the reason should be assessed by the use of prompt questioning to augment the original explanations. For example, asking the child about what the agent believed should tell us whether an aspect of the explanation left implicit in the original explanation is nevertheless appreciated. Secondly, explanations from children in a range of ages will be more informative than restricting the investigation to a narrow band. Both of these measures are exploited in the experiments presented later in this chapter.

5.4.3 Fodor's argument for action prediction based on a desire heuristic

Fodor (1992) proposed a heuristic theory of the child's theory of mind that also reserved a special place for the concept of desire. Fodor's proposal is perhaps the clearest in terms of making explicit why desire is the most important concept for the young child trying to predict action. The reason is that, most of the time, knowing about an agent's desires is enough to allow the prediction of the agent's action; in other words, agents generally do what is objectively conducive to satisfying their desire. This is in turn possible because beliefs are normative (that is, they are supposed to represent what's true) and therefore one can make the assumption that, in most cases the agent will believe what is true. Note that the converse relationship does not hold. Knowing only about what an agent believes will not allow the prediction of action, since one cannot make any useful assumptions about that agent's desires.

Fodor went on to argue that the operation of two ordered heuristics underlies children's performance on the false belief prediction task. As will be recalled from Chapter 3, the first heuristic (H1), is concerned only with an agent's desires while the second heuristic (H2), takes into account the agent's beliefs as well. Younger children accept H1
in any situations where the agent's desires yield a *unique behavioural prediction*. When the action prediction derived from the agent's desire is not unique, the prediction takes into account both the agent's desires and beliefs (H2).

Fodor's view on explanation was that an incorrect search by an agent is viewed by the child as an instance where the prediction of action from the desire heuristic (H1) fails. That is, the desire heuristic predicts search at the location containing the bait, but the action disconfirms this prediction. The child is then in a position much the same as where H1 yields a non-unique, or ambiguous prediction (H1 cannot be used; fails), and in such situations H2 is the heuristic of next resort. H2 is satisfied in the incorrect search scenario only if the agent believes (falsely) that the bait is in the empty location, and so the child ascribes that false belief to the agent as the explanation. This view is more like Wellman's than Davidson's, in predicting belief-based explanations over desire-based explanations in situations where the action frustrates the agent's desire.

It is important to note, however, that Fodor's theory was constructed as a *post hoc* account of performance on theory of mind tasks. With respect to explanation then, Fodor attempted to accommodate the finding of Bartsch & Wellman (1989), which as described in the last chapter (Chapter 4, Section 4.2.1) was that 3-year-olds were able to reason backwards from action to belief. In a later section of this Chapter (Section 5.15), I will return to Fodor's theory and suggest, in contrast to what I've stated above, that it might be closer to Davidson's than to Wellman's.

5.4.4 Changing one's mind: do actions speak louder than words?

Discussion so far has centred on two types of situation: firstly, a type where an agent's desire is satisfied by their action, and secondly, a type where an agent's desire is frustrated by their action. Note that the greatest conflict of opinion has been about the explanations that might be offered in the latter situation. This is because it is in situations where an action fails to satisfy the expressed desire that a realistic clash occurs between.

---

25 Bartsch & Wellman's (1989) study is dealt with in greater detail in Section 5.6.2. Fodor may well have based his prediction on the belief attribution results also presented by Bartsch & Wellman (e.g. 1989, Expt 2). As far as the open ended explanation results are concerned, the authors reported that desire explanations exceeded belief explanations, but it is not clear for which of Bartsch & Wellman's three story types this was true, as no breakdown was presented.
desire and belief explanation. However, there is more than one way in which an action might fail to satisfy an expressed desire.

The false belief situation (Wimmer & Perner, 1983), is characterized by a character searching in a location that previously contained a bait item, but is now empty. Consider, though, a situation similar to the original false belief situation that is simpler in one respect, but more complex in another.

There are different bait objects in two separate locations. A character who has expressed a desire for one of these bait objects searches in the wrong location, that is, the location that contains the other bait object. This new situation preserves a critical part of the false belief situation that provoked interest: an agent performing an action that frustrates that agent's desire. Now, however, in term of explanation, the action that has been performed, despite frustrating the Agent's expressed desire, is compatible with an alternative desire. In addition, it is compatible with the original desire coupled with a false belief about the location of the originally desired bait item. A situation such as this provides an especially good example of desire conflicting with belief in explanation; it amounts to asking whether actions speak louder than words. Recall that the strong prediction from Davidson's primacy of desire observations (Section 5.4.1), was that desire explanations would eclipse belief explanations where they were in conflict.

Alternatively, an action that frustrates the expressed desire, but that is nevertheless consistent with an alternative desire, may reveal a preference for desire explanation more readily than the more usual false-belief situation. Thus, a weaker more refined prediction to be derived from Davidson's observations, is that desire explanations may only eclipse belief explanations where the action is consistent with some other plausible desire. Experiment 5.1 tests both of these primacy of desire predictions.

5.5 Reality explanations as justifications

The final prediction I derive from Davidson (1963) picks up on some points that were discussed in Chapter 4, in the section on explanation coding (Section 4.4). The concern is with a class of non-mental explanation encountered in both Bartsch & Wellman (1989) and Moses & Flavell (1990): reality or outcome explanations. This class refers to cases of explanation where a child refers to some fact about the world in order to explain
an action, seemingly with no reference to a mental state (e.g. "[he looked there]...because the kitten is under there...").

One possible interpretation mentioned in Chapter 4, was that a child using this type of explanation might really be locating the cause of the action outside the agent, with no indication that the action is mediated by a mental state. I argued in the last chapter that this interpretation is difficult to make sense of. To recap briefly, I stressed that it was important to determine just what it might mean for an action to be explained externally to the Agent.

One characteristic of so-called "external" explanations is that they never refer to just any external state of affairs. For example, when I look in the fridge and tell you that I have done this because there's beer there, the explanation is related to a plausible reason: the desire for beer coupled with the belief that there is beer in the fridge. The important point is that, in spite of the wide range of possible external explanations (e.g. that the light is on, that the fridge is painted white etc.), a very narrow range of external facts are cited. This makes implausible the notion that children using reality explanations conceive of actions in a fundamentally different way to those who offer mental state explanations.

There is a subtle way in which this form of explanation might actually be more sophisticated than suggested above. For this possibility I return briefly to Davidson (1963):

Straight description of an intended result often explains an action better than stating the result was intended or desired. "It will soothe your nerves" explains why I pour you a shot as efficiently as "I want to do something that will soothe your nerves", since the first in context of explanation implies the second; but the first does better, because, if it is true, the facts will justify my choice of action. (1963 [reprinted in 1968, p84]).

Explanations of this type then, may really be mental state explanations, which are offered as not only explanation for the action but also as justification for the action. The fact that the fridge contains beer justifies and explains my looking there; I was right to look in the fridge given that I wanted beer.

If this analysis of "reality" explanations is correct, then we would expect certain things to follow. Firstly, as indicated above, reality explanations should also be reliably close to a plausible implicit desire (or indeed a desire already expressed in the course of the situation). It would not be expected that a wide range of references to the external world be observed. If prompted for the agent's desire or belief after the initial explanation, then the implicit desire/belief should be offered. Secondly, it would be expected that such
explanations will occur very rarely in situations where no desire is plausible, such as an agent searching an empty location.

5.6 Existing evidence

Before moving to my own experiments, which test the predictions introduced above, I outline how existing evidence from explanation studies bears on some of the proposals that have been introduced in foregoing sections. During this section, I refer quite extensively to the two prominent explanation studies described in the previous chapter: Bartsch & Wellman (1989, Experiment 1) and Moses & Flavell (1990). Here, however, rather than highlighting the methodology and rationale, I focus on how the results reflect on each set of predictions in turn.

5.6.1 Incompleteness of explanation: evidence

Bartsch & Wellman (1989) referred explicitly, albeit only briefly, to the possibility that explanations would be incomplete. They included in their explanation procedure prompt questions designed to assess whether subjects appreciate elements of the explanation that have been left implicit in the original response.

As far as their results were concerned, Bartsch & Wellman reported that the number of complete (that is, specifying an Agent's belief and desire) explanations was very low: between 1% and 3% of all unprompted explanations. There were other indications in the data that indicated that the explanations were incomplete. Firstly, Bartsch & Wellman provided separate analyses for the unprompted and complete explanations. The number of explanations (over all categories) increased when prompting was taken into account, suggesting that there was information not made explicit in the initial unprompted explanations (1989, graph p953). Secondly, Bartsch & Wellman reported that their raters showed no disagreement over whether psychological explanations were to be scored as beliefs or desires (1989, p952). If explanations had been complete (i.e. referred to both an Agent's belief and their desire), then presumably some disagreement would have occurred.

On the basis of these observations, there is good reason to believe that Davidson's contention, that explanations are incomplete, is correct. In Experiment 5.1, this conclusion will be retested.
5.6.2 The primacy of desire: evidence

Before looking at the specific explanation experiments, it is worth noting some anthropological evidence that adults view desires as important determinants of actions. D'Andrade (1987) interviewed people about their ideas concerning why people do things. The picture from the protocols that he discussed supported Davidson's contention that actions require reasons. In addition, assuming that the views presented by D'Andrade were at all representative, the commitment to desire was also apparent: an agent who wanted to do a particular action would do that action: if they did not:

"There would have to be a reason why the person didn't do it if they wanted to do it... It wouldn't be that they just wouldn't do it" (subject quoted in D'Andrade, 1987, pp 131-133).

A non-overridden desire to perform an action then, seems to cause that action.

Moving to the explanation experiments, Bartsch & Wellman reported significantly more desire responses than belief responses among the Unprompted explanations (1989, Experiment 1, p954). This is consistent with the general primacy of desire prediction made in Section 5.4.1. However, Bartsch & Wellman did not present a breakdown of how stories of each particular type (neutral, anomalous desire, anomalous belief) were explained. They focused, instead, on the general competence children showed in offering "relevant belief and desire explanations" across all story types. In another paper, however (Wellman & Woolley, 1990), Wellman did mention that some young children only ever referred to desire in their explanations (6 out of 23, as compared with one who only ever referred to belief).\(^{26}\) This pattern would seem to be compatible with a strong primacy of desire notion, but Wellman interprets this as evidence that these children only had desire explanations, rather than preferred their use (see Sections 5.4.1 and 5.4.2).

Wellman & Woolley (1990; see also Chapter 1 and this chapter, Section 5.4.1) provided a situation in which an action, though it frustrated an expressed desire, was directed at a location consistent with an alternative desire. They did not present explanation data, but their results seemed to indicate that children did not change the character's desire on the basis of the outcome of the action. For these children, actions did not speak louder

\(^{26}\) Note that this difference is in fact not quite significant (binomial, p=0.062).
than words. However, the situation presented in Wellman & Woolley was not an explanation situation, and the desire expressed by the agent was contained within a larger desire (to take the bait item to school). It is not clear then that this finding is relevant to actions speaking louder than words in explanation (e.g. as in Section 5.4.4).

The other explanation experiment described in Chapter 4, Moses & Flavell (1990), is slightly more helpful. Recall that their experiment was concerned only with false-belief-based search, and so some information relevant to the Davidson predictions, for situations where belief and desire explanations clash, can be derived. The results showed that desire explanations were indeed preferred: 46% of explanations were desire-based as opposed to 12% belief-based. However, the Moses & Flavell stories involved two locations which both contained objects. The action that frustrated the agent's desire was, for these stories, consistent with the desire for the object in the searched location. It is not clear, however, whether Moses & Flavell took into account the content of the desire attributed by the children. Moses & Flavell gave as an example of a desire response, an explanation that maintained the original desire, and their result could thus be taken as support for the strong primacy of desire prediction (Section 5.4.1). Alternatively, given that there was not good evidence that the children were able to attribute a false belief to the Agent at all, the preponderance of desire explanations may indicate that desire was the only means by which these children could explain (Section 5.4.2). This again makes clear the need to include prompt questions after the initial explanation has been evoked, and to assess the explanations of children who are capable of attributing a false belief.

5.6.3 Reality explanations

In Bartsch & Wellman (1989), it was not possible to tell how many explanations referred to reality, or for which stories reality explanations were offered, although it is clear that there were cases of such explanation.

In Moses & Flavell (1990), 21% of explanations were outcome explanations. All of these explanations were given in cases where the action had frustrated the desire expressed by the agent, but there are two difficulties in drawing conclusions from this fact. Firstly, Moses & Flavell did not present the children with stories where an agent's desire was satisfied by the action, so no comparison can be made. Secondly, the example that Moses
& Flavell gave for this explanation type was of a child referring to the current contents of the container. As mentioned above, Moses & Flavell used two filled locations. Given that no subsequent prompt questions were asked to establish whether the child had attributed a desire for the current contents of the container to the agent (but left this implicit), it is unclear what interpretation can be made for the reality explanations in Moses & Flavell. The evidence described in this section has shed some light on the predictions that have been derived from the primacy of desire proposals of several theorists (Section 5.4). The predictions have fared relatively well, but there remain doubts about the precise interpretation of some of the results, especially with respect to the explanations of younger children. This is due to the fact that the experiments have had slightly different aims. In Experiment 5.1, I attempt to provide a systematic test of children's explanations for simple search actions under different conditions.

5.7 Experiment 5.1: Introduction and summary of predictions

Experiment 5.1 was an investigation of young children's explanation of simple search actions, within Davidson's reason-action framework. Of particular importance was how children would resolve the clash of desire and belief explanation constructs under various conditions.

The situations used in this experiment were intended to be computationally extremely simple. This was in order to dispense with some of the preamble events that accompany more standard false belief situations, such as the character hiding the bait in the first place, the character being out of the room and the whole transfer event. It was hoped that restricting the situation in this way would lead to a smaller range of explanations being offered by subjects than in previous experiments (see Chapter 4, Section 4.3.2). Instead, explanations were expected to be more focused on the categories of interest: beliefs and desires. In addition, in order to address the issue of desire explanation being the primary rather than only means of explanation, the age range of children used was expanded to include ages at which competent belief reasoning could be expected.

Children watched situations involving an agent and two containers. In each situation the agent made an utterance that expressed a desire for a bait object that was contained in one of the locations. The other location was either filled or empty. The agent then
approached one of the two locations. The children were required to explain the agent's search action (initial explanation), and were then prompted for the agent's desires and/or beliefs. The first prediction within the framework was that initial explanations would be incomplete; they would refer to a belief or desire, but not both (see Section 5.3).

In some cases, the location at which the agent searched was consistent with the desire that the agent expressed. The second prediction then, was that successful actions such as this would be explained initially with reference to the agent's expressed desire (see Section 5.4.1) or with reference to the contents of the location (see Section 5.5). The desire prompt would be expected to reveal any desire left implicit in the initial explanation.

In other cases the agent performed an action inconsistent with the expressed desire. This action could be directed at either a filled or empty location. In cases where the location towards which the inconsistent action had been directed was empty, the strong prediction derived from Davidson was that desire explanation would be retained. Belief prompting would reveal the implicit component of the reason. By contrast, the predictions derived from Wellman and Fodor were that belief explanation would be favoured under these circumstances (for Wellman so long as prompt questions confirmed belief explanations were available). Alternatively, the inconsistent action could be directed at a filled location. The refined prediction from the primacy of desire observations suggested the explanation under these circumstances would be especially likely to refer to desire, but a desire consistent with the outcome of the action performed rather than the original desire (see Section 5.4.4).

5.8 Experiment 5.1: Method
5.8.1 Design

The explanations that children give for an action performed by an agent were examined under different conditions. In each story the child was presented with two distinct containers, one or both of which could contain an item. The children were then introduced to a protagonist, who made an utterance that expressed a desire, and then acted by searching at one of the two locations. Two factors were manipulated: firstly, whether one or both of the locations contained a desirable object, and secondly, whether the action that followed the agent's utterance was consistent or inconsistent with that utterance. The design was therefore a 2(no. of locations filled) x 2(consistency of action) factorial design. The first
factor was between subjects, yielding two conditions: **1 Location** and **2 Location**. Children were randomly assigned to one of these two conditions. The second factor was within subjects, with each child receiving one **Consistent** action and one **Inconsistent** action story. Order of presentation for the two stories was counterbalanced, with half the children in each condition receiving the consistent action story first and the other half receiving the inconsistent action story first.

The dependent variable of primary interest was the type of explanation that the child offered for the action. Immediately after the action had been performed by the agent, the children were asked the explanation question: "why did <agent> look there?". The children's answers were scored according to the scheme presented in Section 5.8.6, which is based on the guidelines discussed in Chapter 4.

Following Bartsch & Wellman (1989), the initial explanations were supplemented with answers to prompt questions. The prompting depended on the children's answers to the initial explanation question. In cases where the child made no reference to either the agent's beliefs or desires in their initial explanation, two prompt questions were asked, a desire prompt: "what did <agent> want?", followed by a belief prompt: "where did <agent> think the <desired object> was?". In cases where the child referred only to the agent's desires in the initial explanation, then just the belief prompt was asked. Similarly, if just belief information was referred to in the initial explanation, then the desire prompt alone was used. Finally, children were also asked to recall the agent's utterance.

**5.8.2 Subjects**

The subjects were 38 pre-school children all attending the same nursery school in West Finchley in London. Most of the children's first language was English, but it was not a requirement for inclusion in the study. It was assumed that children meeting the control question requirements (see procedure, Section 5.8.5) had sufficient English to understand the stories. The children were assigned randomly to one of the two conditions. Testing continued until a total of 32 children had met the inclusion criteria. This resulted in 16 children in each condition, with the age and sex distributions in the conditions as follows: **1 Location** (nine boys and seven girls, aged between 41 and 58 months [mean=49 sd=5.6]) and **2 Location** (six boys and ten girls, aged between 42 and 57 months [mean=49 sd=5.4]).
5.8.3 Stories

The stories used in each condition differed in the number of locations filled. In the 1 Location condition, Location A (the box) was filled with bananas (or chocolate) and Location B (the pot) was left empty. In the 2 Location condition Location A was filled with bananas (or chocolate) and Location B was filled with apples (or ice-cream). In the Consistent story in each condition, the Agent then approached the container that matched the desire expressed in the utterance. In the Inconsistent story in each condition, the Agent approached the other container. The stories were thus as shown in Table 5.1.

5.8.4 Materials

The two containers were the same for each story. They were a matchbox, which had a flower pattern on the outside, and a small black plastic pot (5cm tall, 3cm diameter), with a lid. There were two specific sets of materials, one set for each story, that consisted of bait items and a protagonist. The bait items were made with Plasticine, and were small enough to fit inside either container. Materials in Set one were two small bananas and two small apples as bait, and a female plastic play-person (5cm tall), dressed in white, as the protagonist. In Set two, the bait items were a small block of chocolate and an ice cream, while the protagonist was a male play-person dressed in red. The sessions were recorded using a Sony portable tape recorder.

5.8.5 Procedure

Each child was tested individually. The child sat down opposite the experimenter at a low table. The child was told that (s)he was going to play some short games, and that the games would be recorded. The children were shown the two containers, inside and out, and told that the game would be played with these containers. Each story was divided into three phases: the Item phase at the beginning, followed by the Agent phase and finally the Question phase. These phases are dealt with in turn.

**Item phase:** The children were told to watch carefully as the experimenter placed bait in the containers. For each child, the materials were always used in the same order: Set one first and Set two second. Material set was thus counterbalanced across conditions. For the one location conditions, the bananas (or chocolate) were placed in the matchbox, and
the pot was left empty. For the two location conditions, the bananas (or chocolate) were placed in the matchbox and the apples (or ice cream) were placed in the pot. The containers were closed and the children were asked to recall the contents of each pot. Any errors at this stage resulted in the experimenter repeating the initial phase, reminding the child to watch carefully. If the child made a further error, then that child was excluded from the study.

Agent phase: The children were introduced to the protagonist and told that her name was "Chloe". ("Sammy" in the case of materials Set two). She was placed on the edge of the table, midway between the two containers, facing the child. The subjects were told that Chloe was going to say something and that they should listen carefully. Chloe then expressed a desire for the bait item in one of the locations (see Stories section above). The child was asked to repeat what Chloe had said. This procedure was then repeated. If the child had any difficulty repeating the protagonist's utterance, the experimenter repeated this procedure a maximum of four times. If the utterance had not been reproduced at least once by this time, the child was dropped from the study.

After hearing and repeating the protagonist's utterance, the children were told to watch. The experimenter acted out the protagonist's action with the model. The protagonist was made to walk forward to a point midway between the two containers, and then turned and approached one of the containers. The protagonist was left next to the container and the experimenter remarked on where she had gone (e.g. "Look, she's gone to the <container> ").

Question phase: The child was then asked the questions described in the design section. Firstly: "why did Chloe look there..?": then if required, the prompt questions were asked (see design): "what did Chloe want..?" followed by "where did Chloe think the <desired item> was..?". Finally, the memory question was asked: "what did Chloe say right at the beginning..?". The answers to the questions were noted during the experiment, and the whole session was also recorded, so that any answers that were difficult to hear or that were omitted from the written record could be checked later.

The children were then told that was the end of the first story, and asked if they would like to hear another. The materials from Set one were removed and the Items phase for the second story commenced. The procedure for the second story was identical to that used for the first.
Table 5.1 Summary of stories used in Experiment 5.1

<table>
<thead>
<tr>
<th></th>
<th>1 Location</th>
<th>2 Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(box - bananas; pot - empty)</td>
<td>(box - bananas; pot - apples)</td>
</tr>
<tr>
<td><strong>Consistent Action</strong></td>
<td>Chloe says - &quot;I want to eat some bananas&quot;. Look, Chloe has gone to the box.</td>
<td>Chloe says - &quot;I want to eat some bananas&quot;. Look, Chloe has gone to the box.</td>
</tr>
<tr>
<td><strong>Inconsistent Action</strong></td>
<td>Chloe says - &quot;I want to eat some bananas&quot;. Look, Chloe has gone to the pot.</td>
<td>Chloe says - &quot;I want to eat some bananas&quot;. Look, Chloe has gone to the pot.</td>
</tr>
</tbody>
</table>

5.8.6 Categorization of initial and prompted explanations

The explanations were then scored into one of the following four categories:

1. **DESIRE-BASED**: Explanations that referred explicitly or implicitly to a person's desire and the object of the desire. These explanations were most often explicit, making use of the term wants, but might equally well have left the desire term implicit, by mentioning either another mental state such as trying or simply referring to an object as being a goal.

Recall that in the last chapter, it was argued that the crucial factor that defined dispositional (desire-based) explanations was that the explanation make reference to the object of an agent's mental state. In order to meet this demand, imposed by the discussion in Chapter 4, the explanation had to contain reference to the object of the desire, as well as the implicit or explicit indication that this object was a goal, to be scored in this category.
2. BELIEF-BASED: Explanations that referred to the person's beliefs. Usually the term thinks was used, but again, it was possible that beliefs would be referred to in other ways, perhaps by using another term, such as knows. To be scored in this category, the belief explanation had to specify both an object, and something about the object (e.g. "because he thinks the apples are in there.", was fine, but neither "cos he doesn't know.." or "cos he's thinking about apples" were accepted).

Again, as we saw in Chapter 4 (Section 4.4.4), the crucial factor that distinguishes belief-based mental state term explanations from desire-based mental state term explanations is the specificity with respect to some relationship between the object and the world. It is this element that the scheme presented above attempts to preserve. "Thinking of" type explanations were not scored in the belief-based category, because of the reservations expressed in the last chapter about the status of these explanations. Recall that although an explanation such as "because he's thinking of apples" makes explicit reference to belief, it's status as a belief explanation is questionable, because it links the agent to an object only, rather than to a proposition about an object. An explanation that only specifies the object of an agent's mental state, is closer to the dispositional type explanations described above than it is to the epistemic states that this category was attempting to capture. "Ignorance" explanations were also not scored here. Again, the reasons for this were described at greater length in the last chapter. To recap, we saw that an explanation type such as "not knowing that...", is epistemic, because it specifies both the object of the belief and information about that object. By contrast, simply referring to an agent's simply "not knowing" or "being wrong", does not specify either of these pieces of information, and accordingly is scored differently.

3. REALITY: Explanations that referred only to a fact about the world such as an object being or not being in a specific location, with no indication that the object or location was a goal.

This category is controversial, as we saw in the introduction to this chapter. It was included
in an attempt to resolve the issue of "external" causes for action outlined in Section 5.5.

4. OTHER: Any other explanations. This category included any explanation that could not be confidently placed in one of the above three categories. Examples include drives that did not specify an object ("he's hungry"), as well as where there is no response, or a "don't know" response.

The final category was intended to capture those explanations that did not clearly fall into one of the categories of interest.

Note that many of the specific categories introduced in the scoring scheme presented in Chapter 4 do not appear in this scheme. This is because the additional explanation types (such as not having seen a transfer of an item or having been tricked) appear only in the context of the more standard false-belief scenario, where there are more degrees of freedom with respect to the explanation. The situations used in this experiment were deliberately devised to simplify the range of explanations (see Section 5.7).

Subsequent Categorization: Once the explanations had received their initial categories, those in the desire category were further scored to ascertain the exact content of the desire explanation. Desire explanations were divided into those that had a content that was the same as the initial utterance and those that had a different content. For example, if the character in the story had expressed a desire for bananas, and the explanation for her action was that she wanted bananas, then the explanation received a Dsame categorization. Conversely, if the character had expressed a desire for bananas and the explanation for her subsequent action was that she wanted apples, then the explanation received a Ddiff categorization. This procedure concluded the scoring of the initial unprompted explanations.

The answers to the desire prompt question were categorized exactly as described above. For the belief prompt question, the answer was scored as a true belief if it attributed to the agent a belief that was consistent with reality. If on the other hand the child attributed to the Agent a belief that did not match the real situation, then this answer was scored as a false belief.
5.9 Experiment 5.1 Results

A total of eight children were excluded from the study for failing control questions. Three of these eight had failed to recall the locations of the bait items, while the other five were unable to repeat the protagonist's initial utterance after more than four attempts. These children finished the story that they had attempted, but no data were recorded.

5.9.1 Initial Explanations

Children's explanations were all incomplete. Of 64 explanations offered, there were no cases where the explanation referred to the Agent's belief and desire. 75% of explanations referred to either a belief or desire, and of the 25% that were scored into the other categories, about half were cases where children offered no explanation at all.

The first hypothesis that will be dealt with here is that concerning the interpretation of reality explanations. The reason for dealing with this first is that on Davidson's interpretation, reality explanations are really instances of implicit mental state justification, where the desire and/or belief is left implicit. If there is no evidence for this contention, then these explanations can be scored into the "Other" category.

The incidence of these explanations was very low, which makes their interpretation more difficult, but the importance of that interpretation less significant. There were very few cases of children referring to a fact about the world in order to explain an agent's action in this experiment: a total of six, evenly distributed between the 1 Location Consistent and 1 Location Inconsistent conditions. The fact that these explanations appear at all in the inconsistent condition is of note, because such an explanation (e.g. "He looked there because the box is empty"), in fact implies a desire for nothing.

The expectation expressed in the introduction was that if these explanations were implicit references to desire and/or belief, then they would be more frequently invoked in order to explain actions that had been successful (i.e. consistent actions) than to explain unsuccessful actions (where the action is directed toward an empty location), because in consistent actions, a plausible desire is readily available. Clearly, such a small number of instances of this explanation type does not allow resolution of this issue, but in any case, no great differences to the overall pattern of results would occur if these explanations were recategorized. Consequently, in the subsequent analyses, these reality explanations will be
included in the "Other" category. I will return to the issue of reality explanation in the discussion (Section 5.10.2).

The frequency of each explanation category for each story in both conditions is presented in Figure 5.2. Inspection of Figure 5.2 shows that the pattern of explanation appears to have been affected by the various story conditions. In the conditions where the action matched the desire expressed in the utterance (1 and 2 Location Consistent), the predominant explanations were in terms of the desire expressed in the initial utterance.

![Initial Explanations](image)

**Figure 5.2** Frequencies of each explanation category in Consistent and Inconsistent stories in 1 and 2 Location conditions, Experiment 5.1.

The pattern changes somewhat in the conditions where the action did not match the initially expressed desire (1 and 2 Location Inconsistent). Under these circumstances, the initial desire was not preserved in the explanation. However, the pattern was not the same...
over both inconsistent actions. In the 1 Location condition, where the action was not consistent with another desire, beliefs were the most common explanation. Where the action was consistent with a different desire (2 Location Inconsistent), the most common explanation was in terms of a desire that matches the action performed, rather than the initially expressed desire. In the following sections, this pattern of explanation is examined statistically.

For actions consistent with the Agent's expressed desire, the predominant explanation offered by children in both the 1 and 2 Location conditions was to cite the Agent's expressed desire. The patterns of explanation in the 1 and 2 Location conditions (represented in Figure 5.2) were compared ($X^2[N=32, df=2] = 3.33$, two-tailed $p>0.15$). There were no significant differences in explanation then, between the 1 and 2 Location Consistent actions.

For inconsistent actions the explanations were more diverse. The frequencies of each explanation type in the 1 and 2 Location conditions (represented in Figure 5.2) were compared, and there was a significant difference between the conditions ($X^2[N=32, df=3] = 12.94$, two-tailed $p<0.01$). Partitioning the chi-squared table to discover the location of the difference(s) revealed a difference in the frequency of Dsame versus Ddiff explanations ($X^2_{1, cor}[N=32, df=1] = 6.40$, two-tailed $p<0.02$) and a further difference in the frequency of Desire(same and diff) versus Belief explanations ($X^2_{2, cor}[N=32, df=1] = 5.88$, two-tailed $p<0.02$). The difference in the frequency of reason (desire and belief) versus all other explanations was not significant ($X^2_{3, cor}[N=32, df=1] = 0.66$, two-tailed $p= N.S.$).

The next analyses look specifically at the difference in explanation that results from manipulating whether or not the action is consistent with the expressed desire. This factor was manipulated within subjects, and the analyses therefore focus on the concordance between children's explanations across the Consistent and Inconsistent actions in both 1 and 2 Location conditions.

The effect on the explanation of whether the action was consistent with the expressed desire was assessed by looking at the concordance between each child's response on the Consistent and Inconsistent versions of the action, for each level of the Location factor. The pattern of responding apparent in Figure 5.2 suggests that children offered predominantly desire(same) explanations for the consistent actions, but not for the...
inconsistent actions, and that the proportion of belief explanations was highest in the 1 Location Inconsistent condition. Table 5.2 shows the consistency of subjects with respect to desire explanations in the 1 Location condition. Each child was classified according to whether their explanations referred to a desire or not for each of the two action types. There are thus four possible classifications: desire explanation for both types of action, desire explanation for the consistent action and any other explanation for the inconsistent action and vice versa and any other explanation for both actions.

Table 5.2 Children's consistency in the use of desire explanation in 1 Location condition in Experiment 5.1

<table>
<thead>
<tr>
<th>Inconsistent Action</th>
<th>Consistent Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Desire Explanation</td>
</tr>
<tr>
<td>Desire Explanation</td>
<td>4</td>
</tr>
<tr>
<td>Other Explanation</td>
<td>6</td>
</tr>
</tbody>
</table>

The majority of children were consistent with respect to whether or not they offered a desire explanation. However, there were six children who offered a desire explanation for only one of their two stories, and all these children offered the desire in the story where the action was consistent with the utterance. This is a significant difference (McNemar test of change, binomial, p=0.016), indicating that desire explanations were less likely to be offered for inconsistent actions than for consistent actions. All the desire explanations in the 1 Location condition referred to desires that were the same as the expressed desire (Dsame categorization).

The children in the 1 Location condition were also categorized according to their consistency as regards belief explanation. This categorization is presented in Table 5.3, which shows children's consistency in the use of desire explanation for Consistent and...
Inconsistent actions in the 1 Location condition. There were fewer belief than desire explanations overall. However, it can be seen that children were specific in their use of belief explanation, since only two children used belief terms to explain both consistent and inconsistent actions. Five children used belief explanation for the inconsistent action and another explanation for the consistent action, while no children show the reverse pattern (McNemar test of change, binomial, p=0.031). This indicates that belief explanations were more likely to be used when an inconsistent action was to be explained than when a consistent action needed explaining.

Table 5.3 Children's consistency in the use of belief explanation in the 1 Location condition in Experiment 5.1

<table>
<thead>
<tr>
<th>Inconsistent Action</th>
<th>Consistent Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Belief Explanation</td>
</tr>
<tr>
<td>Belief Explanation</td>
<td>2</td>
</tr>
<tr>
<td>Other Explanation</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.4 shows children's consistency in the use of desire explanation in the 2 Location condition of Experiment 5.1. The concordance pattern observed for the 1 Location condition with respect to the use of desire explanation was not evident in the 2 Location condition. Most children here were consistent in their explanation, and most explained both consistent and inconsistent actions with reference to a desire. There were a small number of children who explained just 1 of the actions with a desire, but these children did not explain the same action using the desire (McNemar test of change, binomial, p=0.312).
Table 5.4 Children's consistency in the use of desire explanation in 2 Location condition in Experiment 5.1

<table>
<thead>
<tr>
<th>Consistent Action</th>
<th>Desire Explanation</th>
<th>Other Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent Action</td>
<td>Desire Explanation</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Other Explanation</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5.4, however, does not allow us to assess the content of the desire attributed by the children. Recall that one of the predictions derived from Davidson's observations was that young children might appeal to a desire other than the desire expressed by the agent in cases where the action performed was consistent with an alternative desire.

From Figure 5.2, it can be seen that all the desires offered by children in the 1 Location condition were desires that matched the utterance. However, the picture for the 2 Location condition is different. In Table 5.5, the eleven children who offered desire explanations for both stories have been further categorized according to the content of the desire that they attributed (i.e., whether the desire attributed had been categorized as Dsame or Ddiff).

Table 5.5 Children's consistency with respect to the content of desire in 2 Location condition in Experiment 5.1

<table>
<thead>
<tr>
<th>Consistent Action</th>
<th>Dsame</th>
<th>Ddiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent Action</td>
<td>Dsame</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Ddiff</td>
<td>9</td>
</tr>
</tbody>
</table>
Inspection of Table 5.5 shows that although children in the 2 Location condition were concordant in their use of a desire term generally across the two types of action, their answers were discordant when the content of the desire is taken into account. Nine of the eleven children offered a different desire content for each of the two actions within the 2 Location condition. All nine of these offered a Dsame explanation for the consistent action and a Ddiff explanation for the inconsistent action (NcNemar test of change, binomial, p=0.002).

The concordance analyses then, confirmed the picture that emerged from the initial analysis: individual subjects showed differences in explanation dependent on whether the action had been consistent with the expressed desire or not, and these differences were mediated by whether or not the location toward which the inconsistent action was directed contained an alternative object.

5.9.2 Prompted explanations

After the initial explanations had been gathered, the experimenter prompted the children for the agent's desire or belief, or both, depending on the initial explanation they had offered.

Desire prompts were given when the initial explanations failed to make reference to the agent's desire. The answers were scored as described in Section 5.8.6. The answers to this prompt question were used to augment the initial explanations, with the following results.

Consistent Actions: In the 1 Location condition only six children required prompting; after the prompt, all children who hadn't previously done so attributed to the Agent a desire that matched the expressed desire (and the action that had been performed). In the 2 Location condition just two prompts were used, both of which were answered in terms of a desire matching the expressed desire. The three children who offered a "reality" explanation offered Dsame answers to the desire prompt.

Inconsistent Actions: In the 1 Location condition, 12 desire prompts were used. Ten of these (including the three children who had offered "reality" explanations) resulted in children attributing a desire that matched the expressed desire, one child attributed a general
desire to the agent (e.g. "to eat something..."), and one child attributed a desire that was inconsistent with the expressed desire, but consistent with the action that the agent had performed ("she wants the pot..."). In the 2 Location condition, five desire prompts were used. All of these resulted in children attributing a desire that matched the initially expressed desire, but was inconsistent with the action. Note that this increases the number of D_same responses in the 2 Location condition. These data are presented in Figure 5.3.

Inspection of Figure 5.3 shows that once prompting has been taken into account, there is still a marked difference in responding between the conditions. The proportion of children preserving the original desire (D_same response) is significantly greater in the 1 Location condition than in the 2 Location condition ($X^2_{corr}[N=31, df=1] = 6.59$, one-tailed $p<0.01$).

![Desire Prompts](image)

**Figure 5.3** Children's answers to the desire prompt question for the Inconsistent action in each condition in Experiment 5.1.
Belief prompts were given when the child's initial explanation made no reference to the agent's beliefs. The belief prompt always concerned the location of the object of the desire previously attributed to the agent, either in the initial explanation or in the desire prompt.

Consistent Actions: 14 belief prompts were used in the 1 Location condition, all of which resulted in a consistent belief being attributed to the agent. That is, the agent was assumed to believe that the item that (s)he wanted was where (s)he thought it was (and in these cases also where it really was). 16 belief prompts were also required in the 2 Location condition, and again, all of these were answered by attributing a true belief to the agent (a belief consistent with the expressed desire and the real location of the object of the expressed desire).

Inconsistent Actions: Nine belief prompts were required in the 1 Location condition. Of these, three children attributed a belief that was consistent with the agent's action (but not with the real location of the desired object, i.e. a false belief). Five children attributed a belief that was inconsistent with the agent's search action (but consistent with the real location of the desired object, i.e. a true belief). The last child attributed a general belief to the agent: "she's thinking of bananas". In the 2 Location condition, 14 belief prompts were required. Five of the belief prompts were asked in the context of the initially expressed desire, which had been preserved in either the initial explanation or in answer to the desire prompt. Two of these resulted in answers that attributed a belief consistent with the agent's action in the context of the originally expressed desire (false belief attributions), while the other three resulted in a belief attribution that was inconsistent with the agent's action in the context of the attributed desire, but consistent with the true location of the object (true belief explanations). The remaining Nine belief prompts in this condition were asked in context of the desire consistent with the agent's action, and all of these resulted in belief attributions that were consistent with both the desire that had been attributed and the real location of the object of that desire (true belief explanations).

The above details are summarized in Figure 5.4. The graph shows the proportion of false belief responding in each of the two conditions once prompting is taken into account. Note that in the 2 Location condition the "true belief" column contains cases of both appropriate and inappropriate appeals to true belief. This is because some (nine) of these
true belief attributions are in the context of the original desire having been changed in the initial explanation, and so the true belief is consistent with the desire that has been attributed. In the other cases (three), the true belief is offered in the context of the originally expressed desire, and so can be considered inappropriate.

**Figure 5.4** Children's answers to the belief prompt question for the Inconsistent action in each condition in Experiment 5.1.

When belief prompting has been taken into account, the difference in appeal to false belief that was demonstrated between the conditions for the initial explanations is still apparent: the proportion of false belief answers was significantly greater in the 1 Location condition than it was in the 2 Location condition ($X^2_{\text{cont}}[N=30, \text{df}=1] = 3.17$, one-tailed $p<0.05$).
5.9.3 Memory for the initial utterance

Children were also asked to recall the agent's initial utterance. The utterance was divided into two components: object and the desire. Children were scored with 2 points if they mentioned both the correct object and mental state (desire). If they mentioned only the object, or the object and the wrong mental state (e.g. think), they received 1 point. If they said nothing, or mentioned only the mental state, or recalled the wrong object, they received a score of nil. The mean scores and standard deviation for the memory question in each condition appear in Table 5.5 below.

Scores on the memory question were quite good, with most children scoring at least 1 point, indicating that they correctly remembered the object of desire. This was true even for the 2 location inconsistent condition, where many children had explained the action with reference to a desire that was not the same as the utterance.

Table 5.6 Means and standard deviations of scores (out of 2) on memory for initial utterances in Experiment 5.1

<table>
<thead>
<tr>
<th></th>
<th>1 Location</th>
<th>2 Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent Action</td>
<td>1.75 (0.45)</td>
<td>1.63 (0.72)</td>
</tr>
<tr>
<td>Inconsistent Action</td>
<td>1.56 (0.73)</td>
<td>1.44 (0.81)</td>
</tr>
</tbody>
</table>

A 2(number of locations filled) x 2(consistency of action) split plot ANOVA was performed on these data, which revealed no significant differences in memory between the conditions, and no interaction (F<1 in all cases). Children's memory for the initial utterance was no worse for consistent than inconsistent actions, or for 2 versus 1 Location conditions.
5.10 Discussion

The results of Experiment 1 support most of the predictions that were derived from Davidson (1963). I deal first with the notions of incompleteness and justification, and then move on to the central issue of primacy of desire. I will also discuss a possible problem in the design of Experiment 5.1, which will be addressed in a second experiment.

5.10.1 Incompleteness of explanation

In terms of the reason-action framework introduced in Section 5.2, the explanations offered by children were all incomplete. There were no cases of children offering both the belief and desire component of the reason to explain the action. 75% of all the explanations referred to either the belief or the desire of the agent. This proportion is greater than that reported in both Bartsch Wellman (1989), where only 48% of explanations invoked a belief or desire, and Moses & Flavell (1990), where the proportion was 58%.

These differences indicate that the attempt to streamline the range of explanations offered (see Section 5.7) was successful. There is good evidence here then, that children were operating quite specifically within an action-reason framework.

5.10.2 Reality explanations as justifications

The experiment did not, however, provide good evidence for Davidson's suggestion that references to reality, or facts about the outcome of the action, might be made by way of justifying the action in the context of an implicit desire. There were some instances of reality explanations for the consistent action stories, where the answer to the desire prompt question was consistent with a justification interpretation. However, these responses were few, and there were as many instances of reality explanations being offered for the inconsistent actions; in these cases the answer to the desire prompt was not consistent with the justification interpretation. The prediction that reality explanations would appear predominately in the context of successful actions was not supported.

5.10.3 Primacy of desire explanation relative to belief explanation

Actions consistent with the agent's utterance were almost exclusively explained with reference to the desire expressed in the utterance. This is precisely in accord with the first
prediction made on the basis of Davidson's primacy of desire observation.

When the action performed was inconsistent with the expressed desire, the explanation offered depended on the contents of the location toward which the action had been directed. Where the location of the search was empty (1 Location condition), the clash between belief and desire explanations pits a desire that has been frustrated by the action against a false belief about the desired object's location. The results showed that the strong prediction derived from Davidson was not upheld in this case. There were significantly fewer explanations in terms of desire than there had been for the consistent actions. Instead, children explained these actions by appealing to the agent's beliefs. When the location searched contained a bait object, however (2 Location condition), the clash between belief and desire is more complex. There are two possible desires: the desire expressed in the utterance (which has been frustrated by the action) and the desire based on the outcome of the action. There are also two possible beliefs: a (false) belief concerning the location of the object of the originally expressed desire and the (true) belief concerning the location of the object in the searched location. The attenuation of desire-based responses observed for the 1 Location inconsistent action disappeared in the 2 Location condition. However, the desire attributed to the agent in these cases was the changed desire, not the desire that the Agent had originally expressed. Significantly fewer belief explanations were offered for the Inconsistent action in the 2 Location condition than had been for the same action in the 1 Location condition. The 2 Location condition then, provides support for the weaker, refined prediction derived from the primacy of desire argument in Davidson (1963).

Wellman's simple desire psychology was intended only to operate until the third year. From examination of Bartsch & Wellman (1989), we saw that Wellman predicted that reasoning backward would show that children viewed actions in terms of both desires and beliefs. The results of the 1 Location condition were consistent with Wellman's expectation that false belief explanations would be offered in cases where desires were frustrated by actions. This replicates Bartsch & Wellman (1989), who described their results as indicating false belief understanding not evident in previous studies (1989). Indeed, Wellman went on to conclude on the basis of the other experiment in the 1989 paper that false belief in an
explanation scenario is easier than false belief attribution in a prediction scenario.\textsuperscript{27}

The 2 Location condition in this experiment also involves an action that frustrates a desire, and unlike Wellman & Woolley (1990), children appeared to rate the Agent's action as indicating a changed desire, rather than a false belief coupled with the original desire. The reason for this difference is probably that the desires expressed in this experiment were not placed within a wider desire (the desire to take the item elsewhere) as they were in Wellman & Woolley's study (see Section 5.4.2).

As it stands, Fodor's hypothesis, like Wellman's, can explain the false belief responses evident in the 1 Location condition; his explanation has been described already in Section 5.4.3. As has been noted however, Fodor does state explicitly that the theory is a post hoc theory. Therefore with the 2 Location condition, it is not clear whether Fodor's post hoc "prediction" holds. On one reading, the action performed is not that which H1 would have predicted on the basis of the expressed desire, and so H1 can be considered to have failed. On this reading the child should offer the H2 based belief explanation. This is Fodor's prediction as described in Section 5.4.3.

In a sense however, the result observed for the 2 location situation is more in the spirit of Fodor's theory than his own post hoc prediction. Fodor argues that the child is operating with desire based heuristics because they are simple but quite reliable, and that computing beliefs is costly. The action performed by the agent in the 2 Location condition is inconsistent with the H1 prediction based on the expressed desire, but consistent with the (presumably H1) prediction from a desire different to that expressed. It does not seem unreasonable to suggest that H1 need not have failed in this situation. Attributing a desire on the basis of H1 is a simpler strategy than resorting to H2, which involves preserving the original desire and calculating the content of the belief that must have accompanied it. This modification of Fodor's heuristic theory will be incorporated into a model of explanation processes to be presented in the general discussion. I will postpone further discussion until that point.

\textsuperscript{27} As we saw in Chapter (Section 4.2.1), Wellman was later to modify his position (Wellman 1990). In the general discussion of this chapter I will outline and evaluate his revised position.
5.10.4 Problems with the primacy of desire interpretation

One alternative to the accounts offered above is that children gave primarily desire explanations for the actions that they observed because they were presented with desire information alone. This possibility is unlikely, given that for at least one story (1 Location Inconsistent) there were relatively few desire responses. Nevertheless, the suggestion is tested in Experiment 5.2, where children will be presented with initial information about beliefs rather than desires. If the pattern of results obtained in Experiment 5.1 is due to the initial desire information, then the pattern of results expected in Experiment 5.2 will be very different; instead of predominantly desires, a much higher frequency of belief responses should be observed in all conditions. A cross experimental comparison will be made to assess the relative frequency of belief and desire explanation. The remaining predictions for Experiment 5.2 are the same as those stated for Experiment 5.1.

5.11 Experiment 2 Method

5.11.1 Design

The design for the second experiment was the same as for Experiment 5.1. The same factors were manipulated, only this time in the context of an initial utterance that expressed the agent's belief. The details of this appear in the design section of Experiment 5.1. Again, children were randomly assigned to one of the two between-subjects conditions. Counterbalancing and questioning were as for Experiment 5.1.

5.11.2 Subjects

The subjects were 40 preschool children all attending the same nursery school in West Finchley in London. The school was the one used for Experiment 5.1, but no children who took part in Experiment 5.1 were used in Experiment 5.2. Testing continued until a total of 32 children had met the inclusion criteria. This resulted in 16 children in each condition, with the age and sex distributions in the conditions as follows: 1 Location (nine boys and seven girls, aged between 41 and 57 months [mean=49, sd=4.8]) and 2 Location (ten boys and six girls, aged between 42 and 59 months [mean=50, sd=6.0]).
5.11.3 Stories

The stories used in each condition only differed from those used in Experiment 5.1 in the initial information given (about agent's beliefs rather than desires). The stories appear in Table 5.7.

5.11.4 Materials

As Experiment 1.

5.11.5 Procedure

Again, this was almost exactly the same as Experiment 1. This time, however the agent phase involved the agent making a specific statement about his/her belief. The statement was in the form "I think the <bait object> is in the <location>", and the statements were always accurate. However, the actions that followed were not always consistent with the location mentioned in the utterance: sometimes the action was towards the other location.

The rest of the procedure was as detailed in Experiment 5.1.

Table 5.7 Summary of stories used in Experiment 5.2

<table>
<thead>
<tr>
<th>Belief</th>
<th>Match</th>
<th>1 container full</th>
<th>2 containers full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(box - bananas</td>
<td>(box - bananas</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pot - empty)</td>
<td>pot - apples)</td>
</tr>
<tr>
<td></td>
<td>Sammy says -</td>
<td>Sammy says -</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;I think the bananas are</td>
<td>&quot;I think the bananas</td>
<td></td>
</tr>
<tr>
<td></td>
<td>in the box&quot;.</td>
<td>are in the box&quot;.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Look, Sammy has</td>
<td>Look, Sammy has</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gone to the box.</td>
<td>gone to the box.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sammy says -</td>
<td>Sammy says -</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;I think the bananas are</td>
<td>&quot;I think the bananas</td>
<td></td>
</tr>
<tr>
<td></td>
<td>in the box&quot;.</td>
<td>are in the box&quot;.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Look, Sammy has</td>
<td>Look, Sammy has</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gone to the Pot.</td>
<td>gone to the Pot.</td>
<td></td>
</tr>
</tbody>
</table>

139
5.12 Results

A total of six children were excluded from the study for failing control questions. One of these six had failed to recall the locations of the bait items, while the other five were unable to repeat the protagonists initial utterance after more than four attempts as described in the procedure for Experiment 5.1.

Again there were very few reality explanations (eight) which were evenly distributed among the conditions. As in Experiment 5.1, this makes interpretation difficult, but the effect of recategorization in any case would be small. As in Experiment 5.1, reality explanations will be scored as "Other".

As in Experiment 5.1, children’s explanations were typically incomplete; there were no instances of children offering an explanation that might have been considered complete from a total of 64 explanations. 73% of explanations referred to a belief or desire. This figure is consistent with the picture from Experiment 5.1.

5.12.1 Initial explanations

![Initial Explanations](image)

**Figure 5.5** Number of children in each explanation category for each action type in each condition of Experiment 5.2.
As before, the responses to the initial explanation questions were scored by the experimenter into categories, and these data are presented in Figure 5.5. As Figure 5.5 indicates, the consistent actions again evoked D_{same} explanations, and though the inconsistent actions evoked more diverse explanations than in Experiment 5.1, the same general pattern (Beliefs in 1 Location, D_{diff} in 2 Location) is evident.

For Consistent actions, the pattern of explanation for the 1 and 2 Location consistent actions did not differ ($X^2_{\text{corr}}[N=32, \text{df}=3] = 2.68, p=N.S.$)\(^{28}\) This is the same as observed for Experiment 5.1. For Inconsistent actions, unlike the equivalent comparison in Experiment 5.1, there was no significant difference ($X^2_{\text{corr}}[N=32, \text{df}=3] = 3.26, p=N.S.$). This would appear to be on account of there being both slightly fewer belief explanations and slightly more D_{diff} explanations in the 1 Location condition in Experiment 5.2, than in the equivalent conditions in Experiment 5.1.

As in Experiment 5.1, the effect on explanation of whether the action was consistent with the expressed desire was also assessed by looking at the concordance between each child’s response on the Consistent and Inconsistent versions of the action, for each level of the Location factor.

Table 5.9 shows the consistency of subjects with respect to desire explanations in the 1 Location condition. As in Experiment 5.1 (Table 5.2), each child was classified according to whether their explanations referred to a desire or not for each of the two action types. It can be seen from this table that most children were consistent with respect to whether or not they offered a desire explanation. However, there were six children who offered a desire explanation for only one of their two stories, and all these children offered the desire in the story where the action was consistent with the utterance.

This is a significant difference (McNemar test of change, binomial, $p=0.016$), indicating, as seen for Experiment 5.1, that children were more likely to offer a desire explanation for the Consistent action.

---

\(^{28}\) The sparse matrix in this case indicates that the significance level may be affected by a poor match between the sampling distribution $X^2$ and the $\chi^2$ distribution (see e.g. Cochran, 1954). However, the test yielded non-significance, and the conclusion to be drawn on the basis of the test is not controversial; the data in Figure 5.2 clearly indicate no difference between the pattern of explanations in the 1 and 2 Location conditions.
The children in the 1 Location condition were also categorized according to their consistency as regards belief explanation, as in Experiment 5.1. The results of this categorization appear in Table 5.10. There were fewer belief responses in this Experiment (contrast Table 5.3 from Experiment 1). However, it can be seen that children were specific in their use of belief explanation, since the only children to use belief explanation did so only for the inconsistent action. This however, is not quite a significant difference (McNemar test of change, binomial, p=0.062), but it is close and in the right direction.

### Table 5.9 Children's consistency in the use of desire explanation in 1 Location condition (Experiment 5.2)

<table>
<thead>
<tr>
<th>Consistent Action</th>
<th>Desire Explanation</th>
<th>Other Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent Action</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

### Table 5.10 Children's consistency in the use of belief explanation in the 1 Location condition (Experiment 5.2)

<table>
<thead>
<tr>
<th>Consistent Action</th>
<th>Belief Explanation</th>
<th>Other Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent Action</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>
Table 5.11 shows the children's consistency with respect to desire explanation in the 2 Location condition of Experiment 5.2. It can be seen from the table that most children were consistent in their explanation, and most explained both consistent and inconsistent actions with reference to a desire. This is the same pattern evident in the 2 Location condition in Experiment 5.1 (see Table 5.4). Only one child explained just one action with a belief, and there is consequently no evidence that the conditions are different (McNemar binomial, p=N.S.).

<table>
<thead>
<tr>
<th>Consistent Action</th>
<th>Desire Explanation</th>
<th>Other Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent Action</td>
<td>Desire Explanation</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Other Explanation</td>
<td>1</td>
</tr>
</tbody>
</table>

The ten children who offered desire explanations for both stories were further categorized according to the content of the desire that they attributed (i.e. whether the desire attributed had been categorized as Dsame or Ddiff). Frequencies resulting from this categorization appear in Table 5.12 (compare Table 5.5 from Experiment 5.1).

As in Experiment 5.1, although children were consistent with respect to the general type of explanation offered in the 2 Location condition, they were inconsistent with respect to the content of that explanation. Seven of the ten children changed the content of their desire explanation according to which type of action they were explaining. All seven of these offered a Dsame explanation for the consistent action and a Ddiff explanation for the inconsistent action. (McNemar test of change, binomial, p=0.008).
Table 5.12 Children’s consistency with respect to the content of desire in 2 Location condition, Experiment 5.2

<table>
<thead>
<tr>
<th>Consistent Action</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D\textit{same}</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>D\textit{diff}</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

5.12.2 Prompted Explanations: Experiment 5.2

![Graph: Desire Prompting](Figure 5.6 Children's answers to the desire prompt for the Inconsistent action in 1 and 2 Location conditions in Experiment 5.2.)
Desire prompt, consistent actions: Most children responded to this question by offering a desire that matched the desire expressed in the utterance. This was as observed in Experiment 5.1.

Desire prompt, inconsistent actions: Response to the desire prompt are shown in Figure 5.6. The pattern is the same as for the desire prompt question of Experiment 5.1 (compare Figure 5.3). There proportion of Dsame responses in the 1 Location condition was significantly higher than that in the 2 Location condition ($X^2_{cor} N=30, df=1 = 3.49$, one-tailed $p<0.05$).

Belief prompts, consistent actions: Children responded for the most part with true beliefs in both conditions. Belief prompts, inconsistent actions: Similarly to Figure 5.4 of Experiment 5.1, Figure 5.7 shows the proportion of false belief responding in each condition.

**Figure 5.7** Children's answers to the belief prompt for the Inconsistent action in 1 and 2 Location conditions of Experiment 5.2.
Unlike the results of Experiment 5.1, however, the difference here is not significant ($X^2_{cor}[N=32, df=1] = 0.51$, one-tailed $p=N.S.$), indicating the proportion of belief responses was no greater in the 1 Location condition than it was in the 2 Location condition (contrast Figure 5.4).

5.12.3 Memory for the initial utterance

Children were also asked to recall the Agent's initial utterance. For this second experiment, the utterance was divided into three rather than two components: the object, the mental state (thinks) and the location specified in the utterance. Children were scored with three points if they mentioned all three components correctly. Two points were given for the correct combination of object and location (missing the mental state), while one point only was given for any of: the object only, or the object and wrong mental state, or the correct location only. If they mentioned the wrong object or the wrong location or mentioned only a mental state or said nothing, they received 0 points.

<table>
<thead>
<tr>
<th>N=16</th>
<th>1 Location</th>
<th>2 Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (sd)</td>
<td>Consistent Action</td>
<td>1.94 (1.81)</td>
</tr>
<tr>
<td></td>
<td>Inconsistent Action</td>
<td>1.63 (1.26)</td>
</tr>
</tbody>
</table>

The mean scores and standard deviation for the memory question in each condition appear...
in Table 5.13. Memory for the initial utterance in this second experiment was not particularly good. (contrast Experiment 5.1, Section 5.9.3). Children rarely scored more than 2 out of 3 points. A 2 (no. of locations) x 2 (consistency of action) split plot ANOVA performed on these data revealed no main effects of either number of locations filled or consistency of action, and no interaction (F = 1.23, 1.0 and 1.0 respectively). The most common errors here were for children to recall the think statement as a statement about desires.

5.12.4 Cross Experiment comparison

The pattern of responding was very similar in both experiments. There is no reason to believe that the overall pattern of explanation in Experiment 5.2 was any different to that in Experiment 5.1 (For Consistent actions, X^2[N=64, df=3] = 1.76, two tailed p = N.S.; for Inconsistent actions, X^2[N=64, df=3] = 0.89, two tailed p = N.S.).

5.13 Experiment 5.2 discussion

The general pattern of results in Experiment 5.2 was very similar to that observed in Experiment 5.1. Moreover, an important consideration is that the differences that do exist are not on account of an increase in belief responding in Experiment 5.2. If anything, one reason the exact pattern of explanation observed in Experiment 5.1 was not replicated was that fewer belief explanations were offered in one condition. Overall, there was no difference between the experiments in terms of the relative proportion of belief and desire explanations.

5.13.1 Belief explanations in 1 Location condition

The most notable difference occurred in the number of (false) belief explanations offered in the 1 Location condition. Recall that in Experiment 5.1 there were a large number of (false) belief explanations offered for the Inconsistent action in the 1 Location condition. These were supplemented by a number of children responding with a false belief answer to the belief prompt, having explained the action initially with a desire. For both these

---

29 Again, this is based on a sparse matrix. However, as argued in footnote 28, the test in any case yielded non-significance. Inspection of Figures 5.2 and 5.5 reveal that there is clearly no question of the interpretation that resulted.
measures, the number was lower in the 2 Location condition than it was in the 1 Location condition. In Experiment 5.2, by contrast, there was no significant difference between the conditions for this action. Children were more inclined to explain the story by referring to the agent's expressed desire (Ds\textit{same}) in the initial explanation. In addition, on being prompted, fewer attributed a false belief than had done so in Experiment 1.

One possible reason is that in Experiment 5.2, unlike Experiment 5.1, the children were required to attribute a belief that was inconsistent not only with reality, but also with the agent's initial utterance. In Experiment 5.1, the belief was inconsistent with reality, but not with the agent's utterance, which contained no indication of what they believed. This possibility, that there were two sources of information that contradicted the inference that the child could have made from the misplaced action as to the content of the agent's belief, is consistent with the "performance" theories of the child's difficulty with false belief. I will return to this observation in Section 5.15.4.

5.13.2 Memory for the initial utterance in Experiment 5.2

Another difference between the results of Experiment 5.1 and Experiment 5.2 was in the success that children had with the memory question. In terms of the scoring, absolute comparison is difficult, because there were more elements in the belief utterance than in the desire utterance. It is possible that the difference in children's recall simply reflects the extent to which the desire-based utterances provided less overall stress on the child's memory than the belief-based utterances.

However, looking more closely at the errors that children made, a more subtle interpretation can be supported. A common error Experiment 5.2 was that children recalled the utterance as being an \textit{utterance about desires rather than about beliefs}. 12 children in the second Experiment made one or more such errors, as opposed to none in the first Experiment, where the errors were usually of omission. My interpretation of this is based again on the notion that desires are closely causally related to action, as Davidson argues. Children who found it difficult to recall the initial utterance used the action as a cue to recall it, but because actions are more closely causally related to desires than they are to beliefs, the children's inference is that the initial utterance must have expressed a desire.
5.14 General Discussion

The results of the two experiments, taken together, demonstrate firstly that children do appear to reason about action with respect to a framework based on the attribution of beliefs and desires. About 75% of all explanations offered made reference to an agent's belief of desire. Secondly, evidence has been gathered that supports some of the contentions made about explanation in Davidson (1963). Explanations were typically incomplete; very few explanations referred to an agent's beliefs and desire. Where an agent's actions were consistent with that agent's utterance, desires were almost exclusively used to explain the action, irrespective of whether the agent had explicitly stated the desire or stated their belief.

In situations where agents performed actions that frustrated their expressed desire, the outcome depended on the exact conditions under which the desire had been frustrated. Crucially, these conditions depended on reasoning about desire. If the action performed was inconsistent with the expressed desire, and not consistent with some other plausible goal, then desire based explanation was rarer. Instead, the explanations were commonly in terms of the agent's (false) belief. However, even under these conditions, there were children who retained desire explanation, but who augmented this with correct answers to a belief prompt. The strong prediction derived from Davidson (1963; see also Section 5.4.1) required that the latter pattern would predominate, and therefore there is only very limited support for this prediction. However, the weaker prediction introduced in Section 5.4.4 was strongly supported. Actions inconsistent with the expressed desire but consistent with an alternative desire provoked children to attribute a new desire to the agent in their explanations. For these children the agents' actions speak louder than their words.

5.15 A simple model for explanation

During the discussion of Experiment 5.1, I outlined briefly how notions of desire primacy other than Davidson's fared in the light of those results. Here, I wish to pick up the discussion again, and focus on reasons why explanation should be desire-based. Though I have derived the predictions tested in this chapter from Davidson (1963), his observations are simply that. They offer no reason as to why desires should be the primary means of explanation. By contrast, the desire-based heuristic theory presented in Fodor (1992) does offer a reason. I intend in this section to develop Fodor's ideas to form a simple model of
children's explanation, compatible with CC approaches to theory of mind. As well as
drawing on Fodor (1992), I will incorporate some specific ideas on the information
processing demands of theory-of-mind calculations detailed by Leslie (Leslie, 1994, in press;
Leslie & German, in press; Leslie & Thaiss, 1992; Roth & Leslie, 1991; see also Chapters
1 & 3).

I will recap briefly the basics of Fodor's (1992) theory once again, to show how it
provides the basis for the explanation model.

5.15.1 Fodor's heuristics in prediction

According to Fodor, knowing only about an agent's desires is enough for the
prediction of action, but knowing only about that agent's beliefs is not. This in turn is
because beliefs are normative: they are "meant" to be true. If I assume that a given agent has
true beliefs, then I can predict what that agent will do so long as I know what desires (s)he
has. Desires are not normative, however, so one cannot assume anything about them as
reliably as one can for beliefs. Therefore, if I know only about an agent's beliefs then I
cannot predict action. This difference amounts to the desire component of the reason having
a closer causal relationship to action than the belief component.

Fodor argues that younger children make use of this fact when they are faced with
action prediction problems. Their computational resources are smaller and so they do not
routinely compute belief to predict action, but instead predict on the basis of a desire-based
heuristic H1. This strategy is less reliable than computing beliefs, but also less costly. In
some situations, the action prediction derived from the agent's desires fails (for example,
it could be ambiguous; see Chapter 3). Under these circumstances the children are forced
into using H2, which results in them computing the agent's belief. The explanation model
I wish to develop assumes a similar ordering in the computations.

5.15.2 Explanation

The close causal relationship between desire and action is built into the model by
drawing on the distinction made in the adult causal reasoning literature between the "cause"
of an event and a "condition" that enables that event (Cheng & Novick, 1991). Cheng &
Novick illustrate their point with an analogy:
In answer to the question, "What caused the plane to crash?", aviation-accident investigators are unlikely to reply, "The gravitational pull of the earth". Rather they are likely to reserve the title of "cause" for factors such as pilot error. [Gravity, they might say, was merely a condition that enabled the crash to occur. (Cheng & Novick, 1991, p84).

I propose to make this distinction for beliefs and desires respectively with respect to action (desires are the causes of actions, beliefs the conditions that enable actions). I will attempt to elaborate on this proposal during the exposition of the model.

The starting point is to assume that heuristics (such as Fodor's H1 and H2) used in action prediction also operate over explanation. When an action requires explanation, this is in the first instance a search for a desire/cause. H1 generates candidate desire/causes that would have predicted the action, from (at least) two sources of information:

1. Information from the outcome of the action (e.g. information about the location/object toward which the action is directed).
2. Prior information about the agent's desire (e.g. an utterance made by the agent or perhaps information about the agent's likes/dislikes preferences etc.).

When the information from these sources is consistent, explanation is solely desire-based. The action is explained with reference to the desire specified by the two sources of information. When the sources provide different information, the belief-based heuristic requires the calculation of the agent's belief/condition. On the basis of this calculation, an explanation is generated. The actual explanation offered depends on the results of the belief calculation. I will assume that belief calculation involves the usual performance resource requirements posited by CC versions of theory-theory (See Chapter 1). Most importantly, there will be a prepotent default belief content: that the agent believes what is true (see also Section 5.4.3). In cases where the default true belief content is attributed, the desire/cause will be offered. In some cases however, the prepotent response will be rejected and a different belief content will be attributed. Under these circumstances, the belief/condition
is hypothesized to block the desire cause explanation.\[31\]

5.15.2.1 Consistent Actions

In simple situations where the actions are consistent with the utterances made beforehand, such as the Consistent actions used in Experiments 5.1 and 5.2, the calculation of the desire cause is simple. Both sources of information described above suggest the same candidate desire under H1. Since there is no conflict, this desire is attributed as the cause of the action. Figure 5.8 shows an example of this type of event.

\[31\] In terms of the analogy with causes and conditions in a plane crash, the case where a true belief is calculated is analogous to a case where gravity is operating - it is the default. In the case of a false belief however, it is instead as if gravity had been found to be absent. This deviation from the default would block the usual cause attribution.
5.15.2.2 Inconsistent Actions

In cases where the utterance made by the Agent does not match the action performed, the candidate causes from the two sources of desire information conflict.

In the case where the agent wants the bait object but searches an empty location, the "outcome" source of desire information offers no plausible candidate cause. The utterance information, however, does suggest a possible cause. The inconsistency between the outcome of the action performed and the candidate desire suggested by the utterance provokes the calculation of the belief content. The prepotent belief content (truth) is not consistent with the candidate desire here, but a false belief about the location of the bait item would have enabled the utterance-desire to have caused the action. This belief/condition, being a deviation from the default contents, is ascribed as the explanation. This case is represented in Figure 5.9.

![Inconsistent Actions: 1 Location](image)

**Figure 5.9** The explanation model in the case where the action is inconsistent with desire information, and the location searched contains no object.
The inconsistency again provokes the calculation of the enabling belief/condition. This time, the prepotent belief content (truth) enables a plausible desire/cause, and so the child attributes that desire as the cause. An example of this type is represented in Figure 5.10.

![Inconsistent Actions: 2 Locations Diagram](image)

Figure 5.10 The explanation model in the case where the action is not consistent with the desire information and the location searched contains another object.

5.15.3 Initial belief utterances

It is not entirely clear what role the initial utterance plays when it expresses belief rather than desire information. In terms of the model, I have argued that it can have played two roles: setting up candidate desire causes but also affecting the belief content calculation for inconsistent actions.

Firstly, the desire heuristic H1 is capable of setting up a candidate desire from the utterance information even when the utterance did not explicitly express a desire. Recall that
the memory for the initial utterance was poor in Experiment 5.2, and the errors suggested that at least some children had taken the initial utterance as a cue for desire.

Secondly, the difference in the pattern of explanation for the 1 Location inconsistent action between Experiments 5.1 and 5.2 may be a result of the belief utterance information having been taken into account to some extent. Children were less likely to attribute a false belief to the agent for the 1 Location inconsistent action in Experiment 5.2 than they had been for the same task in Experiment 5.1. When children in the second experiment calculated the belief condition for this action, there was more evidence in favour of the prepotent default belief content: evidence from the true location of the item and also evidence from the utterance expressed by the agent. In the first experiment, by contrast, the initial utterance contained no belief information to affect the belief calculation.

5.16 Conclusions

5.16.1 When do actions speak louder than words?

The model presented above offers a simple account of the way that explanation might proceed in young children. Implicit in the model is the notion of desire primacy; explanation is primarily about locating the desire/cause; the explanations for the consistent actions, where the sources of information about desire tell the same story, were almost exclusively desire-based. In situations where the child is faced with conflicting information from the utterance and the outcome sources, then we can meaningfully ask when actions speak louder than words for children.

In the 1 Location task, the action and utterance information conflict, but only the utterance information yields a plausible desire. The only desire consistent with the action in this case is in fact a desire for nothing, or perhaps a desire to find out what is there. As noted, this strategy was very occasionally adopted, with a few children explaining the action in terms of the agent wanting to see what was inside. For these children, the utterance information was discarded. For most children however, the utterance based desire is the only plausible cause, but because the default belief contents would not enable this cause, the actual belief condition must be calculated. Words are retained in the face of conflicting actions in this case because the desire/cause enabled by the default belief content is not plausible.
In the 2 Location task the action and utterance information once again conflict, but this time both suggest plausible desire causes. The action utterance conflict is resolved here on the basis of the calculation of belief content. The action desire information is consistent with the default (true) belief condition, whereas the utterance desire is not. There is no problem now, in attributing a true belief content, which results in the utterance desire being discarded. Actions speak louder than words in this situation, because the action is consistent with a desire/cause enabled by default belief contents.

5.16.2 What of "strong" primacy of desire

In Section 5.4.1 Davidson's desire observations were used to derive the strong prediction that desire explanations would eclipse belief explanations when they were in conflict, but that appropriate belief information would be left implicit. The results of the experiments provided a few examples of such responding. Some children indeed offered the originally expressed desire as the explanation even when it was inconsistent with the action. Some of these children went on to offer an appropriate false belief response to the belief question, while others completed the explanation inappropriately.

The performance of these children is anomalous as far as the model presented in Section 5.15 is concerned. According to the model, the inconsistency evident in the action with respect to the original desire should provoke the calculation of the belief, which blocks the desire as cause. Why should children persist in attributing a desire that has been frustrated as the cause? The question is especially relevant when one considers that at least some children were perfectly capable of attributing the belief content when asked to do so. In the next chapter I present an experiment that is concerned with just this issue.
CHAPTER 6

Factors affecting the explanation of search action by older children

6.1 Introduction

In the last chapter the distinction between causes and enabling conditions was drawn to characterize the different causal roles that desires and beliefs play in action. The general goal of this chapter is to look more closely at the notion of offering a condition (hereafter belief/condition) versus a cause (hereafter desire/cause) as the explanation for an action.

6.1.1 General goal

It was proposed in Chapter 5 that where desire/cause explanations and belief/condition explanations were in conflict, the desire/cause explanation is under some circumstances preferred as the explanation. This preference for desire explanations over belief explanations was clearest where the desire/cause was enabled by a true belief/condition (see Chapter 5, Figure 5.10). However, as noted in Section 5.16.2, the desire/cause was preferred by some children in situations where a false belief/condition enabled the cause. These instances of desire/cause explanation are particularly interesting because the children concerned were often (though not always) able to complete the desire/cause explanation, when prompted, by offering the appropriate false belief. The possible reasons for this information not featuring in the initial explanation are investigated in this chapter. I will attempt to decide between two possibilities:

i) The children offered cause explanations because of a conceptual limitation in understanding false belief (see Section 6.3).

ii) The children offered cause explanations because they had not calculated the belief/condition (see Section 6.4).
6.1.2 Competence limitations on the calculation of enabling conditions

In Chapters 1 and 5, I outlined Wellman’s views on a "simple desire psychology". The first possibility for why desire/cause explanations might be maintained in situations where the action in fact frustrates the desire can be traced to those views, and subsequent criticism of those views. Wellman stated in 1989 that he expected children's belief understanding to be reflected in their explanation (Bartsch & Wellman, 1989, Experiment 1, p948). Desire-based answers are offered up to the point where the children can interpret the action as resulting from a false belief. Though Wellman argued that the 1989 paper showed that explanation was "easier" than prediction, he later modified this conclusion (1990; see also Chapters 1 & 5). One reason that children might offer desire explanations for false belief based actions, (where the actions frustrates the desire/cause) is that their commonsense psychological competence does not stretch to understanding false belief/conditions; desire is the only means available by which an explanation can be offered (see also Section 5.4.2).

One problem faced by this competence shift (CS) hypothesis is that often when desire/cause explanations were preferred to false belief/condition explanations in Chapter 5, the children in fact went on to answer the belief prompt appropriately. The CS position then, requires more sophistication. One line available is to argue that this correct answer to the belief prompt was not based on false belief understanding (e.g. Perner, 1989, 1991). Indeed, this line of argument motivated Wellman's change of position between 1989 and 1990.

Perner (1989) argued that children may interpret "think" questions as being in fact "thinking of" questions. The difference between "thinking that" and "thinking of" interpretations of think questions has been described in earlier chapters (Chapters 1 & 4). Perner went on to elaborate this in his prelief hypothesis (in press; see Chapter 1), where he discussed children's failure to recognise the defining difference between pretence and belief. Children's answers to belief prompt questions, such as those presented in the Chapter 5 Experiments, could be prelief-based answers, rather than false belief-based answers; the answers may reflect the children's belief that the character is thinking about (pretending that) the object at a specific location. This does not amount to belief competence, because prelief does not involve understanding mental states as representations (see Chapter 1, Section

158
1.7.1 for more detailed discussion of these ideas).

On one reading of the prelief hypothesis then, it is possible to argue that desire/cause explanations are a result of children being at the prelief "stage". Once past the prelief stage, as indexed by passing a belief prediction task, the children would not show this type of explanation pattern, but instead would be expected to offer belief/condition explanations. The prediction on the basis of this reading is that belief/condition explanation and accurate belief-based prediction should go together. To put it another way: the shift from prelief theorising (failing false belief: age three) and representational theorising (passing false belief: age four) as assessed by prediction tasks, should be mirrored in explanation tasks, with desire/cause explanations before the shift becoming belief/condition explanations afterwards. Experiment 6.1 assesses this version of the prelief hypothesis.32

6.1.3 Routine versus forced calculation of the belief/condition

The second possibility introduced in Section 6.2 was that children offered desire/cause explanations in situations where the desire was frustrated by the action because they had not calculated the belief/condition. Recall that in the explanation model presented in Chapter 5 (section 5.15), it was assumed that calculation of a non-default belief/condition blocked the desire/cause. This explained the preponderance of belief/condition explanations in the 1 Location Inconsistent condition in Experiment 5.1. If some children failed to calculate the belief/condition under those circumstances, then the desire/cause explanation would not have been blocked. However, the same children were able to calculate the belief/condition when they were later prompted to do so. This last fact indicates that the desire/cause explanations were not due to an inability to accurately calculate belief content, but suggests rather that the calculation of belief contents may not be a matter of routine. Recall that a distinction was drawn in Chapter 3 between Fodor’s CC theory, where children were only assumed to calculate belief/conditions under some circumstances, and the ToMM-

32 There is another reading of prelief, however, in which even spontaneous references to false belief in explanation circumstances may be indicative of prelief rather than belief understanding. There are two problems with this reading. One is that it is not now clear why explanations should ever refer to desire when the agent is clearly not acting in a way that would satisfy desire (acting "as if"). The other is that under this reading there is no way that a child's explanation could ever tell one that belief rather than prelief understanding was at work. The question becomes one of whether it is plausible that understanding false belief can only ever be indexed by false belief prediction tasks.
SP theory where children were assumed to calculate belief routinely, but also to make errors routinely (see discussion in Chapter 3, Section 3.10.5). This issue is readdressed in the discussion (Section 6.6.5), in the light of the results of Experiment 6.1.

One thing missing from this story so far is an account of how it is that children fail to calculate the belief/condition, given that the desire that causes the search is violated by the location of search. In the model I introduced in Chapter 5, the belief/condition calculation was hypothesized to be provoked by the discrepancy between the desire/cause (utterance information) and the action being directed elsewhere. If children fail to recognize this discrepancy spontaneously then they will consequently not be provoked into calculating the belief/condition. In the next section I make some suggestions as to what might determine whether a given child might recognize the action/desire discrepancy.

6.1.4 Performance resources in explanation

For the child to notice a discrepancy between the desire and the action, (s)he must have adequately represented i) the desire, ii) the location to which the action was directed, and iii) the contents of the Location toward which the action was directed. One possible reason for a discrepancy not being noticed would thus be a failure to keep track of one or more of these characteristics of the situation. Of the three, the desire component is implausible, since the children explain the action with reference to the desire. The location to which the action is directed is also unlikely to be a strain, since at the time of explanation the agent is standing where they searched; this information can be read directly from the situation. This leaves a failure to recall the contents of the location. Unfortunately, the children in Experiments 5.1 and 5.2 were not asked to recall the contents of the locations after their explanations, so this cannot be directly assessed.

I intend to hypothesize that keeping track of the above characteristics is something that places strain on the performance resources that children bring to bear in belief reasoning tasks. This position is in line with the general competence continuity approach that has been favoured so far, and builds specifically on the idea, introduced in Chapter 1 (Section 1.9.2), that performance limitations gradually relax during the fourth year and beyond.

The foregoing discussion gives rise to two predictions. Firstly, whether the discrepancy between action and desire is noticed by children will be affected by the saliency
of the characteristics of the action (e.g. the location toward which the action is directed; the memory for the contents of that location). Introducing manipulations that make the mismatch especially salient should increase the likelihood that the belief/condition is calculated, and reduce the number of desire/cause explanations offered as a consequence. Secondly, the performance limitations that are implicated in explanation are possibly beyond those that are involved in the calculation of belief/conditions, since some of the children offering desire/cause explanations in Chapter 5 had been able to calculate the belief/condition when required to do so. This suggests that children capable of passing a standard false belief task might, nevertheless, fail to calculate belief/conditions in explanation, and instead make reference to the desire/cause.

6.2 Interim summary: predictions

Both of the hypotheses introduced in Section 6.2 to account for why children offer desire/cause explanations for actions that frustrate the desire, suggested that some assessment of the relationship between a child's ability to predict an action and the explanation offered would be important. In the case of the CS compatible position, the expectation was that the relationship between passing a false belief prediction task and offering belief/condition explanations would be good. In addition, because the child's competence explains both passing the prediction task and offering belief/condition explanations, the relationship between these two measures should be a global one. That is, it should be unaffected by local structural changes in the explanation situation. It would be unlikely, according to this view, that children beyond "belief" understanding (i.e. those who pass false belief action prediction tasks) would offer desire/cause explanations.

On the other hand, according to the CC performance position, it is not considered unlikely that children might pass the false belief prediction task, yet offer desire/cause explanations, on the contrary, this is exactly the pattern of performance that is predicted. A further consideration is that according to the performance view, the explanation offered will be affected by local rather than global factors. For example, because what determines the likelihood of belief/condition explanation for a given child has more to do with whether (s)he has calculated it in a given set of circumstance than it has to do with the presumed "competence" stage (s)he has reached, it should be possible to affect the type of explanation
offered by manipulating the *local factors that affect the likelihood that the belief/condition will be calculated.*

In Experiment 6.1, I intend to make use of the fact that correctly predicting an action based on a false belief requires the belief/condition to have been calculated. Children who make a correct action prediction before explaining the action should offer belief/condition explanations, because they will have calculated the belief/condition. Children who are not required to make an action prediction before explanation may well fail to calculate belief and therefore explain with reference to a desire/cause. By assessing children under each of the two conditions (Prediction and No Prediction), it will be possible to assess both the *global* relationship between children's prediction and explanation, and also the local effect that making a prediction has on the explanation of the action predicted. The CS compatible position predicts a strong relationship between prediction success and belief/condition explanation, irrespective of whether the action explained was predicted or not. The CC compatible position, by contrast, suggests that the relationship between explanation and prediction will be local. When the action to be explained has been predicted first, the belief/condition will necessarily have been calculated and will be offered as explanation. Where no prediction is required before explanation, there is more chance that the belief/condition will not have been calculated and the explanation is therefore more likely to refer to the desire/cause. Before presenting Experiment 6.1, I outline evidence that is directly relevant to this issue.

6.3 Evidence: justification and prediction

There is some recent evidence on the relationship between prediction and explanation. Wimmer & Weichbold (1994) report an experiment that assesses Fodor's heuristic theory of the child's theory of mind. This experiment has already been considered in some detail in Chapter 3. Here however, I focus on an additional "explanation" procedure employed by the authors, which was omitted from earlier discussion.

In Wimmer & Weichbold's procedure, children were asked an "explanation" question after they had been required to predict the agent's action. The children watched a false belief situation (similar to that used in Wimmer & Perner 1983, described in Chapter 1) and were asked the test question concerning where the agent would look for the bait. Immediately
afterwards, if their prediction was correct, they were asked: "...and why will <agent> look there?". If their prediction was incorrect, they were first corrected (e.g. "No, look what Maxi does. He doesn't look here (location A) for his chocolate. He goes to this cupboard (location B) to get his chocolate." [1994, p47]), and then asked the explanation question above. The explanation procedure used in Wimmer & Weichbold is thus not strictly comparable to the methodology we have been dealing with, since in the cases where the child makes a correct prediction, the question is arguably not an explanation of an action that has occurred, but more a justification of the prediction. This is the reason that the procedure was not referred to in Chapter 4. However, the results are informative with regard to the issues that have arisen in the present chapter.

The results that Wimmer & Weichbold describe for the justification measure are not very detailed, and cannot support the type of analysis I presented for my own experiment in the last chapter. The authors do report a number of important facts about children's responses however. Firstly, correct answers to the explanation question were given 4% of the time by 3-year-olds, and 56% of the time by 4-year-olds. Secondly, all the correct answers explicitly or implicitly referred to the information condition which specified Maxi's belief; not a single child used the epistemic verbs "think" or "believe". Thirdly, there was a close to perfect association between accurate action prediction and correct explanation. I deal with each finding in turn.

Wimmer & Weichbold do not say how responses were scored as correct or incorrect for the justification measure. Specifically, they do not mention desire responses at all. The results that 3-year-olds perform poorly (4%), and 4-year-olds perform imperfectly (56%), are not surprising. Without further information this result is difficult to interpret.

"Information condition" explanations, according to Wimmer & Weichbold, are explanations in which the child refers either implicitly or explicitly to the information to which the child was exposed; the information which forms the basis of their belief. The examples that Wimmer & Weichbold give of this response type are as follows: for implicit cases, "Because it was there before...": "Because first she (mother) put it there...", for the explicit cases, "Because he saw that mother put it in there...".

A great deal is made of this category of explanation by Wimmer & Weichbold. It is interesting to note that the examples they give for the category are very similar to those I
identified in Chapter 4 as being similar to belief explanations (Section 4.4.5; also Table 4.5). This similarity was *in terms of what the explanations told us about how the children might have processed the task*. Explanations of this type are clearly belief/condition explanations rather than desire/cause explanations, although they do not explicitly mention belief. Looking more closely at the examples above, it turns out that they can be fitted neatly into the scheme that has been developed over the course of Chapters 4 and 5. Explicit cases of "informational access" (e.g. references to "seeing that"), were included among belief-based explanations in Chapter 4. Implicit cases ("Because that's where the sweets were before..."), are no more plausibly implicit cases of information access than they are implicit cases of belief. As far as Wimmer & Weichbold's "explanation/justification" measure is concerned then, the important point is that the belief/condition explanations evoked were unlikely to explicitly mention the terms "think" or "know".

The final aspect of the results to be considered is the close to perfect association that Wimmer & Weichbold observed between prediction and explanation. This sheds some light on the problem identified with "correct" and "incorrect" measures of explanation identified above. My assumption is that incorrect explanations were a mixture of "No Explanation" and "Desire" categories, while the correct explanations were largely explanations that I have previously identified as belief/condition explanations, albeit instances where belief was not explicitly mentioned. Children who failed the action prediction task gave mainly incorrect explanations, while those who had passed the prediction went on to explain correctly.

Wimmer & Weichbold interpreted these data as indicating that young children's performance on belief reasoning tasks was a function of their *increasing understanding of the informational causation of belief*. Children are hypothesized to pass false belief tasks by slowly coming to understand how beliefs are caused. Generally, before four years of age they do not understand, and afterwards their understanding is related to the extent to which information causation is understood (Wimmer & Weichbold, in press; Wimmer, Hogrefe & Sodian, 1988). In Chapter 1, I argued that this can be regarded as a version of CC theory (see Chapter 1, Section 1.8.1).

In terms of the predictions outlined in Section 6.2, these results are only of partial help. The association between prediction and justification/explanation was high, which provides support for the CS compatible position described in Section 6.2 (see also Section
6.1.2). If children who offer desire explanations were poor at offering accurate action prediction, then their desire explanations may well have been a result of their limited competence. The problem however, is that because Wimmer & Weichbold assessed only the \emph{global} relationship between prediction and explanation, it is not possible to evaluate the CC theory (see Section 6.1.3). One reason that the children who passed the prediction task might have offered belief/condition explanations is precisely that they had just passed the prediction task. Without a condition where the children are asked only to explain, it is not possible to assess the \emph{local} effect that prediction might have on explanation, which is central to the CC position. This manipulation is examined in Experiment 6.1.

\textbf{6.4 Experiment 6.1: introduction and predictions}

I argued above that the CC compatible position conjectures that making an action prediction will make computing the belief/condition necessary for children. This however, is assumed to be a local effect of task structure, in that it holds for the specific action predicted. On the basis of this, situations that require children to make an action prediction before explaining the action should result in more belief/condition explanations than situations where no such computation is necessary.

Wimmer & Weichbold (1994) showed a close association between action prediction and belief/condition explanation. The CS compatible position described above predicts that, because this association is hypothesized to be a result of belief reasoning competence, the prediction should have no local effect on explanation. Instead, one should be able to forecast the type of explanation in any situation from the success or otherwise of the action prediction; the relationship, in other words, is a \emph{global} one.

To test these predictions, children will be asked to explain an incorrect search action under two different conditions. One condition will involve the child predicting the action to be performed before it is performed, and then being required to explain it. The other condition will involve only explaining the action performed. The explanations of the children who pass the prediction task will be of special interest here, with respect to showing that any differences between conditions cannot be explained in terms of competence problems. For this reason, children from a slightly older age group were used as the subjects, to maximise the number of children passing the prediction task, and the responses of children who
correctly predict the action will be analysed separately.

An additional phase of the experiment sought to establish whether children who might not initially have explained the action with reference to a belief/condition were nevertheless capable of using a belief/condition in the context of explanation. In Section 6.1.4, it was predicted that making the mismatch between action and desire especially salient would result in the belief/condition being calculated, and there would be consequently more belief/condition explanations. To test this prediction, a further explanation question was asked. The incorrectness of the search was made explicit, first by asking whether the goal item was really in the location searched and then by asking children why the character had not searched the correct location. This second measure was called the "Not" explanation question, and it was predicted that it would invoke belief/condition explanations. This is because there is no plausible desire that might have caused the omission of search at the baited location, and the discrepancy between the original desire and the omission of the action predicted by it should be especially salient.

6.4.1 Design

This experiment had a within-subjects design. Children were presented with stories in which an agent searched the wrong location for an item. The situations were of the standard false belief type, where an unexpected transfer causes the false belief. Each child was asked to explain the search behaviour under two different conditions. In the Prediction condition, the children were asked to predict where the agent would search before being asked to explain the search action. It was expected that in this condition, being required to predict the action would cause children to calculate the belief/condition, making them therefore more likely to offer belief/condition explanations. In the No Prediction condition, the children were not asked to make an action prediction before they were asked to explain the action. To equate task structure, they were instead asked to recall what the agent had been doing when they were away. In this condition it was expected that children would be less likely to refer to a belief/condition explanation because they need not have calculated the belief/condition. Instead children in this condition were expected to offer desire/cause explanations.

The dependent variables were the responses that children gave to two explanation
questions that were asked about the agent’s incorrect search. The first explanation question asked why the agent searched where (s)he did. The second explanation was supplementary, and asked why the character had *not searched in the right place*. It was predicted that asking about the omission of the action predicted on the basis of the desire would make desire/cause explanation extremely unlikely in both Prediction and No prediction conditions, and result in the belief/condition being calculated and offered instead. The order of the two conditions was counterbalanced, so that half the children received the Prediction condition first and the other half received the No Prediction condition first.

6.4.2 Subjects

The subjects were 39 preschool and reception class children, aged between 50 and 64 months. There were two control questions which children had to answer accurately. Failure on either of these resulted in the child being excluded from the study. A further criterion was set such that testing continued until 24 children had passed the prediction task (in the Prediction condition). The resulting group consisted of 12 boys and 12 girls (age range 49-64 months, mean=56.2 sd=4.5). Explanation data were also collected for a further seven children who, though they had passed the memory control questions, had failed the prediction task.

The subjects were all recruited from playgroups and nursery schools in the North London area. The schools served children from a variety of socio-economic backgrounds. All the children received both story conditions, and were assigned to one of the two presentation orders.

6.4.3 Stories

The tasks used were very closely modelled on the Sally-Anne false belief task adapted by Baron-Cohen *et al.* (1985). The tasks were acted out in front of the children using props.

Both tasks involved an agent (e.g. Pippa) hiding an item (her apples) in Location A (a basket) on the table. The agent then was moved to an opaque container (the agent's "house"), in order to do something else (read a book). In the agent's absence, another character appeared on the scene and moved the item to Location B (a box). Two control
questions were asked: "where did Pippa leave her apples..?" and "where are the apples now..?", and then a further question. In the Prediction condition, the next question was the same as in a standard false belief prediction task; the children were asked: "where will Pippa look for her apples when she comes back..?". In the no Prediction condition, the question was not a prediction, but instead: "what is Pippa doing while she is away..?". This question was included in order to preserve task structure relative to the Prediction condition.

Finally, the agent returned to the scene, and the experimenter enacted the agent's search at the original location (A). The final set of questions was now asked. Firstly the explanation question: "why is Pippa looking in there..?". Following this the location was established as the wrong location: "will she find her apples there..?". Lastly, the children were asked the "Not" explanation question: "why didn't Pippa look over here..?" (point to correct location: A).

6.4.4 Materials

Prediction condition: A small female Duplo doll, "Pippa", was the main protagonist, and a small Duplo orange bear enacted the transfer. A small basket and a box were locations A and B respectively, and some small green "apples" made from Plasticine served as bait.

No Prediction condition: A small male Duplo doll, "Sam", was the main protagonist, and a small Duplo polar bear enacted the transfer. A jar and a box were locations A and B respectively, and some small coloured "sweets" made from Plasticine served as bait.

Children's responses were audiotaped using a Sony Professional portable recorder, and a written record was also taken at the time of the experiment.

6.4.5 Procedure

Each child was tested individually in a quiet area of the classroom. The children sat opposite the experimenter at a low table. The experimenter told the child that they were going to play some short games, and then presented the child with the first task. After a short period (about one minute during which the child's responses were recorded, and the new story materials were prepared), the second task was presented to the child. Each of the tasks lasted about three to four minutes, and the whole session was completed in under ten minutes.
As indicated, the answers were written down briefly during the session, and in addition, the sessions were audiotaped and the tapes transcribed later. If a child failed a control question during the testing, then the story was repeated up to that point and the question asked again. If the child failed again then they were excluded from the study. Children were also excluded if they failed to answer the prediction question accurately, and testing continued until a total of 24 children had met these criteria.

6.4.6 Scoring of explanations

In this experiment the main aim was to distinguish the conditions under which desire/cause and belief/conditions might be offered by children. Scoring was largely based around the scheme presented for Experiments 5.1 and 5.2, which in turn were based on the discussions in Chapter 4. See those sections for more details. Explanations were scored into one of the following three categories:

1. **Desire/cause** explanations specified limited information about the action. The action was attributed to a simple cause: a pro-attitude toward a goal object. No reference was made to information about the status of the goal object (e.g. "Because he wants the sweets..."; "Cos he's trying to get the sweets..."; "To find his sweets...").

Desire/cause explanations were expected to refer mostly to desires (where limited information is expressed about the action, see Chapter 4), and were also expected to be more frequent in the No Prediction condition, in which no belief/condition need have been calculated.

2. **Belief/condition** explanations specified more information about the action. The explanation made explicit an object, but also information relating to the status of the object that would enable a pro-attitude toward the goal object to have caused search elsewhere. This category was intended to capture references to belief (e.g. "Cos he thought the sweets were there..."), as well as other explanations that specified a similar level of detail (see...
Belief/condition explanations were expected to refer mainly to the agent's beliefs. Once again, following the guidelines presented in Chapter 4, it was required that the explanation specify both an object and *something about that object*. This was in order to rule out responses that simply referred to an object as the subject of the belief (e.g. "because she's thinking of..."), and responses that referred simply to an agent's ignorance (e.g. "because she doesn't know anything...", "because she was wrong..."). See Chapter 4 (Section 4.4.4; and Table 4.5) for more details. Other belief/condition explanations were also expected. Specifically, references to "non-mental" or external causes that made explicit information about a goal object that would enable the desire/cause. Examples include references to (past) reality (e.g. "because that's where the sweets were before...")\(^{33}\) or the agent's previous action (e.g. "because that's where he left the sweets..."), as well as explicit references to lack of informational access (e.g. "because he didn't see sweets were moved.") Wimmer & Weichbold (1994: see also Section 6.3).

3. No Explanation: The final category captured responses where the child failed to respond, stated that they didn't know the answer, or offered an irrelevant explanation (see Chapter 4, section 4.4.6).

6.5 Results

A total of eight children failed one or other control question in one or both of the conditions. Children only ever had any difficulty with the past location control question, typically responding that the bait item had been where it in fact now was. These children completed the procedure but their answers were not included in any analyses.

\(^{33}\) This category is different to the "reality" category used in the last chapter. In the previous usage, the emphasis was on references to the current reality: to what was in the location searched. This was in order to assess the idea that references to reality were justifications (see Chapter 5, Section 5.5). Here, the emphasis is on references to past reality: to where the bait item was before. The current aim is to assess whether children can specify belief/conditions, one example of which would be a reference to past location. Accordingly, references to "current" reality in this experiment will not be scored along with references to past reality, because they are references to an empty location ("Because there's nothing there..."), and therefore are not explanations that enable the desire/cause. Instead, they will be scored as "No explanation".
The age range of the sample was chosen so that most children could be expected to pass the prediction question. This indeed proved to be the case, with only seven children failing the prediction question. These seven children’s explanations were recorded, and they are dealt with in Section 6.5.2. A question concerning what the agent was doing while he was away was asked to preserve task structure in the No Prediction condition. All the children answered this question accurately.

6.5.1 Correct Predictors

In order to assess the pattern of explanation among children who were able to calculate the belief/condition, the first analyses were performed on the responses of children who passed the prediction task. The data are presented in Figure 6.1.

![Cause versus Condition explanations](image)

**Figure 6.1** Number of children offering cause and condition explanations in Prediction and No prediction conditions in Experiment 6.1.
Figure 6.1 shows the children's explanations, categorized according to the scheme described in Section 6.4.6. It can be seen from inspection of the figure that the explanations fall into the predicted pattern.

Table 6.1 shows individual children's consistency with respect to the type of explanation they offered. Each child was categorized according to how they had explained the action in each condition. Two children were excluded, one who had offered no explanation for either condition, and another who had offered an explanation for only one condition. The remaining 22 children fell into one of four categories as shown.

Table 6.1 Children's consistency in desire/cause or belief/condition explanation across conditions in Experiment 6.1

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Desire/cause</th>
<th>Belief/condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Prediction</td>
<td>Desire/cause</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Belief/condition</td>
<td>0</td>
</tr>
</tbody>
</table>

The data in Table 6.1 confirm the initial conclusions drawn from inspection of Figure 6.1. Many children are consistent in their choice of explanation. However, of those children who do show inconsistent explanation across the conditions, the pattern is in all cases to use a desire/cause explanation in the No Prediction condition and a belief/condition explanation in the Prediction condition. No children show the opposite pattern (McNemar test of change, binomial, p=0.016).

One further issue is that of the order of presentation. Recall that the order in which the Prediction and No Prediction conditions appeared was counterbalanced. There was no difference between the presentation orders in the number of children who were consistent in responding. Of the six cases where children changed their responses, four occurred when
the No Prediction condition came first (Fisher's exact test, two tailed p=0.39).

To assess the frequency of "informational access" explanations (Wimmer & Weichbold, 1994; see also Section 6.3), the belief/condition explanations were broken down into those that were specifically belief related (think and know) on the one hand, and those that were implicit or explicit cases of reference to informational access (see Section 6.3) on the other. Figure 6.2 shows the breakdown of explanations within the belief/condition explanations in each condition in Experiment 6.1. It can be seen here that there were a similar proportion of each in each condition. In addition, among the cases of informational access explanation, most (nine out of eleven) were "implicit" examples (see Section 6.3), referring to the object's past location ("because that's where the apples were before..."), or the agent's previous action ("because that's where she left the apples..."). Explanations that referred explicitly to "informational access" ("because she didn't see the sweets had moved...") were rare (one child: the same child in both conditions).

![Breakdown of Condition Explanations](image)

**Figure 6.2** Number of "Belief" and "Informational access" condition explanations in Prediction and No Prediction conditions in Experiment 6.1.
6.5.2 Incorrect predictors

For inclusion in the analysis, the children in the study were required to have passed the action prediction task. This was to ensure that all the explanation data was drawn from a sample similar in terms of their "belief competence". During the course of collecting the 24 correct predictions, explanations were also collected from seven children who failed the prediction task, although they had passed the two control questions. The explanations from these children were also classified as desire/cause, belief/condition or nil explanations. Four of the children offered nil categorized explanations, one offered a belief/condition explanation and two offered desire/cause explanations.

6.5.3 Not Explanation question

![Diagram](image)

**Figure 6.3** Number of Children offering desire/cause and belief/condition explanations to the "Not" explanation question in Experiment 6.1.
The children were also asked to explain why the character had not searched the correct location. These explanations were also scored according to the desire/cause-belief/condition categorization described above, and the results appear in Figure 6.3.

Two things are apparent from this graph: firstly, that fewer children attempted an explanation in response to this question, and secondly, that there were virtually no desire/cause responses. For both the Prediction and No Prediction conditions, most children attempted at least one of the two questions (Explanation and "Not" explanation). In those cases where only one explanation question was attempted, children always attempted the explanation question, no children attempted only the "Not" explanation question (nine versus zero in No prediction, eight versus zero in Prediction, McNemar test of change, binomial p=0.001 and p=0.002 respectively). It is clear that the number of desire/cause explanations for the "Not" explanation is negligible. When children account for the omission of an action, belief/condition explanations are generated, as predicted.

6.6 Discussion

The results of the experiment clearly showed that children's explanations of search actions could be affected by local changes in the situation leading up to the explanation. Children were often consistent in their choice of explanation, but forcing children to calculate the belief/condition did seem to increase the likelihood that a belief/condition rather than a desire/cause explanation would be offered. This is striking given that children witness essentially the same action in both conditions.

The results indicate that the relationship between the explanation of action and the prediction of action is not straightforward. The majority of the children taking part in Experiment 6.1 demonstrated their ability to correctly predict the action that would result from a false belief. However, in both conditions, there were sufficient desire/cause explanations to undermine the notion that belief/condition explanation and passing the prediction task go hand in hand.

6.6.1 CS theory compatible accounts

There are indications in the data that do suggest some relationship between prediction competence and explanation sophistication (see Section 6.1.2). Among the
children who passed the prediction task, most offered some kind of explanation. There were few children who did not pass the prediction task and, among those, only one child offered a belief/condition explanation, with the majority offering no explanation at all. There is thus only one child who wouldn't fit the suggestion that passing the prediction task is at least necessary for belief/condition explanations.

On the other hand, one could also argue that the results strongly undermine the notion that there is a simple relationship between reaching some "competence watershed" (as indicated by passing the false belief prediction task), and offering information or belief/condition responses to an explanation question. Most of the children in the experiment were able to predict the agent's action correctly, yet many of them went on to offer a computationally simple explanation: in terms of the agent's desire. This pattern was not expected on the basis of the CS compatible position outlined in Section 6.1.2. More problematic still is the finding that the explanation offered depended on the local conditions under which it had been evoked. Whatever association there is between correct prediction and belief/condition explanation does not appear to be a global one, which is problematic for positions that consider action explanation and action prediction to be converging measures of competence. The CS position might be modified in the way suggested above such that being at a certain competence stage (indexed by the passing of false belief prediction tasks), though necessary for belief/condition explanation, is recognized as not sufficient to guarantee belief/condition explanation. This modification to the CS position however, has no account of why children ever explain an action by referring to a desire/cause that has been frustrated by that action. It also cannot explain why changes to task structure have an effect on the explanations offered. These phenomena are easily accommodated in terms of competence and performance.

6.6.2 CC theory compatible accounts

Although problematic for the CS compatible account of desire explanation, the results are consistent with the CC compatible account (Section 6.1.3). Recall that local structural changes to the explanation situation were predicted to have an effect on the children's subsequent explanations. In particular, forcing the children to calculate the belief/condition before explanation was expected to make it available to be used in
explanation, when under usual circumstances it might not have been. Children were indeed more likely to offer an explanation that referred to a belief/condition rather than a desire/cause in situations where they had been forced to calculate the belief/condition first.

However, this cannot be the full story, for two reasons. Firstly, there were many cases observed in Experiment 6.1 that clearly contradict the pattern expected on the basis of the CC account. Although there were no children who showed a pattern opposite to that predicted (desire/cause explanation in Prediction condition, belief/condition explanation in No Prediction condition), which in itself is striking, many children were nevertheless consistent in the type of explanation they offered for each of the two conditions. The fact that desire/cause explanations were offered at all in the Prediction condition suggests that though calculating the belief/condition might have been necessary to block the desire/cause explanation, it was not sufficient. In the explanation model presented in Chapter 5, it was assumed that calculating a belief/condition that did not include the default belief content would block the desire/cause explanation (see Chapter 5, Section 5.15.2, also footnote 31). Secondly, there were children in the No Prediction condition (who had not been required to calculate the belief/condition), who nevertheless offered a belief/condition explanation. This demonstrates that some children calculated the belief/condition irrespective of the condition to which they were assigned.

To account for these findings, some elaboration of the CC position is required. I propose to offer an explanation in terms of children adopting one of two strategies that depend on different processing resource requirements. These two strategies, together with one assumption, can account for the complete pattern of explanation observed in the present experiment.

The children who offered belief/condition explanations in both Prediction and No Prediction conditions consistently offered a more complex explanation. This strategy is hypothesized as having been dependent on the child having had processing resources large enough for the action/desire discrepancy to be recognized irrespective of the condition. Children who followed this strategy were thus unaffected by the Prediction/No-Prediction manipulation: they calculated the belief/condition anyway, as a result of the mismatch between the desire/cause and the action (see Chapter 5, Figure 5.9). I intend to refer to this as internally motivated belief/condition calculation.
The other strategy is hypothesized to be a result of lower resources. The children did not recognize the action/desire discrepancy spontaneously, but were however able to calculate the belief/condition when prompted to do so (for example, by being required to predict action). I intend to refer to this as externally motivated belief/condition calculation. The key assumption is that only internally motivated belief/condition calculation results in the desire/cause being blocked in the way suggested in Chapter 5 (footnote 31). External motivation for belief/condition calculation makes that belief/condition available for use in an explanation, but because the belief/condition was generated externally to the explanation process, the desire/cause is not blocked, and therefore remains available as well.

Figure 6.4 Possible explanations in the No Prediction condition. Low resource children (left) have only desire/cause explanation available. High resource children have both desire/cause and belief/condition explanations available, but the desire/cause explanation has been blocked by the action/desire discrepancy. They thus use the belief/condition based explanation.
Figure 6.4 illustrates how the hypothesis explains the No Prediction condition. The children with the lower resources (left) have not calculated the belief/condition, and as a consequence, the only explanation available to them is the desire/cause. By contrast, the children with higher resources have calculated the belief/condition, and this calculation was *internally* motivated (motivated by the explanation situation). For these children, the desire/cause explanation is blocked, and so they offer the belief/condition as their explanation.

**Figure 6.4**

**PREDICTION CONDITION**

Low resource children have both desire/cause and belief/condition explanations available, and might choose either because the desire/cause explanation has not been blocked. High resource children, who have the desire/cause explanation blocked, offer belief/condition explanations.

Figure 6.5 illustrates how the hypothesis explains the Prediction condition. After having made a prediction, the lower resource child now has the belief/condition available.
However, because the belief/condition calculation was motivated externally (i.e. outside the explanation situation - in order to answer the prediction question), the desire/cause explanation is not blocked. Accordingly, the child has both desire/cause and belief/condition explanations available. The child might choose either, and this would result in some children ending up as having offered desire/cause explanations in both conditions, and the rest having offered a desire/cause explanation for the No Prediction condition and a belief/condition explanation in the Prediction condition. There were indeed about the same number of children showing each of these patterns of explanation. Note also, that one pattern of explanation that would not be expected according to this hypothesis, would be for children to ever explain the No Prediction with reference to a belief/condition, and the prediction condition with reference to a desire/cause. There were no such cases in Experiment 6.1.

I have argued here that in explanation, there are resource requirements beyond those implicated in prediction. In support of this, it would be desirable to show that the children following each strategy (high or low resource) were distinct in terms of some independent measure of resources. Unfortunately, no test is yet available that specifically measures resources, although the Screen task (Roth & Leslie, 1994; see also Chapter 1, Section 1.9.2), is the type of measure that might be ideal, if it were modified so as to be more difficult for 4-year-olds. In place of a specific independent measure of resources, a crude substitute will have to do. A rough guide to the resources available to the child has been assumed under the CC position to be the child's age (see e.g. Fodor 1992; Leslie & German, 1994; also Chapter 1).

6.6.3 Age considerations in experiment 6.1

To assess the suggestions made in the last section, the children in experiment 6.1 were assigned to one of the two resource dependent strategies described. Children who offered belief/condition explanations under both Prediction and No Prediction conditions were assigned to the "high" resource category, and children who either offered desire/cause explanations under both conditions or desire/cause explanations under No Prediction and belief/condition explanations under Prediction conditions were assigned to the "low" resource category.

If the suggestions made above concerning the relationship between resources (as
crudely indicated by age) and explanation strategy are correct, then the high resource strategy should have been generally adopted by the older children, and the low resource strategy should have been used by the younger children. The changing pattern of strategy, from low resource to high resource, that occurred with increasing age was assessed with Kendall's test for ordered contingency tables and found to be marginally short of significance ($S_{corr} = 47.3$, $z=1.63$, one tailed $p=0.0516$). While this result is encouraging, indicating a virtually significant change in strategy with age, clearly a more specific independent measure of resource would allow a better test of the resource hypothesis.

6.6.4 Conclusions

The experiment presented in this chapter has provided some evidence concerning the factors that determine how sophisticated the explanation for an incorrect search action will be. Particular attention was given to the notion of offering cause explanations, in situations where that desire/cause of an action is in fact frustrated by the action. Explanations of this type were suggested to be a result of either some general belief reasoning competence constraint, or alternatively a result of constraints imposed by limited resources. Experiment 6.1 provided results that have been interpreted as consistent with the latter position. The results have also motivated an elaboration of the simple explanation model introduced in Chapter 5.

One further issue arising from the results of this chapter, as mentioned in Section 6.1.3, has been the question of whether children can be assumed to routinely calculate belief contents in explanation situations. This issue was first raised in Chapter 3, where we saw that Leslie & German (in press) argued that Fodor's (1992) assumption that beliefs are only calculated when desire based prediction fails, was unwarranted. We argued that the young child routinely calculates belief in theory of mind problems, but fails to get the answer right under many circumstances. We supported this by drawing attention to children's problems with tasks that ask them to directly calculate the belief content, which indicate that children's difficulty must go beyond simply "not having tried".

The results presented in this chapter indicate that, though routine failure to calculate belief cannot be the whole problem, neither can a failure to accurately calculate belief contents, not in explanation of action at any rate. Children's reluctance to offer be-
belief/condition explanations in Experiment 6.1 cannot be due to a problem in calculating the
correct belief contents, since the children have shown themselves to be capable of predicting
action on the basis of the belief content. Non-routine belief calculation then, may still have
a role to play in explaining performance on certain theory of mind tasks, despite the fact that
other tasks show that problems in the calculation of belief contents are well established.
Elements of both versions of CC may be implicated in the complete picture.
CHAPTER 7

From Actions to Attitudes:
Children's use of their own actions to compute belief

7.1 Introduction

In this chapter the primary concern is the issue of whether theory of mind computations are influenced by whether the subject of the calculation is oneself or another agent. Self-other differences have already generated considerable interest in the theory of mind literature (Perner, Leekham & Wimmer, 1987; Astington & Gopnik, 1988; Gopnik & Astington, 1988; Wimmer & Hartl, 1991), but has been studied exclusively within utterance-based deceptive contents paradigms. The possibility of self-other differences has not been assessed in an action-based paradigm before. I intend to argue that an action explanation type situation is ideally suited to address the issue of self-other differences in belief reasoning, and that the more usual deceptive box methodology has several disadvantages.

Self-other differences have been argued to be relevant to a number of current issues within theory-of mind. Firstly, the theory-theory versus simulation-theory debate (e.g. Wimmer & Hartl, 1991) and secondly, children's use of various cues to calculate the content of mental states. For example, action evidence for belief attribution (Moses & Flavell, 1990), utterance cues to belief (Roth & Leslie, 1991; Riggs & Robinson, in pressa) and representational markers of belief (Mitchell & Lacohée, 1991).

In the following sections I turn first to the history of self-other differences in the theory of mind literature, and touch on how the results bear on the two sets of issues mentioned above. Specifically, I will argue that Self-Other differences cannot in fact tell one very much about radical simulation theory (see Chapter 2), but that they can be useful when viewed in terms of the cues that children might use to calculate belief contents. I then offer some reasons why the deceptive box task may not be the best paradigm in which to look for self other differences, and make the case for studying self other differences within an action based paradigm.
7.1.1 Self versus Other: brief history

The starting point is the deceptive box task introduced by Hogrefe, Wimmer & Perner (1986, Experiments 2 and 3). To recap, the child was shown a typical container (e.g. a "Smarties" box)\(^\text{34}\) and asked about its content. After the appropriate answer (Smarties) had been offered, the actual content (a pencil) was revealed. The box was closed again and the child was asked about the actual content, and his or her belief. Following this, the child was asked what another child would think when shown the closed box and asked about its content.

As far as self-other comparisons are concerned, the picture that has emerged is far from clear. Two early examples are firstly Perner, Leekham & Wimmer (1987, Experiment 2), who found that among 3-year-olds, 72% responded correctly when asked about their own belief (the "self" question) as opposed to 45% correct when asked the question about another child (the "other" question). Secondly, and with a slightly modified version, in which the child's initial expectation was not solicited, Gopnik & Astington (1988) found that children performed more poorly for their own beliefs than those of another child.

Wimmer & Hanl (1991) provided one of the most recent self-other investigations. They made a specific attempt to resolve the contradiction mentioned above, and clearly placed the issue in the context of the philosophical debate to which the simulation/theory debate described in Chapter 2 can be traced. I will dwell on this paper because it brings many of the important issues clearly into focus.

7.1.2 Perspective taking, informational causation or general representation?

Wimmer & Hartl began by describing the perspective taking tradition originating from Piaget (1932). This view conceived of social-cognitive development as a movement from egocentrism to perspective taking, and understanding false belief, on this view, has nothing to do with understanding mental states. Instead, it is seen as a perspective taking problem:

> If the other's position can be taken, then introspection of one's own mind, which is seen as posing no difficulty, guarantees that the other's mental state is understood (Wimmer & Hartl,

\[^{34}\] In Hogrefe et al., the items used were a matchbox that contained chocolate rather than matches. Since then however, the more usual container is a Smarties box.

184
This view shares the Cartesian assumption that concepts of mental states are acquired by introspection (Descartes 1639/1981). One consequence of this assumption is that understanding one's own mind is a precondition for understanding the minds of others (Hobbes, 1651/1968). Hence, the interest in Self-Other differences in the theory of mind literature.

Wimmer & Hartl went on to outline an alternative to the Cartesian view, that focused on the child's conception of the causal relationship between beliefs and the informational conditions that cause them. This proposal initially appeared in Wimmer, Hogrefe & Sodian (1988), aspects of which were discussed alongside Wimmer & Weichbold (1994) in the last chapter. The authors also contrasted this view with the general representational claim (Pemer, 1988, 1991; Forguson & Gopnik, 1988; see also Chapter 1, Section 1.7). Wimmer & Hartl took on board Gopnik & Astington's claim that looking at Self-Other differences could help adjudicate between these possibilities. Firstly, the expectation in the case of the Cartesian assumption is that reasoning about one's own mental states should be easier. Secondly, according to Gopnik & Astington (1988), understanding informational causation (e.g. Wimmer et al., 1988) is not necessary for understanding Own false belief, because one would not need to reconstruct belief contents from informational conditions for oneself: one could simply identify them. However, understanding one's own beliefs was hypothesized to require understanding of the representational relations between mind and world. Gopnik & Astington went on to argue that their result (Self belief harder than Other), suggested that the informational causation view was wrong.

Wimmer & Hartl noted, however, that Gopnik & Astington (1988) modified the Hogrefe et al. (1986) procedure in such a way that a crucial aspect of the above argument was changed. Gopnik & Astington dropped the elicitation of the child's initial expectation from their procedure, and since the children were presented with a number of misleading objects, Wimmer & Hartl suggested that they may have formed no expectation about the content of the box. Under these circumstances the task can be construed as requiring difficult hypothetical reasoning along the lines of: "if I did not know what's in the matchbox then I would think...". It is this possibility that may account for the discrepant findings for own false belief identification among 3-year-olds: 72% in Perner et al., (1987), 30% in
Wimmer & Hartl argued that to be regarded as a false belief identification task, where direct introspection might be possible, and information conditions can be ignored, the content of what has to be identified as a false belief should be *still in mind at the time it is inquired after*. If not, the content is more likely to need to be reconstructed from information conditions. They went on to test exactly this possibility by replicating the Self-Other experiments so far described, but ensuring that the critical condition for false belief identification (initial expectation still in mind) held. This was achieved by soliciting the child's initial expectation, keeping the interval between the expectation and the belief identification short and tricking each child only once.

Their results showed that there was no reliable difference between belief identification (Self) and belief inference (Other) conditions. These results are in agreement with Gopnik & Astington (1988), while being discrepant from Perner et al. (1987). Wimmer & Hartl went on to reject informational causation as being wholly responsible for 3-year-olds problems with false belief on the strength of these results. Instead, they suggested that Perner's (1988) proposal on the child's conceptual limitations with the notion of representation was correct. In addition, they argued that their results were evidence "against the Cartesian view on mind" (1991, p125).

I intend to argue, briefly, that neither position can in fact be rejected on the strength of the data described above. The reason offered above for rejecting both the positions rests on an over-extension of the Cartesian assumption, which Wimmer & Hartl actually noted in their paper. Descartes need not be committed to the view that all past mental states of the Self must be accessible, which is why Wimmer & Hartl took the trouble in their experiment to ensure that the content of the belief was still in mind at the time that the new content was revealed. However, surely the crucial condition required (to meet the Cartesian assumption), is that the belief content is still in mind at the *time it is inquired after*, not the time it is revealed to be false. The very fact that the original belief content has been replaced by the new belief makes it a past belief content, and therefore by definition not in mind at the time it is enquired after. It would appear on this argument, to be impossible to test any assumption about the accessibility of current belief contents using a false belief task.

I have already provided an analysis of the more recent version of the perspective
taking tradition, "radical simulation-theory", and argued that theory-theory provides a more persuasive account of commonsense psychology (see Chapter 2). I intend to leave the question of whether or not current mental states are directly introspected as an issue that could be compatible with theory-theory however it turned out. However, in Chapter 2 (Section 2.2), I mentioned a weaker version of simulation, which viewed "simulating" as one way that children and adults might sometimes calculate the content of belief. This version of simulation theory is addressed during the discussion of the experiments reported in this chapter.

7.1.3 The deceptive box task: some concerns

In this section I outline two concerns with the deceptive box task as a paradigm with which to look for self-other differences, and suggest instead the use of an action based paradigm.

The first concern is that the deceptive box task requires that the child identify a belief that has been both acquired and disconfirmed very recently; the child does not entertain the belief (that the box contains Smarties) for very long. Further to this, the child may in fact have only a limited commitment to this belief; (s)he makes a brief utterance, but that is all. The concern then, is that the child being asked to identify a fleeting belief which we cannot be sure was held strongly in the first place might make the later identification of the belief very difficult. More difficult say, than a belief that the child had held for some time, and which had more meaning. Consider a situation where the child performs some meaningful desire-based action, such as searching a particular location to retrieve an object. The child will have good reason to hold the belief in the first place, and should the belief turn out to be false it will have a salient consequence: the desire will be frustrated. I return to this idea shortly.

The second concern is related to the discussion above. It is based on the issue of what cues the child might use to calculate belief content. The deceptive box task can be considered "utterance-based" (see Chapter 1, footnote 2) because the child's commitment to the belief is in the form of an utterance. The utterance may well be all the child has to go on in order to calculate the belief. This may be one of the reasons that deceptive box tasks are so hard; utterances are transient and fragile things. Mitchell & Lachée (1991)
demonstrated that providing children with additional cues could help them overcome a bias toward reality responses in a deceptive box paradigm. Their posting task was described in Chapter 1, but briefly, the finding was that children who posted a physical "marker" of their belief content at the time of the utterance were better able to recall their belief than children who posted an irrelevant marker. It might be very difficult for a child to use an utterance alone (without the physical marker) to calculate belief, and I now turn to further evidence on exactly this issue.

The evidence is conflicting. On the one hand, Roth & Leslie (1991) showed that 3-year-old children used utterance cues to calculate the content of an agent's belief. In their task, two characters (Yosi and Rina) were playing outside and Yosi had some chocolate that she was going to eat. She placed it in one location (A), and left the scene. Rina transferred the chocolate into a second location (B). Yosi then returned and seeing that her chocolate was no longer where she left it, asked the Rina where it was. Rina replied (deceptively) that the chocolate was in a third location (C). The children were asked where each of the characters thought the chocolate was, and where the chocolate was really. Roth & Leslie found that 3-year-olds often judged that both the speaker and listener would believe that the chocolate was where the speaker had said it was (C). 5-year olds, by contrast, judged that the speaker believed that the chocolate was where it really was (B), and that the listener believed it was where the speaker had said it was (C). This latter pattern indicated that the full deceptive nature of the speaker's utterance had been comprehended. The 3-year-old children were thus more likely to attribute a false belief to an agent under some circumstances, which Roth & Leslie attributed to their having over-extended a strategy of linking utterances with beliefs. Autistic children were also tested by Roth & Leslie; they mostly predicted that both the speaker and listener would believe that the chocolate was in the real location (C), showing no evidence of even the utterance-belief understanding evidenced by the 3-year-olds.

More recently, however, Riggs & Robinson (in pressa) have failed to find evidence that children can use a speaker's utterance as a cue to belief. They used a task where a character was present when an item (a red cup) was initially hidden, but absent when the cup was exchanged (for a white one). When the character returned, he was asked what colour the cup was, and in reply he made an utterance: "it's red". Children were then asked what
colour the character thought the cup was, and what colour he had said it was. They were found to be very poor at using the character's utterance to calculate his belief, despite the fact they could recall accurately what had been said. Riggs & Robinson interpreted their results as indicating that children did not readily make the link between utterance and belief. They suggest that the "deception" structure of Roth & Leslie's task, with the speaker making untrue or "pretend" utterances, might have made the link easier for children to make.\(^5\)

Another possible reason for the discrepancy is that in the Roth & Leslie task, but not in the Riggs & Robinson task, the events are placed in the context of desire based actions. It is possible that the "conversation" between Yosi and Rina in the Roth & Leslie task was part of a situation that had more meaning than the Riggs & Robinson procedure. The child was able to interpret the utterance because it had been embedded in an action-based scenario. In Riggs & Robinson, by contrast, where no salient desires were implicated in the exchange, might have been a more difficult situation to interpret. This possibility leads us nicely away from the consideration of utterance cues, about which there is disagreement, to consider action as a cue to belief.

### 7.1.4 Do actions speak louder than words II?

The possibility that children might be able to use action based cues, presented on video, to calculate belief has been tested for the case of another person's belief (Moses & Flavell, 1990). This experiment was described in detail in Chapter 4 (Section 4.2.2). Recall that Moses & Flavell demonstrated that the more cues the child had to the belief, the better was their performance. However, they rejected this result on the basis that absolute performance was no better than chance. In Experiment 7.1, I adopt a video procedure in order to compare children's ability to use video evidence to identify their own false belief and the false belief of another person.

### 7.1.5 Experiment 7.1 introduction

In Experiment 7.1 children will be induced to play the part of Sally in a standard

---

\(^5\) Recall that in Chapter 1 (Section 1.7.2) I described Perner's prelief hypothesis and its application to "acting as-if" situations. Perner (in press) argues that children may be answering on the basis of the agent's utterance in Roth & Leslie's study because the children use utterance as a basis for calculating the agent's prelief. I return to this issue in Section 7.7.3.
"Sally-Anne" false belief task (Baron-Cohen et al., 1985). They will hide a bait object and leave the scene and in their absence, "Anne" (played by an experimenter), will move the object to another location. When they return they will search the wrong location. Subsequently they will be shown the video of both themselves and another child enacting these events, and asked to explain why they (the other child) looked where they did, and where they (the other child) thought the item was when they looked for it.

Children going through this procedure should entertain a false belief about the bait item's location for longer than the belief is usually entertained in the deceptive box paradigm. The search action is a concrete commitment to their belief, and should be especially salient when it turns out to be wrong.

Following the explanation experiments reported in the previous chapters, where desire based explanations were common, an additional step was included to make salient to children the fact that the search had taken place at the wrong location. This step was partially based on the "Not" explanation procedure used in Chapter 6, and involved asking children whether the bait item really was where they had looked for it, before asking the explanation question. This extra question was expected to make the fact that the search had taken place at the wrong location especially salient and as a consequence, block desire explanations (see Chapter 5, Section 5.15; Chapter 6, Section 6.6.2).

7.2 Experiment 1 Method

7.2.1 Design

The experiment had a within subjects design. Each subject was required to explain the search action of an agent under two conditions. In the Self condition the subject was shown a video film of themselves searching in an incorrect location for a bait item, and asked to explain the action. In the Other condition the subject was asked to explain the action of another child. Children were asked three questions. The first two were asked at the point in the film where the protagonist was searching the wrong location. Firstly: "is that where the item is really..?". Then, they were asked an open ended explanation question: "why are you looking there..?". Subsequently, the children were asked a direct belief attribution question: "where did you think the <item> was..?". Half the subjects received the Self condition first, the other half received the Other condition first.
7.2.2 Subjects

The subjects were 20 children (nine boys and eleven girls) aged between 39 and 44 months (mean = 41.7, sd = 1.5). A further five children were also tested but failed to meet the control question criterion described in the procedure (Section 7.2.4). The children were recruited through the subject pool at the MRC Cognitive Development Unit. The subjects in this experiment were predominantly white, from a range of socio-economic backgrounds. All the subjects had English as a first language. The subjects were assigned alternately to one of the two orders of presentation.

7.2.3 Materials

The false belief situations that the children were asked to explain were presented on a video recorder. The films were created as described in the procedure section below. The materials were a small low table, two playroom chairs, a small box (approx 9cm on a side, that opened at one end), a small basket (12cm diameter, 5cm deep, covered by a small cloth), and some biscuits (Jaffa cakes). The sequence was videoed using a video camera, and there was an opaque screen, on wheels, approx 180cm on a side. The BPVS (short version) was used as the intermediate game, and a video recorder and monitor were used to show the videos to the children.

7.2.4 Procedure

Pre-test procedure:

The subjects were tested in the babylab at the MRC Cognitive Development Unit. There were two experimenters (hereafter E1 and E2). Children arrived in the babylab and were met by E1. Each subject was told that they would be playing some games and would be able to have a biscuit as part of one of the games (permission for this had been obtained from the parent present).

The child was then taken into a room that was laid out as shown in Figure 7.1 above. The child was introduced to E2, who was sitting behind the small table and remained so throughout the procedure. The child was asked by E1 to sit opposite E2 and told that the whole game was going to be filmed, and that later they could watch themselves on film. E1 then asked the child to wave to the camera and started the video recorder.
E2 started the game by showing the child a biscuit, and saying that (s)he could have the biscuit to eat, but that first (s)he was to take part in a short word game with E1. The word game was to take place around the corner (E2 indicated the area behind the screen), and so E2 suggested that the child should hide the biscuit to keep it safe. The child was asked whether (s)he would like to hide the biscuit in the box or in the basket (these items were on the table in front of the child). The child indicated a location and was invited to hide the biscuit there. If the child was reluctant then E2 hid the biscuit in the location suggested by the child and the child was told to watch carefully.

![Figure 7.1 Set-up of experimental room for Experiment 7.1. Positions shown are starting positions. Changes to these are described in the procedure section (7.2.4).](image)

E1 then asked the child to remember where (s)he had hidden the biscuit and told him/her that they would now play a word game round the corner. The child was taken by
E1 to the area behind the screen, and shown the monitor and told that was where they would be able to watch the film later. The word game consisted in the first few training plates of the British Picture Vocabulary Scale. The child was then congratulated on how well (s)he had done and asked if (s)he wanted to get the biscuit.

While E1 and the child were behind the screen, E2 ostentatiously removed the biscuit from its hiding place and placed it in the other location, ensuring that it was in plain view for a few seconds.

The child was encouraged to return and retrieve the biscuit. E1 followed and was on hand to prompt the child to search on occasions where the child was reluctant to search spontaneously, and also to display surprise and mystification when the child uncovered an empty location. E1 quickly began "wondering out loud" what might have happened and encouraged the child to search somewhere else for the biscuit. If the child did not look in the other location on the table, E1 suggested it and the biscuit was eventually retrieved. E1 asked the child if (s)he wanted to find out what happened, and indicated that the whole thing had been filmed, and that this would therefore be possible. The child was taken back to the area behind the screen by E1, and the films were watched.

Test Procedure:

In the Self condition, the child was presented with the film of the events (s)he had just participated in, with E1 commenting on significant moments. The children were encouraged to describe the transfer event while it took place. E1 stopped the film at the point where the child was searching the empty location. The following questions were then asked: "is that where the biscuit really was?", followed by "why are you looking there?", and finally "where did you think the biscuit was?".

In the Other condition, the child was shown the video of the child that had been tested previously. The child was told that the little boy (girl) in the film had played the same game as they had. The questioning procedure employed was the same as that described for the Self condition, but with "(s)he" substituted for "you" in the questions.

When the order of presentation was Self-Other, the children were shown the video

\[36\]
Note that this prompting was only to induce the child to search. No direction as to which location to search was given, only encouragement that the child was free to retrieve the biscuit.
of themselves, and then simply asked whether they would like to see another film of another boy (girl) that had played the same game. Children were happy to watch another film. The other order (Other-Self) was slightly more problematic, as the children had been offered the chance to find out what had happened in their own game. The video was "accidentally" rewound too far, and the children were told that the first film was of a little boy (girl) who had played the game before. They were asked if they would watch this film first, and told that the film of themselves came on afterwards. Children who received this order of presentation were in fact quite happy to watch the other film first. A written record of the child's answers was taken, and the whole questioning procedure was also audiotaped.

7.2.5 Explanation question scoring

The responses to the explanation questions were scored into the same categories as were used for Experiment 6.1. The scheme is re-presented here:

1. **Desire/cause** explanations specified limited information about the action. The action was attributed to a simple cause: a pro-attitude (e.g. desire) toward a goal object (or proposition). No reference was made to information about the status of the goal object. Examples included: "because he wants the sweets.."; "cos he's trying to get the sweets.."; "to find his sweets..".  

Cause explanations were expected to refer mostly to desires.

2. **Belief/condition** explanations specified more information about the action. The explanation made explicit an object, but also information relating to the status of the object that would enable a desire for the goal object to have caused search elsewhere.

Condition explanations were expected to refer mainly to the agent's beliefs. Following the guidelines presented in Chapter 4, it was required that the explanation specify both an object and something about that object. This was in order to rule out responses that simply referred to an object as the subject of the belief (e.g. "because she's thinking of..."), and
responses that referred simply to an agent's ignorance (e.g. "because she doesn't know anything...", "because she was wrong...") see Chapter 4 (Section 4.4.4, also Table 4.5) for more details. Other condition explanations were also expected including references to (past) reality (e.g. "because that's where the sweets were before...") or the agent's previous action (e.g. "because that's where he left the sweets..."), as well as explicit references to lack of informational access (e.g. "because he didn't see sweets were moved...").

3. **No Explanation**: The final category captured responses where the child failed to respond, stated that they didn't know the answer, or offered an irrelevant explanation.

7.3 Results

![Initial explanations](image)

**Figure 7.2** Number of children offering Desire/cause, Belief/condition and Nil explanations in Self and Other conditions in Experiment 7.1.
Five children were excluded from the study for failing the control question. Figure 7.2 shows the initial explanations observed in each category for each condition. Note that the crucial explanations in this experiment are the belief/condition explanations. Steps were taken to minimize desire/cause explanations. It can be seen from this figure that the incidence of condition explanations was not especially high, and there was little evidence of a difference between the Self and Other conditions on this first measure. Virtually all the condition explanations were belief based. There were no instances of past location or information access explanation. Of the few desire explanations, about half were references to a desire to find out what was inside the location.

In Table 7.1 children are categorized according to their consistency in belief/condition explanation across the two conditions. They fall into one of four categories: offering belief/condition explanations for both Self and Other conditions, any other explanation for both conditions, or offering a belief/condition explanation for just one category (Self or Other).

**Table 7.1 Number of children offering belief/condition explanations for Self and Other conditions in Experiment 7.1**

<table>
<thead>
<tr>
<th></th>
<th>SELF</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief/condition Explanation</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Other Explanation</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

Five children offered belief/condition explanations for both self and other conditions and eleven children offered other explanations for both conditions. There was no indication from those children who offered a belief/condition explanation for only one of the two conditions that the belief/condition was any more likely in the Self condition than it was in the Other condition or vice versa (McNemar test of change, binomial, \( p > 0.3 \)).
Figure 7.3 below shows the number of children who were correct on the belief attribution question in each condition (Self and Other). To be scored as correct a child was required to point or indicate verbally the original location (the one that the agent was looking into in the video.

![Belief attribution](image)

**Figure 7.3** Number of children correct on the false belief attribution question in each condition in Experiment 7.1.

Two things are immediately clear from this graph. Firstly that performance in the two conditions is the same, and secondly that the absolute level of performance is quite good. The consistency data bear out this picture. They appear in Table 7.2, which shows the number of children passing and failing the belief attribution question in each condition in Experiment 7.1.
Table 7.2 Number of children correct on belief attribution question for Self and Other conditions in Experiment 7.1

<table>
<thead>
<tr>
<th></th>
<th>SELF</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td>OTHER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Incorrect</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Eleven children were correct in both conditions, and there was again no indication that children correct in only one of the two conditions were more likely to be correct for the Self condition than for the Other condition or vice versa (McNemar, binomial, \(p>0.6\)). Only three children were incorrect in both conditions.

In addition to this, the absolute performance here comes very close to being significantly better than chance in both conditions (binomial, \(p=0.058\)). Chance performance is a very conservative benchmark to use in false belief tasks, but is included here because standard task performance was not assessed in these children.

Since children received the two conditions in quite a short space of time, their answers to the initial explanation and belief attribution questions were also assessed for the possible effect of which condition was presented first. The children in each of the two presentation orders (Self-Other and Other-Self) were treated as separate groups for this analysis. If presentation order had no influence, then the proportion of belief condition explanations and accurate answers to the belief attribution question should be the same in each group.

For the initial explanations: of the eight children offering belief/condition explanations for the self condition, five were in the Self-Other group and three were in the Other-Self group (Fisher's exact test, two-tailed \(p>0.41\)). Among the six children offering a belief/condition explanation in the other condition, five were in the Self-Other order group and one in the Other-Self group. This difference falls just short of significance (Fisher's exact test, two-tailed \(p=0.076\)).
Figure 7.4 shows the number of children correct in each condition under each presentation order. It can be seen that performance in both conditions was better when the Self condition precedes the Other condition. This trend was in the same direction as observed for the initial explanation question presented above, and each differences falls only just short of significance (Fisher's exact test, two tailed p=0.076 for both conditions).

Though these differences were not quite significant, the group sizes for this comparison were relatively small (N=10), so one cannot dismiss out of hand the notion that the children's performance might have been affected by the order in which they received the tasks. It appears that children did better on both Self and Other questions when the Self condition came first. The possible implications of this are outlined more fully in the next sections.
7.4. Experiment 7.1 discussion

7.4.1 Explanations

The explanations offered in Experiment 7.1 did not fall into the same pattern as has been observed over the last two chapters. Firstly, the number of desire explanations has dramatically fallen. Most of the explanations offered by children in this situation fell into either the belief condition category or the nil category, and there were not an especially large number of belief condition explanations (40% for Self, 30% for Other). The children often were unable or unwilling to offer an initial explanation at all. Clearly this situation was one in which the initial desire, frustrated by the action performed, was not considered an appropriate explanation. This contrasts with the experiments presented in the last two chapters, where children sometimes offered desire explanations even in situations where the desire is frustrated by the action.

This finding is perhaps not as surprising as it may first appear. Recall that one of the aims of the procedure used in this experiment was to make more salient the incorrectness of the action (see Section 7.1.5). In this experiment, the child performed a desire-based action on the basis of the false belief. For this reason, the fact that the action turned out to be wrong might well have been especially salient, as suggested in the introduction. Moreover, the absence of a difference between the Self and Other conditions indicated that children appeared to find an action that frustrated another agent's desire equally salient, which was not expected. I will postpone further discussion of the Self-Other issue until I have dealt with the effect of presentation order (see Section 7.4.3).

7.4.2 Belief attribution

Performance on the belief attribution question was very good for both conditions. There was no indication of an advantage for the Self condition over the Other condition, considering performance alone (but see next section). The performance on this measure was almost better than the strict "chance" benchmark adopted in Moses & Flavell (1990), who reported that additional cues to help children reason backwards to a false belief were not effective (when assessed against chance). The present results suggest in contrast that at least in this scenario, where children had taken part in the situation themselves, extra cues were helpful to the child attempting to calculate belief contents.
7.4.3 Order effects

There was no indication of any Self-Other difference in the results of the first experiment. However, there were indications that the within subjects design, with each child receiving both conditions, might have resulted in an effect of the order of presentation. None of the differences were significant, but for the explanation measure performance in the Other condition tended to be better if the child had already received the Self condition. For the belief attribution measure, performance on both measures appeared to be better if the Self condition had been presented first. One interpretation of this effect is that an advantage of Self over Other was being masked by the within subjects design. Children were more likely to get the Other question right if it followed the Self question. Where other preceded Self, however, not only did performance drop in the Other condition, but this may have attenuated any advantage that might have been enjoyed by children if they had answered the Self question first. If children's recall of their own desire was important in establishing their good performance for the belief question, then receiving this question second may have lessened the effect that their desire being frustrated had.

Because of this possibility, a further experiment was run, with an extremely similar methodology. The aim was to supplement the data collected in Experiment 7.1. If each child's first task only is taken into account from Experiment 7.1, a between-subjects effect of Self versus Other begins to open up. In addition, I took the opportunity to make some other improvements which are outlined in the next section. The results of Experiment 7.2 will be presented first in their own right, and then used to supplement the data from Experiment 7.1.

7.4.4 Problems and improvements

As well as the possible order effect mentioned above, a number of procedural problems existed in Experiment 7.1. Most importantly, there was no control exerted over exactly what the children saw in the Other condition. Because the children were simply shown the film of the child who had been tested previously, the two conditions that a child saw could have differed in a number of respects (e.g., different sex child; different absolute timing of the events; different actual audio record; different location of search etc.).

For this reason, it was thought desirable in the follow-up experiment to control at
least some of these parameters. In Experiment 7.2 two standard films for the Other condition were created: one for each location of search. In this way the children in the Other condition could be assigned to see either a child searching in either the same place as or the opposite place to where they had searched. In addition, all the children in the Other condition would see films that had the same timing in the sequence of events. Finally, to control the auditory aspects of presentation, it was decided that the films would be presented with no sound, and that the experimenter would provide an identical commentary for each film.

There was one other significant change made. This was to include a prediction/recall question before the explanation, in order to test the possibility that one's own action might be easier to recall than another child's action is to predict. It was also hoped that the prediction question, following the results from Experiment 6.1, would increase the likelihood of belief/condition explanations overall. The effect of adding the prediction/recall question will be assessed in itself, between experiments, before the results are amalgamated.

7.5 Experiment 2 method
7.5.1 Design

This was a between subjects design, with two conditions. In the Self condition, subjects watched a video of themselves going through the same procedure as described in Experiment 7.1, culminating in their searching the wrong location for a bait item. They were then asked prediction and explanation questions pertaining to their own incorrect search. In the Other condition, children were taken through the same procedure as children in the Self condition, but were shown a video of another child's incorrect search. They were asked the same questions about the search event depicted on the video.

There were three dependent measures. The first was an action-prediction task; the film was stopped following the transfer of the bait, and the subject was asked to state where the agent in the film would look for the bait item on their return to the scene (Recall of own action for Self; Prediction of other action for Other). The next question was asked with the film stopped at the point where the agent was searching the empty location; the child was asked why the agent was looking there. Finally, immediately after their answer to the explanation question, the children were asked a direct belief attribution question: "when
<agent> looked for the <bait item>, where did <agent> think the <bait item> was?".

7.5.2 Subjects

The subjects were 28 preschool children aged between 3:3 and 3:9 (mean = 3:6). They were divided into two groups of 14. The Self group consisted five boys and nine girls aged between 39 and 45 months (mean = 42.1, sd = 1.8). The Other group consisted 5 boys and 9 girls aged between 39 and 45 (mean = 41.4, sd = 1.8). A further five children were excluded from the study for failing to complete the procedure, or failing control questions (see Section 7.5.3). The children were recruited through the subject pool at the MRC Cognitive Development Unit.

Subjects were assigned to either the Self or Other conditions on the basis of the order they were tested. Subjects were run in the Self condition first, in order to find suitable films to be used in the Other condition (see materials section). Once two films had been made, subjects were assigned to conditions in advance, keeping the age and sex distributions the same.

7.5.3 Materials

The false belief situations were presented to the children on video. The videos were created as described in the procedure section below. In the initial procedure, the same set of materials was used as in Experiment 7.1. For the Other condition, two videos were chosen from the initial subjects run in the Self condition. This was to ensure i) that the number of children watching a child hide the biscuit in the same location as they had was the same as the number watching a child hide the biscuit in a different location and ii) that children in the Other condition watched videos that were as similar as possible to one another in terms of the timing of the sequence of events (see discussion of Experiment 7.1, Section 7.4.4).

7.5.4 Procedure

Pre-test procedure:

The pre-test procedure adopted here was exactly the same as in Experiment 7.1.
Test procedure:

In the Self condition, the child was presented with the film of the events that (s)he had just participated in. The video was presented with the sound turned down and E1 commenting on significant events. The children were encouraged to describe the transfer event while it took place. The film was stopped after the transfer event and the following questions were asked. Firstly two control questions: "where did you hide the biscuit right at the beginning..?" and then, "where did E2 move the biscuit to..?". If the child answered either of the control questions incorrectly, the film was rewound and shown again. If the error persisted, E1 continued with the procedure but the child's data was not included. Finally the test question was asked: "when you came back into the film, where did you look for the biscuit..?".

The film was re-started and the next events were shown. The film was stopped again at the point at which the child was searching the empty location. The following questions were then asked: "is that where the biscuit really was..?" followed by the explanation question, "why are you looking there...?". Immediately after the child had answered the explanation question, the final belief attribution was asked: "when you were looking for the biscuit, where did you think the biscuit was...?".

This was a between subjects design, and so there was no need in the Self condition to show another film to the child. In the Other condition, half the children were shown a video of another child who had hidden the biscuit in the same place that they had. The other half were shown a video of a child who had hidden the biscuit in the different location. The test procedure was very similar to that described for the Self condition. Children in this condition however, went through an additional questioning phase during the film. When they were watching the other child hide the biscuit, they were asked firstly to recall where they had hidden the biscuit, and then to describe where the child in the video was hiding the biscuit. At this point E1 commented as to whether these locations were the same or not, and sought the child's agreement.

If a child made an error at this stage the video was replayed from the beginning, and the procedure continued once the child had successfully recalled the location that (s)he had hidden the biscuit. Only one child needed to be reminded of the location of his own search. No children failed to accurately describe where the child in the video had hidden the biscuit,
or to agree on EI's judgement about whether the locations were the same or not. Many children spontaneously commented on the events in the video as being similar to or different from what had happened to them. The only other difference between the procedures was substituting "(s)he" for "you" in the questions.

7.6 Results

For the Recall/Prediction question, children were scored as correct if they predicted search at the original location. The graph in Figure 7.5 shows the number of children correct in each condition.

![Graph showing the number of children correct on Recall/Prediction task in Self and Other conditions in Experiment 7.2.]

**Figure 7.5** Number of children correct on Recall/Prediction task in Self and Other conditions in Experiment 7.2.
It can be seen from this graph that performance appears better in the Self condition. However, this difference is not significant ($X^2_{1,28} = 1.28$, one-tailed $p>0.1$). Children then, are not significantly better at remembering their past actions than at predicting false belief based action in another child.

Other(same)/Other(diff) subgroups: In the Other condition, half the children saw the other child hide the biscuit in the same place they did, and half saw the other child hide the biscuit in the different location. There was no significant difference between the performance in each of these groups (one out of seven for same versus four out of seven for diff, Fisher's exact test, two-tailed $p=0.143$).

The explanations were scored into the same categories that were used in Experiment 7.1. The pattern of explanation in each condition appears in the Figure 7.6.

![Explanations](image)

**Figure 7.6** Number of children offering Desire/cause, Belief/condition and Nil explanations in each condition in Experiment 7.2.
Again, there appears to be a difference between the conditions in the explanations that are offered. Nearly half the children in the Self condition offered belief/condition explanations (and these were in fact all belief based conditions as in Experiment 7.1). In the Other condition, there were only two belief/condition explanations (one belief based, one that referred to the past-location\(^7\) of the object). This difference however, once again, falls short of significance (Fisher's exact test, one tailed p<0.1).

The final measure was the belief attribution question. The children's performance on this question in each condition is shown in Figure 7.7.

![Belief attribution](image)

**Figure 7.7** Number children correct on the belief attribution question in Self and Other conditions in Experiment 7.2.

\(^7\) See description in Chapter 6.
Here there is a marked difference between the two conditions; a difference which is significant ($\chi^2_{2\text{df},N=28}=5.39$, one tailed $p<0.02$). The performance in this condition is also significantly better than chance (Binomial, $p<0.01$). Other(same)/Other(diff) subgroups: As was shown for the prediction question, the performance in the Other(same) and Other(diff) subgroups was not significantly different (two versus three out of seven, Fisher’s exact, two tailed $p=0.633$).

7.7 Results across both experiments

![Explanations](image)

**Figure 7.7** Number of children offering Desire/cause, Belief/condition and Nil explanations in each condition in Experiments 7.1 and 7.2 combined.

Firstly the effect of the prediction question on the initial explanation was assessed. The proportion of belief/condition explanations among the ten children who received the
Self condition first in Experiment 7.1 was compared to that among the 14 children in the Self condition of Experiment 7.2. The proportions (five out of ten versus six out of 14) were not significantly different ($X^2_{corr}[N=24, df=1] = 0.005$, N.S.). The same calculation was performed for the proportions in the Other condition (one out of nine versus two out of 12), with the same result (Fisher's exact test, two-tailed $p=0.629$).

The results of each child's first trial from Experiment 7.1 were therefore added to the relevant group in Experiment 7.2 (ten to the Self condition and ten to the Other). The results appear in Figure 7.7, which shows the number of children offering desire/cause, belief/condition and nil explanations in Self and Other conditions in Experiments 7.1 and 7.2 combined. When the data from the two experiments are combined, the pattern of explanation clearly differs according to condition ($X^2_{corr}[N=48, df=1] = 4.94$, two-tailed $p<0.05$).

![Belief attribution](image)

**Figure 7.8** Number children correct on the belief attribution question in Self and Other conditions in Experiments 7.1 and 7.2 combined.
The effect of the prediction question on the belief attribution measure was also checked. The proportions of correct responding among the two Experiment 7.1 groups were compared to their complement groups in Experiment 7.2, as described for the explanation question. Prediction had no significant effect in either Self or Other groups (12 out of 14 versus nine out of ten for Self, Fisher's exact test, two-tailed p=0.629; five out of 14 versus five out of ten for Other, $X^2_{\text{cont}}[N=24, \text{df}=1] = 0.07, \text{N.S.}$).

The data from the two experiments was combined for this measure; it appears in Figure 7.8. Analysis once again revealed a strong effect of condition ($X^2_{\text{cont}}[N=48, \text{df}=1] = 9.11, \text{two-tailed } p<0.01$), and the performance in the Self condition was also significantly better than chance (binomial, $p<0.001$).

### 7.7 Experiment 7.2 discussion

Overall, Experiment 7.2 is an improvement on Experiment 7.1. More measures were taken to control the exact parameters of each condition. Despite these changes, and the inclusion of a prediction question, the pattern of explanation in Experiment 7.2 was similar to the pattern observed assessing each child's first task only in Experiment 7.1.

The difference between performance on the Recall question in the Self condition and Prediction question in the Other condition was not significant. This result contrasts with that reported in Riggs & Robinson (in press $b$), where children were accurate at recalling both their own and another person's false belief-based actions. I will return in more detail to the results of Riggs & Robinson (in press $b$) in the general discussion (Section 7.7.1).

As far as explanation was concerned, there was firstly no effect of prediction on explanation (assessed between Experiments 7.1 and 7.2), unlike the experiment presented in Chapter 6. Recall that measures were taken in this experiment to reduce the number of desire-cause explanations: specifically the inclusion of the "where is it really..?" question before the explanation question. This prompt had been used before the "Not" explanation question in Experiment 6.1, and the results of this experiment indicated that it was effective in reducing desire/cause explanation here as well.

There was a marginal difference between Self and Other on the explanation measure in Experiment 7.2, an effect that was clearer when the results of the two experiments were combined. The proportions of children offering belief/condition based explanation were
46% and 12.5% in the Self and Other conditions respectively. In terms of the rate of belief/condition explanation that has been observed in Experiments 5.1 and 6.1, 46% is quite good performance for children much younger (mean age = 41 months) than those used in the previous chapters (compare 44% in 1 Location Inconsistent condition, Experiment 5.1 with children of mean age 49 months; 58% in the Prediction condition of Experiment 6.1, with children of mean age 56.2 months).

The belief attribution measure revealed a genuine Self-Other difference. Children were very good at identifying their own past false belief when given video evidence cues to help them do so. The same video cues did not appear to help the children in the Other condition.

7.8 General Discussion
7.8.1 Relation to other findings

These results differ from those reported in Wimmer & Hartl (1991). Although there was no difference between Self and Other on the "concrete" question ("what did you say...?" in Wimmer & Hartl, "where did you look...?" here) in either study, there was no difference for the belief question either in Wimmer & Hartl, whereas there were differences in both explanation and belief questions over Experiments 7.1 and 7.2 combined.

The results also differ from those of Riggs & Robinson (in pressb). I will describe this study in more detail, since it is similar in many respects to Experiments 7.1 and 7.2. These authors used a paradigm similar to the deceptive box paradigm, but asked children to retrieve an object from a deceptive container rather than to state what was in it (utterance task). They then compared children's memory for their own action and their ability to identify their own false belief with the corresponding measures for another person. They found that children were able to recall both where they themselves had searched for the item and also where another child had searched, but performed poorly when asked about Self and Other's belief. By contrast, the current experiment found better results on the belief attribution measure for the child's own belief. There were, however, two obvious differences between the two paradigms that might account for this difference.

One difference is that the Riggs & Robinson procedure, although it used an action as the child's commitment to their belief, did not use an action that was strongly desire-
The action in Riggs & Robinson was desire-based only to the extent that the child wanted to satisfy the experimenter's request ("can you get me an egg...?"). The action in Experiments 7.1 and 7.2 by contrast, was based on the very salient desire to retrieve the biscuit. Recall that the absence of strong motivation for the children when they made their commitment to the initial belief in the deceptive box paradigm was suggested in Section 7.1.3 to make later identification difficult. I will return to the possible role that desire might play shortly.

A second difference to the Riggs and Robinson procedure was the presence of additional (in this case video) cues in order to help the calculation. One possibility is that video cues in Experiments 7.1 and 7.2 played a role similar to that played by the "posting" procedure adopted by Mitchell & Lacohée (1991). The video cues in Experiments 7.1 and 7.2, however, only appeared to give an advantage to children calculating their own belief. Similarly, Mitchell & Lacohée (1991) demonstrated superior performance in the posting condition only for children reasoning about their own past beliefs; reasoning about the beliefs of another child was not assessed in their study. It is possible that Mitchell & Lacohée's procedure would improve performance on a question assessing the calculation of another child's belief contents. However, another possibility is that the advantage can only be gained for Self, as demonstrated here.

A further piece of evidence consistent with this latter picture is that Moses & Flavell (1990; see also Section 7.1.3) found that the overall performance on belief attribution for Other was not better than chance when video cues were provided to children. Taken together, these findings suggest that providing cues (video or posting) alone, is not sufficient to improve belief reasoning performance: other factors may be involved. How then, might the advantage that children enjoyed in using cues to calculate belief for Self, demonstrated across these studies (Moses & Flavell, Mitchell & Lacohée and Experiments 7.1 and 7.2), be explained?

### 7.8.2 The role of desire

I intend to offer an explanation that is based on the general approach that has been adopted in previous chapters. I will again make use of the ToMM-SP model (Leslie and colleagues, see Chapter 1, Section 1.9).
In the ToMM-SP model the improved performance on certain tasks is held to be a result of less stress being placed on the Selection Processor. One possible reason why children performed well for the Self conditions with posting or video cues (Mitchell & Lacohée; Experiments 7.1 and 7.2), may be that the memory of having entertained the belief itself can contribute to the selection process. On this view, the memory of the false belief is an additional cue that lightens the load on SP in the case of Self, but not for Other, where the memory cannot be as well evoked.

However, an alternative possibility, that seems especially plausible for the Self advantage enjoyed with video cues (Experiments 7.1 and 7.2) may well be that children were reminded not of their past belief but of their own desire for the bait item. The calculation of the belief condition may well have been made easier in the presence of both video evidence and the memory of the desire. In terms of the ToMM-SP model, recalling the desire may well have helped identify the action as false belief based, and also have helped in the selection of the appropriate belief-condition. For the case of the other child, the video would not have evoked the child's recall of own desire to the same extent, and therefore would not have been as much help.

These two possibilities are not mutually exclusive and I wish to highlight one similarity and one difference between them. In terms of similarity, both make the prediction that the advantage of Self over Other should be relatively short-lived, since the effect relies on the evocation of a recent memory which is either used as additional evidence, or provides an extra bias in selection during the belief calculation. This source of evidence would become weaker as time elapsed between having held the belief/desire and being asked the test question. A test of this hypothesis would be to show children the videos of their own actions after various time delays. The Prediction is that the memory of having entertained the false belief or the memory of having entertained the desire for the bait object would become a less salient cue as time passed.

The desire hypothesis, however, raises the possibility that video evidence might not even have been necessary for improved performance. The standard tests of Own false belief have been carried out using non desire-based paradigms, and no advantage for Self over Other has been consistently observed. This indicates that the memory of Own belief alone is not sufficient (see e.g. Gopnik & Astington, 1988). It is possible, however, that simply
asking children about their own belief when a desire-based action turns out to be incorrect may well be enough to provoke accurate calculation of belief. If the memory of the desire itself can help SP in belief calculation, then we might expect exactly such an advantage, even in the absence of additional video cues. We have seen that Riggs & Robinson (in press) failed to find evidence for in a paradigm that, though action-based, did not use a strongly desire-based action.

It is not completely implausible that the standard deceptive box task might have an element of desire to it (e.g. the children may say "Smarties" because they want some), but in Riggs & Robinson (in press) this is less likely because the deceptive box was an eggbox rather than a Smarties tube. It is possible then, that a Sally-Anne type action situation, such as that used in Experiments 7.1 and 7.2, would provoke good recall of false belief, even with no extra video cues.

7.8.3 Self-other differences and the prelief hypothesis

The results of Experiments 7.1 and 7.2, taken together, are not readily explicable in terms of the prelief hypothesis (Perner, Baker & Hutton, 1994; Perner, in press). Perner has argued that children might offer "think" answers to belief attribution and explanation questions on the basis of their concept of prelief, rather than on the basis of true belief understanding (see for example Chapter 1, Section 1.7.1; Chapter 6, Section 6.1.2 and this chapter, footnote 35). This objection is a possible response to some of the results reported here.

I intend to argue that it is not a plausible objection for two reasons. Firstly, as far as the explanation measure in particular is concerned, the belief-condition explanations referred to a range of belief-based states, including for example, "because that's where I remembered the biscuit was...". "because I thought the biscuit was still there..." and "because I didn't know that the biscuit had been moved...". Perner argues that "think" may be understood in terms of a mixture of pretend and belief, but has not made the claim that such a wide range of epistemic attitude verbs may also be used to cover the same simpler prelief conception. If this were to be his claim, then some account would be required of just how one could ever tell from belief-based explanation alone that a child (or adult for that matter), had reached beyond the prelief stage.
Secondly, though both the explanation question and especially the belief attribution question might be vulnerable to a prelief objection, Perner offers no reason to expect there to be a Self-Other difference in understanding prelief. Since Perner's account, being a CS theory, is competence-only rather than competence-and-performance based, any explanation would need to invoke some extra conceptual requirement that Other belief reasoning has over Self belief reasoning. However, in no CS theory, including Perner's (1991, in press), is reasoning about others assumed to be conceptually more demanding than reasoning about the self.

7.8.4 Simulation versus theory revisited

I argued in Section 7.1.2 that Self-Other differences were not relevant to the Cartesian assumption on mind: the issue of whether current A consequence of this is that an advantage for Self reasoning over Other cannot help adjudicate in the simulation versus theory debate (Davies & Stone, in press: see also Chapter 2). However, there is an alternative sense in which "simulating" may play a role within a theory-theory approach to commonsense psychology. I alluded to this in Chapter 2 and the introduction to this Chapter (Section 7.1.2), and I address it in more detail now.

This weaker sense of simulation views "simulating" as one of the abilities used to deploy theory of mind knowledge. Leslie & German (in press) suggested that "putting oneself in another's shoes" might be one way of calculating people's belief contents, predicting their actions, or imagining their complex emotional reactions. Recall from Chapter 2 that this is knowledge based simulation because the reasoning takes place within a framework where theoretical entities (specialized representations of propositional attitudes), are being attributed. For example, one way for me to form expectations about what a given person will do is to imagine that I share that person's beliefs and desires, and decide what I would do in their place. If in this situation I were to witness a person who was going through an experience which I had recently been through, I might be able to exploit the similarity between my recent experience and the situation I was modelling. This would make the "simulation" process easier than if I had never encountered the situation before.

Essentially, the children in the Other condition in Experiments 7.1 and 7.2 were faced with exactly such a possibility. It is conceivable that if they were capable of such a
simulation process, and were inclined to make use of it, their performance in explanation
would be better than children who had not been exposed to the situation at all. An
improvement that could be made to the Experiments presented in this Chapter would be to
present videos of other children's incorrect search to a group of children who have not been
through the procedure themselves. If this "non-exposed" control group were to perform
worse at belief attribution than a group who received the Experiment 7.2 Other condition
(that is, having been through the procedure themselves), then one could argue that the
children in the Other group made use of their experience to help their belief calculation:
evidence for a form of simulation.

One further avenue that this new Self-Other paradigm might offer then, is a way of
testing specific hypotheses associated with the idea of simulation as a theory of mind based
ability.
CHAPTER 8

Summary and conclusions

8.1 Introduction

In this last chapter I will draw some conclusions on the basis of the seven chapters that have gone before. I will do this by presenting a summary of the main empirical findings: evidence for early competence (Chapter 3); desire-based explanation (Chapter 5); causes and conditions in explanation (Chapter 6) and Self-Other issues (Chapter 7), and drawing these findings together in terms of two themes that have run through the thesis.

The first theme concerns the relative merits of two theory-theory approaches; accounts in terms of continuity of competence coupled with a recognition of performance factors have been more consistently supported in the thesis than accounts that seek to explain shifts in theory of mind performance in terms of shifts in theory of mind competence alone. The second theme concerns young children's understanding of desire, and how it affects the calculations involved in theory of mind tasks under different conditions. The two themes are not unrelated: the use of desire reasoning alone (in prediction and explanation) has been hypothesized here to be less resource demanding than reasoning that takes belief into account. As well as the papers discussed throughout the thesis, the "lag" between belief and desire understanding in the preschool years has been the topic of very recent debate (e.g. Wellman & Bartsch, in press; Harris, in press). In the last section of this chapter, I will go on to make some speculations about how the ideas presented in this thesis could be developed with special reference to the second, desire related theme described above.

8.2 CS versus CC theory

In the first chapter I described the theory of mind literature within the framework of two main classes of theory-theory. Although specific theories were detailed, the emphasis was placed on the general approach adopted by each class of theory in explaining the robust shift in performance between age three and four on a wide range of theory of mind tasks. I recap on each approach here and develop the summary of my findings in terms of this
general distinction between types of theory-theories, rather than in terms of specific theories.

On the one hand, the theories I termed competence shift (CS) theories were described as offering an explanation in terms of stage-like developmental change. It is considered that the child's shift in performance, on this view, reflects a genuine shift in theory of mind competence: the younger child has fewer, or less well developed concepts than the older child, and development is about the younger child acquiring these absent or deficient concepts. The shift to a representational understanding of mind is indexed by the passing of standard false belief tasks. The "consensus" view among developmentalists in this area is claimed to be some or other version of CS theory (e.g. Flavell, 1988; Ferguson & Gopnik, 1988; Perner, in press; Wellman, 1990; Gopnik, 1993a).

By contrast, the class of theories that I termed competence continuity (CC) theories was described as offering an explanation for the performance shift in terms of a stable competence and gradually improving efficiency in the deployment of that competence. This view maintains that the younger child has essentially the same concepts as the older child, but is only able to adequately use these concepts under a restricted set of conditions. The theories differ in precisely which conditions are hypothesized to make a difference to performance (e.g. Fodor, 1992; Leslie, 1994; Mitchell, 1994; Russell et al., 1991; Wimmer & Weichbold, 1994).

8.2.1 Prediction tasks

A common aim in the theory of mind literature has been to assess the extent to which young children’s conceptual competence is masked by other characteristics of the tasks used to assess it (e.g. Mitchell & Lacohée, 1990; Siegal & Beattie, 1991). In this vein, in Chapter 3 I presented a test of Fodor’s (1992) CC theory which hypothesized that children’s performance on false belief tasks was explicable in terms of their following certain, computationally simple, desire-based predictions; in situations where the agent’s desire offers a unique behavioural prediction, the child fails even to consider the agent’s belief. This strategy is in fact made possible by the normativity of belief. Beliefs ought to be true, and because for the most part they are, to predict on the basis of desires alone is a simple and quite reliable way of predicting action. Children were hypothesized to exploit this simple strategy in most situations, but also to be able to compute belief in cases where the desire-
based strategy fails.

The results in Chapter 3 were encouraging for Fodor specifically, and for CC theory-theories in general. Children were shown to be better able to predict action in situations where the strong desire-based prediction was not available. In concluding Chapter 3, in the light of possible "dumb strategy" accounts of the improved performance, I discussed the problems inherent in building sufficient controls for "guessing strategies" into false belief tasks while at the same time not overloading the task's complexity. It was suggested that the risk of observing false positives should not be overemphasized to the point where the task is in greater danger of producing false negatives.

8.2.2 Explanation tasks

Attention was then turned to backwards reasoning tasks, and in particular the explanation of agents' actions, as a measure that might converge with the false belief prediction tasks used in Chapter 3. However, in the light of the discussion of desire-based prediction above, immediate questions were raised concerning the possibility of desire-based explanation. In Chapter 4, I reviewed the backwards reasoning literature, which is mostly concerned with belief attribution measures, paying specific attention to experiments where open-ended explanation questions were employed. One principal achievement made in Chapter 4 was the construction of an explanation scoring methodology that would allow explanations to be assessed in terms of what their resource requirements were. The categorization scheme for children's explanations was designed to be sensitive to how much information the children had brought to bear on the explanation task.

In terms of CS theories, predominantly desire-based explanations could be taken to indicate that younger children had not yet reached a stage of belief competence. Indeed, some versions of CS theory had been identified in Chapter 1 as including a specific "desire psychology" stage. For CC theories, the interesting concern was whether desire-based explanation, on account of its computational simplicity, might be regarded as the primary means of explanation rather than the only means, as suggested in CS theory.

The experiments presented in Chapter 5 were presented within Davidson's (1963) reason-action framework. The simple experiments reported there addressed ideas of desire
primacy, derived not only from Davidson's observations but also from the CS and CC compatible ideas expressed above. Children's explanations were collected for a range of simple actions, presented in situations less complex than the standard false belief task in order to reduce the variety of explanations that children would offer.

The results of the Chapter 5 experiments indicated that in most circumstances desire explanations were preferred over belief explanations, even by children at an age where proficient belief attribution should have been expected. Belief explanations were more common in some situations, but only situations where the simplest desire explanation strategy was not available. These results were used to produce a simple model for explanation based on the idea that there was computational complexity associated with correctly calculating belief contents (e.g. Fodor, 1992), and that explanation therefore relied on the calculation of simplest cause.

Chapter 6 presented an elaboration of the Chapter 5 model, and looked specifically at the relationship between prediction of action and subsequent explanation. The results showed very clearly that the explanations that children offer for a given action can be affected by the immediate structure of the task. An important aspect of the results was that all the children assessed were beyond the "competence-shift", as indexed by passing the false belief prediction task. This allowed the strong development of the CC theme because the shift between failing false belief and passing false belief, the shift to a representational understanding of mind, was therefore unavailable as an explanatory construct in this experiment. By contrast, the idea that performance constraints relax gradually over the first four years of life and beyond is central to the CC position. The explanation strategies actually offered, in terms of hypothesized increasing resource requirement, were indeed related to children's increasing age, again in line with CC and performance notions.

The results in Chapters 5 and 6 have an unusual slant in comparison to other findings that support CC over CS versions of theory-theory. As I indicated above, a common means of casting doubt on the stage-like model of development, is to find tasks which, though sharing the same conceptual features, have structures which allow 3-year-olds to perform better than one would expect given their stage of development as indexed by a standard false belief task. This strategy was adopted in Chapter 3 in the test of Fodor's theory. By contrast, the strategy used in Chapters 5 and 6 has applied the same logic, but has concentrated on
showing that in explanation tasks, older children capable of passing false belief, show local
differences in the pattern of performance that undermine the idea that passing the false belief
task in itself guarantees a certain level of performance in other theory of mind tasks.

It is of course open to the CS theorists to suggest that performance factors have an
influence after (or before and after) a shift in competence but this strategy then brings into
question the need for a shift at all. One of the attractions that CS theory has (for CS
theorists), is the presumed ability to dispense with "information processing" requirements.
Indeed, such performance accounts are claimed run the risk of being "unparsimonious"
(Gopnik, 1993a, p9), compared with the simple all-encompassing performance shift.

In Chapter 7 the study of explanation came full circle. Chapters 5 and 6, as described
above, showed that the open-ended nature of explanation procedures could sometimes result
in children beyond the hypothesized competence-shift offering computationally simple
explanations, and at some other times more complex explanations. These chapters allowed
us to discover something about the performance factors that could affect the specific
explanations that children offer. In Chapter 7, an explanation paradigm was used, but the
emphasis was replaced on the attribution of belief, in order to address the question of Self-
Other differences in theory of mind in children before the competence shift. This appeared
at first glance to be somewhat of a departure from what had gone before, but the themes
that I have identified as running through the thesis were also in evidence in Chapter 7.

Chapter 7 used an action explanation procedure, coupled with presenting children
with video evidence, to demonstrate a difference between Self and Other on two theory of
mind measures: belief-based explanation and belief attribution. The experiment made use of
advantages that action-based paradigms were argued to have over deceptive box paradigms
in helping a child recognize his or her past belief. This advantage was traced in part to the
fact that action-based tasks fix the false belief in the context of a salient desire on the part
of the child. Findings in both Chapters 5 and 6 had suggested that desire frustration needed
to be especially salient in order to provoke calculation of the agent's belief.

CS accounts of theory of mind were argued to have no principled reason to expect
Self-Other differences. Such positions are thus embarrassed by these findings, since no
plausible conceptual difference exists that could explain different performance for Self and
Other in this experiment. However, as far as CC theory accounts are concerned, any
manipulations that reduce the performance requirements of the task are expected to improve performance. Children's memory for their own past desire was hypothesized to have helped them in the selection process when asked about their own past belief.

8.3 Further research questions

In this section I outline some further questions that arise from the experiments presented here. I will divide the suggestions into those that pick up on immediate consequences of the experiments presented here, including modifications that would have improved the force of the arguments developed in the thesis, and those that extend the ideas that have been developed here into wider areas. Specifically, in pursuit of the second goal, I will speculate about the relationship between causal reasoning in the theory of mind domain and causal reasoning in general.

8.3.1 Improvements and extensions in theory of mind explanation

One of the weaknesses of the support for the themes presented in this chapter is that, despite the fact that performance factors have been shown to affect younger children in prediction (Chapter 3) and younger and older children in explanation (Chapter 7 for younger; Chapters 5 and 6 for older), there is no single demonstration that the same performance factors that affect younger children have a similar effect on older children in an explanation or prediction experiment. For example, in Chapter 3 the effect of manipulating the Ambiguity of the agent's desire was shown in children younger than four but not in the oldest group of children. Likewise, the effect of making a successful action prediction on subsequent action explanation was demonstrated in 4-year-olds (Chapter 6) but not in younger children. In fact, in Chapter 7, the cross-experiment comparison (Experiment 7.1 versus Experiment 7.2) indicated that no effect of prediction on explanation occurred in the 3-year-olds.

To resolve this problem it would be desirable to show that for example, the performance factors hypothesized to be responsible for children adopting one or other explanation strategy in Experiment 6.1 (see Sections 6.6.2 and 6.6.3), could be shown to have an effect in children the same age on some types of prediction task. The children identified in Chapter 6 as having "high resources" (see Figures 6.4 and 6.5) would be
perhaps expected to pass even more complex versions of belief attribution tasks: for example the second order belief attribution task (Perner & Wimmer, 1985; Perner, 1988). Conversely, children identified as low resource would be expected to do worse on more resource demanding theory of mind prediction tasks. In addition, the type of manipulations that have been effective in reducing the performance requirements of theory of mind tasks (e.g. overcoming prepotency in Chapter 3; helping with selection in Chapter 7), would be expected to help (the low resource) children who failed more complicated theory of mind tasks such as those mentioned above.

Testing the predictions expressed in the above paragraph would be one step that could be taken in confirming the picture that has been pieced together in this thesis. If the results turned out to be favourable, the CC position would have support for the notion that the same performance resources are responsible to patterns of task performance in explanation and prediction, evidence which is not provided in this thesis.

On a slightly different note, the results of Chapter 7 could be used to develop recent attempts to teach autistic individuals certain theory of mind concepts such as "emotion", "play" and "belief" (e.g. Hadwin, Baron-Cohen & Brown, 1993; Swettenham, 1993; Whiten, Irving & MacIntyre, 1993). A strategy that has not so far been used for teaching is to confront children with their own false belief-based actions. The attempts have instead for example, focused on children learning by observation (Whiten et al., 1993). The results in Chapter 7 suggest that confronting children with direct experience of false belief may be a profitable way of progressing, although the prospects of succeeding such that autistics are then able to transfer strategies they learn to everyday situations are not certain.

8.3.2 Domain specificity in causal explanation

As far as extending the work presented in this thesis to other areas is concerned, I intend to make some speculations about domain specificity in theory of mind. The principal evidence for the domain specificity of theory of mind come from the research on children with autism (see Section 1.9.3). Initial research yielded contradictory findings when autistic individuals were assessed for their understanding of desire. A potential problem for the domain specificity account would exist if autistics were shown to be unimpaired on some desire-based tasks (e.g. Tan & Harris, 1991; Baron-Cohen, 1991b; Tager-Flusberg, 1992).
This evidence for simple desire understanding in autism has raised the possibility of a fractionated ToMM (Leslie, in press) or indeed of a multi-modal basis for theory of mind (Baron-Cohen, 1994).

Some very recent findings, however, indicate that autistic children do find more complicated desire-satisfaction and desire-change tasks difficult (Phillips, Baron-Cohen & Rutter, 1994), find reading desire from eye-direction difficult (Baron-Cohen, Campbell & Walker, 1992) and finally have difficulty with tasks where children are required to understand the more complicated, although desire-based mental state of intention (Phillips, 1993).

In this vein, the explanation tasks presented in this thesis might serve to shed more light on the autistic child's theory of mind. Desire explanation was demonstrated to be common even among children who pass false belief tasks, and this was hypothesized to be a result of performance considerations in normal children. The question of whether autistic children can offer simple desire-based explanations for their own actions (Chapter 7) and the actions of others (Chapters 5 and 6) is thus an interesting one with respect to what are the conclusions that can be drawn about the normal development of theory of mind. The expectation on the basis of the tasks described above that showed good performance on aspects of desire (e.g. Baron-Cohen, 1991; Tager-Flusberg, 1992) might be that though children with autism may be able to refer to desire in explanation, belief explanation would not be invoked by changes in task structure, as was observed in normal children. Alternatively, on the basis of the more pessimistic picture of desire understanding in autism (e.g. Phillips et al., 1994), it might be expected that even desire-based explanation may be too difficult for autistic individuals.

The final suggestion for further research is built on the claims made in Chapters 5 and 6 about theory of mind explanation being causal explanation. The idea would be to assess the domain specificity of theory of mind by looking at how theory of mind explanation relates to more general causal explanation. In addition to the work discussed in this thesis on theory of mind explanation, there is a wide literature on children's sensitivity to physical cause-effect relations from early life (see for example Leslie & Keeble, 1987) right through to school-age (e.g. Schultz, 1982). This work, however, has not assessed how causal reasoning might be affected by the presence of goal-directed agents.
An interesting possibility raised by Premack (1990, 1991) concerns the partitioning of objects in the world according to whether they are governed by cause-effect relations or intention-action relations. Independently, Nichols (1992) suggested that the notions of causality that underlie folk psychological judgements may be very different from modern philosophical positions on causality (e.g. Mackie, 1974). The suggestion here is that these possibilities may be addressed by looking at how children locate cause in situations where physical causality and intentional causality co-occur. After all, situations often involve both the action of intentional goal-directed agents and effects of simple physical causes. This would be possible by constructing causal chains, not unlike those used in Baron-Cohen, Leslie & Frith (1986), in which mental state causes and physical causes are implicated, either alone or together, in creating some action. The pattern of explanations observed in a variety of such situations would shed light on the interaction of domains in causal explanation.

8.4 Concluding comment

Two main themes have run through this thesis which has investigated pre-school children’s theory of mind, specifically their explanation of action. First, explanation experiments have been used to develop one of the two main theoretical approaches to theory of mind: the competence continuity approach. This approach explains developments in terms of stable conceptual competence coupled with specification of the performance demands associated with theory of mind tasks. Second, the role of the concept of desire has been demonstrated to be important in both explanation and prediction, which in turn relates to the computational simplicity of desire compared to belief; theory of mind prediction and explanation that is desire-based (assuming belief contents as default), is less resource demanding than prediction and explanation where beliefs are calculated.
References


227


230


Pratt, C. & Bryant. P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development, 61,* 973-982.


