A Defence of Simulation Theory

Timothy Lawrence Short

Submitted to UCL for the degree of PhD

March 1, 2016

Thanks to Ian Phillips, who supervised the first three terms of writing this thesis. Thanks also to Lucy O'Brien who supervised the next four terms, and to Daniel Rothschild who supervised the last two. Maarten Steenhagen asked a crucial question at the beginning. Kevin Riggs made valuable comments on the first full draft. Mark Lancaster helped at the beginning in 2007. Kathrine Cuccuru read the book. Although I have never met Rebecca Saxe, it will be apparent to all readers that I am very much in her debt.

I would also like to thank the other founder members of the Gordon Square Dining Club: Margaret Hampson, Jerome Pedro, Lea-Cecile Salje, Alex Sayegh, and Tom Williams. They will have to stand in for all of the staff and students of the UCL philosophy department who have made the last eight years such tremendous fun.

Declaration

I, Timothy Lawrence Short, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

In this thesis I defend the Simulation Theory of Mind against the Theory Theory of Mind. I do this in two major ways. Firstly, I set out the logical space available to accounts of Theory of Mind and suggest that there are many valuable options available to simulational accounts. I also canvas serious objections to Theory Theory which have not I contend been resolved. I will argue that hybrid theoretical accounts do not resolve all of these objections. Further types of hybrid accounts which add in some simulational capacities, some of which involve both theory and simulation, are complex and unparsimonious and so a different approach is needed. I argue for a specific weak hybrid approach which is very close to pure Simulation Theory. This avoids all of the objections. Secondly, I provide an answer to a challenge to Simulation Theory which is widely considered to be its single most significant problem. That challenge, termed the 'argument from error,' is that while Simulation Theory can account for frequent error in Theory of Mind, it cannot account for the systematic nature of those errors. My response is a novel Bias Mismatch Defence. This suggests that the systematic errors can arise because cognitive biases, such as Confirmation Bias, can have differential effects in the person simulating and the person being simulated.

Word count: 79444 (from LaTeX statistics; excluding bibliographical appendices as is permitted by UCL regulations)

"...οἰα καὶ ἡμιν Ζεὺς επι εργα τίθησι διαμπερες εξέτι πατ ρων.
οὐ γὰρ πνγμαχοι εἰμὲν αμύμονες οὐδὲ παλαισται ἀλλὰ
ποσι κραιπνως θεομεν..."

"I want you to be able to tell your noble friends that Zeus has given us too a certain measure of success, which has held good from our forefathers' time to the present day. Though our boxing and wrestling are not beyond criticism, we can run fast ..."

Homer: The Odyssey, Book VIII[1]

[1]Since this text appeared at Short (1992, p. 3), it appears here also.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We seem to understand one another. How do we do it? When does it go wrong? These are the two questions I will explore in this thesis. Humans seem to be able to predict one another's behaviour and explain it. Indeed, we spend much of our time happily engaged in these activities. The label for this way in which we predict and explain each other is 'Theory of Mind.' This term is perhaps slightly unfortunate; as Dennett (2007, p. 396) comments, it conjures up too much "theorem-deriving" and "proposition-testing." I will be arguing in this thesis for less theoretical and more imaginative answers to the questions as to how we know each other and ourselves. I will be arguing for an account whereby we understand others by putting ourselves in their shoes.

The term Theory of Mind is generally agreed to originate in the seminal Premack and Woodruff (1978) which asked "Does the chimpanzee have a theory of mind?" The question there was whether the chimpanzee has the ability to predict or explain the actions of others on the basis of beliefs or perhaps quasi-beliefs about the mental states of those others. It was taken as read that humans have those abilities: persons can in fact so predict and explain. Humans do have, then, a Theory of Mind, or at least Theory of Mind abilities. People know each other because of it, or they think they do.

Theory of Mind abilities have also been known as 'mind-reading' or 'mentalising,' because on some views, persons predict behaviour by first ascribing mental states such as beliefs and desires to others and then working out what people with those beliefs and desires would do. Accounts of Theory of Mind that explain how people predict each other's behaviour have fallen into two competing types: Simulation Theory and Theory Theory. This thesis will defend Simulation Theory of Mind against Theory Theory of Mind. These terms are often shortened to 'Simulation Theory' and 'Theory Theory.' Something should be said at the outset about these terms, since at least the latter one looks somewhat odd.

The oddity of the term 'Theory Theory' derives from its repeating the word 'theory.' This is intended to drive home the two domains of theory involved. Firstly there is the theory in Theory of Mind which is just the label for whatever mechanism I use to predict your behaviour from the theoretical knowledge that you have a mind which, presumably, means you have beliefs and desires as well. The second usage of the word 'theory' serves to underline that on the Theory Theory view, how I predict your behaviour —how I can use my Theory of Mind —is that I employ a theory to do so. The contrast is with Simulation Theory, which says that I predict your behaviour not by employing a theory of people, but by simulating you. My Theory of Mind on the simulationist account would be more like 'that's what I would do if that were me' and less like 'as a rule, people in situation X do action Y.' It would be more human and less scientific in construction.

The theory or simulation underlying Theory of Mind should not imply flawless performance. We need to explain the observed performance of Theory of Mind, which varies from good under some circumstances to poor under others. For example, I believe that if I see you going into a coffee shop, I have a good picture of some of your desires and beliefs: viz. you desire coffee and

you believe that you will be able to get some in the coffee shop. So I can explain your behaviour when you go in. On the other hand, you may well be involved in more complex scenarios that defy my Theory of Mind abilities. I may be mistaken about your purposes in going in to the coffee shop; perhaps you do not desire coffee but you believe you will meet a friend. Indeed, errors in Theory of Mind are legion, and it is consideration of these errors that will form a major part of this thesis. That is because there is a serious challenge from Saxe (2005a) as to how one explains the systematic nature of these errors. She says that the inability of Simulation Theory to explain the systematic errors combined with the ease with which Theory Theory can explain the errors is a major reason to prefer the latter over the former. I agree with her that this is a serious challenge, but I disagree that Simulation Theory cannot explain the systematic errors. I will argue that not only can Simulation Theory explain the systematic nature of these errors, but it can do so better than Theory Theory, because it is more parsimonious and more plausibly ascribed to children who have a serviceable Theory of Mind by the age of five at the latest, among other reasons. Simulation Theory alone is clearly more parsimonious than the current consensus position which is a poorly specified 'Strong Hybrid' of simulation and theory. I will explain the differences between Strong and Weak S/T Hybrid accounts in more detail at the beginning of Ch. 4, but the basic distinction is that Strong S/T Hybrid accounts allot significant roles to both simulation and theory while Weak S/T Hybrid accounts do not. Weak S/T Hybrid accounts could either be mostly theory with a minor amount of simulation or vice versa: it is this latter 'pure simulation plus minor theory' account for which I will argue in this thesis.

Using Theory of Mind is part of 'folk psychology.' This is distinct to scientific psychology, which is the sort of activity conducted in university research laboratories. Both sorts of knowledge aim at understanding people, but the

first one is conducted by everyone more-or-less all the time, while the second one is a specialised academic discipline. I will be aiming in this thesis to make a contribution to the second by providing a new approach to the first. Or more precisely, to provide a previous approach to understanding the first, Simulation Theory, with the resources to defeat its most serious challenge. This will also enable the necessary defence of the position I favour, Weak S/T Hybridism against the same charge.

I have one task in this thesis; I will engage with this task by pursuing two major and linked aims. The task is, approximately, to defend simulational accounts against the more mainstream theoretical accounts of Theory of Mind. The two major aims intended to provide this defence are as follows. The first aim is to support simulational accounts against theoretical accounts by noting the serious objections to the latter that the former can avoid. The second aim of this thesis is to respond to a systematic error challenge to simulational accounts.

To begin the pursuit of this first aim, I will be setting out in detail the logical space which defines possible accounts within the simulational/theoretical domain. This will show that there are more options available in the simulation space than have currently been explored. I will be clearly setting out the debate between simulation and theory and making more clear what the relevant variations of these positions are.

I will then establish that the theoretical accounts so far canvassed come in two major variants. I will be assuming that all theoretical accounts postulate that there are a set of rules or axioms or generalisations which represent the body of theoretical knowledge that underpins Theory of Mind. Where the two sorts of theoretical accounts differ is on the source of these generalisations. For some proponents of a theoretical account, these generalisations are learned, while for others, they are innate. I will consider three objections to the first

type of account and three further objections to the second type of account. (I will set out what these objections are in more detail in the chapter outlines below.) These will show that there are serious problems with both types of pure theoretical account. I will then show that combining the two theoretical accounts does not resolve all of the objections.

I will begin the pursuit of the second aim by setting out the systematic error challenge to simulational accounts as propounded by Saxe (2005a). Although for the sake of specificity, I will generally use Saxe's position as the one which I oppose, her view is a mainstream one which is widely defended. For example, Apperly (2008, p. 268) writes, "many authors now argue for a hybrid account in which both Simulation and Theory play a role." Saxe is within the mainstream as a Strong S/T Hybrid theorist who sees major roles for both simulation and theory in accounts of Theory of Mind. It is this entire mainstream consensus that I challenge; including its significant reliance on theory.[1]

The central support for this consensus, as Saxe (2005a, p. 175) argues, derives from the fact that there is "occasional systematic error" in ToM. This argument is known as the 'argument from error.' The sort of case she means may be exemplified by the notorious experiments in which participants believed that they were giving severe electric shocks to others. The Theory of Mind error is that no-one predicts that the subjects will give the shocks. The errors are also systematic in that they seem to occur repeatedly: every time a naive subject makes a prediction about how people will behave in the Milgram (1963) experiments, that prediction will be wrong. I will not dispute that these errors occur, nor that they are systematic in nature. I will instead seek to provide additional resources to Simulation Theory in a parsimonious fashion to allow it to explain the systematic nature of the errors. This will

[1] Though see Wilkinson and Ball (2012, p. 265) for the suggestion that the "hybrid consensus is perhaps more apparent than real."

also constitute a defence of the Weak S/T Hybrid account which I favour since Weak S/T Hybrid accounts are very close to pure Simulation Theory.

Saxe holds that the systematic nature of these errors is easily explained on Theory Theory and not at all explicable on a Simulation Theory basis. She is joined here by a large number of writers including Apperly (2008, p. 268) again, who goes on to observe that "cases where people make systematic errors [...] are seen by many as good evidence" for Theory Theory. He gives only two citations in support of this claim, of which Saxe (2005a) is one. Many other authors make similar comments about the unique importance of Saxe's argument in bolstering support for theory and thereby reducing support for simulation. For example, "as Rebecca Saxe (2005) argues, there are both new and old data that speak strongly in favour of a substantial theoretical component to our folk-psychological capacities" (Godfrey-Smith 2005, p. 8). Dimaggio et al. (2008, p. 786) set out a simulationist approach to Theory of Mind, and single out the argument from error as an obstacle, beginning with the phrase "[o]f note there are limitations to this view." Saxe (2005a) is the major anti-simulation argument considered by P. Mitchell, Currie, and Ziegler (2009b, p. 535). Morin (2007, p. 1069) writes that ST "is largely accepted in the literature (but see Saxe, 2005)." Grafton (2009, p. 109) describes Saxe (2005a) as "an important review [that] provided a detailed analysis of behavioural errors in intentionality decoding experiments [which] is a strong argument against the conclusion that simulation" is sufficient to explain the decoding. In sum, as Doherty (2008, p. 47) points out, the " 'argument from error' (Saxe 2005a) is one of the most powerful arguments against" ST.[2]

Clearly, responding to this charge that Simulation Theory cannot explain systematic error is of the first importance. However, as far as I can see, there

[2]Cf. also Bello and Cassimatis (2006, pp. 1014–1015); Kaplan and Iacoboni (2006, p. 182); Oberman and Ramachandran (2007, p. 316); Nico and Daprati (2009, p. 233); Gallese and Sinigaglia (2011, p. 512).

has been no significant response at all to this challenge from the Simulation Theory side, although Saxe (2005a) is comprehensive, clear and widely cited. This lack of a response to Saxe (2005a) has driven the consensus in favour of Strong S/T Hybrid views of Theory of Mind involving both simulation and theory. The absence of a comprehensive response from the Simulation Theory side lets the Theory Theory side win by default. In this thesis, I will supply this lack.

My main response to Saxe (2005a) is going to be that cognitive biases, to which persons are all subject, explain the systematic errors. As an example of a cognitive bias, I mean such effects as confirmation bias. This is the tendency people all have to seek only information confirming what they already believe. Often, the application of these biases is caused by emotional reasons. For example, most people want to believe positive things about themselves, and sometimes people do that by ignoring evidence to the contrary. If the person doing the simulation has different emotional responses to the person being simulated, they may well not apply the same biases. For instance, someone else might be emotionally involved in maintaining their own positive self-image, but I might not be. If that emotional involvement leads them to apply any cognitive biases, that bias may not feature in my simulation. Thus my simulation will exhibit systematic error. I will use this approach to explain a wide array of experimental data to which Saxe appeals to back her Strong S/T Hybrid consensus view. I term this defence the Bias Mismatch Defence because it relies on the simulator and the person being simulated applying different biases to explain the errors in theory of mind and their systematic nature.

The arguments for Simulation Theory that rely on the discovery of 'mirror neurons,' put forward for example by Gallese and Goldman (1998), lie outside the scope of this thesis. While convincing, the results are heavily disputed by

Theory Theorists. I suspect that a full consideration of the current state of this evidence and the surrounding arguments could only be adequately done in a separate book-length treatment.[3]

I will proceed as follows. In Ch. 2, I will give an overview of the various accounts of ToM. I will set out very briefly at this stage some initial motivations for considering simulational as opposed to theoretical accounts. I will give descriptions of the two main types of each of Simulation Theory and Theory Theory. I will show how the logical geography of Simulation Theory results in an array of possible variants of Simulation Theory, for some of which arguments have been given. This brings out a major risk: that collapse of Simulation Theory back into Theory Theory. Such collapse would mean that ST is not a separate defensible position from Theory Theory. I give reasons to think that this threat can be avoided. I close this chapter by considering the important problem of 'setting the bar too low.' This involves Theory Theory proponents proposing too easy a test for whether Theory of Mind has involved theory use.

In Ch. 3, I consider six objections to pure Theory Theory accounts. There are three objections given to each of the two theoretical accounts of Theory of Mind: the variant on which the generalisations are learned and the variant on which they are innate. The three objections to the learned variant of theoretical accounts are as follows. Such accounts a) implausibly ascribe mastery of complex and difficult sets of generalisations to very young children; b) require that a solution of the intractable frame problem be embodied within the generalisations and c) entail convergence between the Theory of Mind across different persons and different cultures which is empirically false. The three objections to the innate variant of theoretical accounts are as follows. Such accounts a) cannot explain the observed development in Theory of Mind ca-

[3] P. Mitchell, Currie, and Ziegler (2009b) survey the mirror neuron evidence for Simulation Theory.

pacities; b) cannot explain how persons usually start Theory of Mind tasks by assuming that other persons share most of their beliefs and c) lack a parsimonious explanation of certain features of autistic subjects. I conclude that none of these pure accounts can overcome the objections to them. This leads on to the project of the next chapter, which is to consider whether hybrid views involving mixtures of various types of account can avoid these objections.

In Ch. 4, I begin by examining whether the six objections noted in the chapter above can be avoided by the combination of learned and innate theoretical accounts. This move obviously has some costs in terms of the explanatory power versus simplicity value metric of accounts generally, but does pay some dividends in terms of the six objections. I will conclude however that the combination leaves some serious objections unresolved. I then go on to note that the consensus nowadays is for a Strong S/T Hybrid position, which holds that both simulation and theory play a major part in Theory of Mind. There are two sets of problems for this view. One set relates to its inheritance of all of the problems set out in the previous chapter for pure theoretical accounts. The second set of problems derives from the Strong S/T Hybrid nature of the consensus, which means an account of interaction between theory and simulation is required. Will they answer separate questions, or somehow work together? I contend that all of these problems taken together mean that Theory Theory and strong hybrids are unsuccessful, and weak hybrids which are very close to pure simulation accounts are the best remaining option. This means that a major unanswered problem for simulational accounts must be answered, which will be the project of the next five chapters.

In Ch. 5, I will outline the unanswered problem for Simulation Theory: the 'argument from error.' Saxe (2005a) argues that Simulation Theory cannot account for systematic errors in Theory of Mind in certain circumstances, because if people use their minds to simulate other minds, the simulators

should be accurate. This chapter aims to give Saxe her best case in two of the areas she considers: occasions when Theory of Mind is too cynical, others when it is too rosy.

In Ch. 6, the question as to why we need a new defence is answered by agreeing with Saxe that the existing defences do not work. The Bias Mismatch defence is introduced: 'simulation may not accurately model bias' is the central idea. A list of biases that will be employed, for example Confirmation Bias, is given and each is outlined.

In Ch. 7, three reasons why biases may not be simulated are given. There are two main ways: Affect Mismatch and system mismatch. In the first, the emotional impact on the target is not fully felt by the simulator. In the second, they use different reasoning systems.

Ch. 8 covers an array of 'too rosy' evidence introduced by Saxe (2005a), which arises in situations where people are systematically over-optimistic in predictions of the rationality or morality of ourselves and others. For example, no-one predicts the way participants in the Milgram experiment are prepared to give out severe electric shocks to strangers for minor infractions. These data are explained by appealing to Conformity Bias, the tendency to do what one is told. A set of 12 experiments Saxe cites in support of her challenge is described and explained using the Bias Mismatch defence in similar fashion to above.

In Ch. 9, I turn to the opposing sort of data introduced by Saxe (2005a); it covers occasions when persons are systematically too cynical in Theory of Mind. For example, persons on different sides of vexed political questions often form very harsh evaluations of their opponents. They see their opponents as biased and unwilling to examine the evidence or assess it impartially. This is explained using the Bias Mismatch defence with the bias in question being Confirmation Bias. People might be more sympathetic to their opponents if

they realised that people all fall victim to it. Nine further experiments are similarly explained.

In Ch. 10, I will examine whether the Inaccurate Generalisation Defence can allow theory theory to explain the systematic theory of mind errors which represent the explanatory problem. This defence claims that in every case where systematic theory of mind errors are observed, this results from an inaccurate generalisation in ToM. For example, in the case of the Milgram data, the inaccurate generalisation might be something like 'people will not generally harm others without justification.' While this is successful, I will argue that in other cases no plausible inaccurate generalisation can be found and there is no reason to expect adults to possess such an inaccurate generalisation. I will therefore conclude that the Inaccurate Generalisation Defence of theory theory fails. I will close by offering brief Conclusions in Ch. 11.

Henceforward, I will throughout adopt abbreviations and terminology common in the literature. Theory of Mind becomes ToM. I will generally use the abbreviations ST and TT in common with Harris (1992, p. 120), who writes of the debate "between advocates of the simulation theory (ST) and the theory-theory (TT)." I will adjust citations where necessary to reflect this usage. I will also follow Harris (1992, p. 121) when he suggests that we "suppose that a simulation allows the subject (S) to identify the particular emotion, desire or belief that another person (O) currently entertains." What this means is that a person, the subject or S, is using ToM to predict the behaviour of a person, the object of ToM or O. S and O could also be Self and Other, but note that O could really be another person, or equally S at a different time or in a counterfactual situation. The idea is that persons also use ToM to predict what they themselves might do in the future, for example. Often in the literature, authors will refer to the simulator and the simulatee; the subject and the object; the person who is simulating and the target of the simulation; a

person considering what they themselves might believe and desire at different times or under counterfactual circumstances. As said, I will for the benefit of clarity replace all of these terms with the use of S and O.

I will consider two types of TT in this thesis, which I will term TT(Scientific) and TT(Innate). The first form of TT holds that ToM generalisations are learned via processes that are usefully analogous with the processes involved in making scientific progress. The second form of TT holds that the generalisations are learned. So both of the forms of TT which have been widely supported in the literature are based on generalisations. I will not consider in this thesis the possibility of new types of TT which do not involve the use of generalisations. One reason for this is that one might well think that the existence of generalisations is essential to theories; I am sympathetic to that view. It is also the case that this thesis is in large measure a response to Saxe (2005a) which is based on TT(Scientific). If a form of TT can be constructed without generalisations, then it would remain to be seen how plausibly it could still retain the theoretical characteristics required to be a form of TT and how well it performed as an account of ToM. Slaughter and Gopnik (1996, p. 2967) offer a definition when they state that "[i]ntuitive theories are defined as coherently interrelated systems of concepts that generate explanations and predictions in a particular domain of experience." This looks very much as though generalisations will be central since the process of generating an explanation or a prediction will proceed by generalising from the concepts involved. If S generates prediction X in scenario Y, then presumably S will do so every time scenario Y or similar occurs: this is a generalisation.

Chapter 2

ToM Accounts: Overview

2.1 Introduction

The question as to whether simulation or theory form the basis of ToM abilities has been heavily debated the last couple of decades and arguably much longer; it remains open and important.[1] I will begin consideration of that debate in this chapter by analysing the competing theories. It is essential to consider TT for its own sake, but by doing so we can also learn about ST, since it was developed as a skeptical alternative to TT.[2]

There are several variants of each of TT and ST. Keeping all of the variants clear and separate is important, since there is a 'collapse risk' between the various theories. By this term is meant the possibility that one of two apparently separate theories entails elements of another, so that anyone espousing one is committed to the other even if they do not wish to be. For my project in this thesis, collapse risk between TT and ST would be a serious problem, while collapse risk between different sorts of ST would not be serious. The reason for this is that I am seeking to defend ST (or Weak S/T Hybridism) against TT, and that project would be complicated if ST and TT were found to be

[1] Cf. Nagel (2011, p. 14).

[2] Cf. Apperly (2008, p. 268).

linked in this problematic way. There would not be a separate position to defend. On the other hand, if there turned out to be a real collapse risk between two variants of ST, that would still leave some variants of ST as viable and separate from TT, which is all that is required by a defence of ST.

I will proceed as follows in this chapter. I will first in §2.1.1 complete these introductory remarks by sketching some initial motivation for considering ST. After that, in the following four sections, I will examine the two most important variants of each of our two competing theoretical and simulational accounts of ToM. First, I will in §2.2 consider the scientific variant of TT, under which ToM is theoretically based and the theory used is akin to a scientific theory. This account is widely supported in psychology and is the only one discussed by Saxe (2005a). At points in the past, it has been called just 'theory-theory,' but I will not use that term to avoid confusion, since we now have more than one theoretical account of ToM. Then in §2.3, I will examine the innate variant of TT, which also claims that ToM is theoretically based, but denies that the theory is like a scientific theory. On this Modular account, which is also known as ToMM, humans are born with the theory that underlies their ToM. Turning to the simulationist views, I will in the subsequent two sections outline the two major variants of ST. Perhaps the major difference between them is whether when S simulates, S becomes like O or rather 'becomes' O. On the first 'replication' variant, discussed in §2.4, what happens is that S examines the situation of O, places himself in that situation, introspects his own consequential mental states, ascribes them to O and then predicts what O will do, if O has those mental states. I will then in §2.5 discuss the rival 'transformation' account. Here, S simply places himself in imagination in O's situation and 'acts' accordingly, with the exception that the acts are to be ascribed to O rather than actually implemented. The transformation account denies that humans have introspective access to their own mental states.

This division into two of the simulationist accounts will involve three claims, which are either asserted or denied by the two rival theories. This suggests further possible simulationist accounts, all of which are of interest and some of which have received support in the literature. I will sketch in §2.6 what these accounts look like, but will not select a champion. As said, all that is needed for the project of this thesis is that at least one variant of ST is plausible and distinct from TT. I will provide some real-life examples of how simulation works; this task will also indicate further potential types of ST. A further possible logical space for ST theories is provided by consideration of whether they are on-line or off-line. I discuss this in §2.6.1. The idea on some ST accounts is that beliefs in the simulation context must be quarantined from normal beliefs of S, or 'held off-line' in order to prevent them from directly causing S's behaviour, which is not a ToM function.

At this point, we will have arrived at a good picture of the various competing accounts of ToM, and so we can turn in §2.7 to the collapse risk between ST and TT. Proponents of TT have laid charges at the door of ST which I will aim to refute. These are of the sort that if simulation employs any theoretical concepts, such as beliefs or desires, then it is really TT. I think this is unreasonable, because no account of human mental lives can get far without beliefs and desires. I call making this charge 'setting the bar too low,' because it is too easy for TT proponents to insist that part of ToM is theoretical if the use of beliefs and desires in ToM is enough to be theory use. If that is indeed theory use, then it is not such in any interesting fashion.[3] I will likewise have little truck with claims of the type that all simulation is theoretical, because using it involves applying the theory that 'simulation works' or 'S is like O.'

[3]See also Blackburn (1992) for discussion of overly promiscuous application of the term 'theory' in accounts of ToM.

2.1.1 Why Consider ST?

I will introduce some main motivating factors here; this topic will be covered in more detail in Ch. 3.

The central claim of ST, as set out by Friedman and Petrashek (2009, p. 115), is that "reasoning about mental states often requires attempting to make one's cognitive system mimic or replicate (simulate) another person's thoughts and feelings."[4] The motivation for pursuing the ST approach, as Stone and Davies (1996, p. 127) put it, is the fact that "when we try to understand other people, we are trying to understand objects of the same kind as ourselves." So why not assume that people use their access to their own minds to understand other minds? We do not need to introduce extra machinery here. By contrast, Saxe (2005a, p. 174) sets out the opposing TT position as the contention that "when asked to predict or explain an inference, decision or action, children and adults do not simulate the other person's beliefs in their own mind, but instead deploy an intuitive theory of how the mind works." An important motivation for ST then is one of parsimony or explanatory power with minimal 'working parts.' This will be my working definition of parsimony: lack of complexity or few moving parts combined with significant explanatory power. We can explain ToM by postulating that persons exploit the fact that they all have similar minds so we need not introduce additional theoretical machinery to explain ToM.

A second advantage for ST over TT derives from the fact that we are trying to explain ToM, which is quite advanced in five-year-old children. The claim that children have developed scientifically a more or less complete body of psychological knowledge by the age of five is already difficult to accept. That difficulty is increased if various experimenters (Onishi and Baillargeon

[4]Strictly speaking, the inclusion of 'often' means their view is technically a Hybrid, but their statement is still a good exposition of ST.

2005), (Helming, Strickland, and Jacob 2014) are correct when they report that 15-month old infants have sufficient ToM to appear to be surprised by behaviour that is not consonant with false beliefs of others. The implausibility of this scientific approach, with children or infants selecting, confirming and disconfirming hypotheses, was one motivation for TT adherents to propose the alternative innate TT account, but there are problems with that as well, as I will outline below.

ST has received significant empirical support. For example, one study looked at children with SLI —Specific Language Impairment. Farrant, Fletcher, and Maybery (2006) investigated Visual Perspective Taking or VPT, which refers to such tasks for S as stating whether O can see something from O's position, irrespective of whether S can. They note that Harris's version of ST "predicts that the development of VPT will be delayed" in subjects with SLI. Farrant, Fletcher, and Maybery (2006, p. 1844) point out that Harris "argued that language facilitates the development of the ability to simulate another's perspective because conversation involves a constant exchange of differing points of view." Thus, SLI subjects should exhibit developmental delay on both VPT and ToM tasks. This is what is indeed found: Farrant, Fletcher, and Maybery (2006, p. 1842) report that their "results supported Harris's theory and a role for language in ToM and VPT development."[5]

A further empirical argument for ST explains emotional empathy[6] in infants. Gordon (1995b) notes an observation that a six-month old exhibited

[5]Simulationists such as Gallese and Goldman (1998) and Goldman and Sebanz (2005) have also appealed to the discovery of 'mirror neurons' to support their claims. These neurons are activated when an action is performed or observed, which lends itself to simulationist accounts. This type of evidence is outside the scope of this thesis.

[6]Preston and De Waal (2002, p. 9) give an account of empathy which is "not in conflict" with ST; I would go further and say that their account is in fact highly supportive of ST since it matches perception and production of an emotion.

facial signs of sadness on seeing his nurse pretend to cry. Gordon explains this by suggesting that even from a young age, people can experience the same emotion as someone else due to a motor mimicry process. When persons observe the facial movements of someone else, this will produce similar motor activity in them. They may not know that fact directly because the motor activity can be sub-threshold i.e. insufficient to cause actual facial movement. Gordon (1995b, p. 729) notes that "motor activity, especially the movement of facial muscles, can drive the emotions." This is then a mechanism whereby people can 'catch' the emotions of others even when they could not say what those emotions were, even when they are unaware that they have done so, and even when they are six months old. All of these empirical claims are hard to explain on a TT basis but consistent with ST. It would not be the only scenario in which sub-threshold motor activation is held to explain people's understanding of others. On the Motor Theory of Speech Perception,[7] people perceive the speech of others by micro-activation of their own speech production musculature.[8]

Goldman (2006, Ch. 6) discusses several further forms of empirical support for ST, including studies of some subjects who have deficits in both experiencing and recognising certain emotions, suggesting that they have damage in a single area responsible for both.

[7]See for example Liberman (1985), Ivry and Justus (2001), Fadiga et al. (2002) and D'Ausilio et al. (2009) for the Motor Theory of Speech Perception including the sub-threshold activation elements thereof.

[8]Rochat (2002, p. 45) cites speech perception as one example among many of "common code between perception and action systems," suggesting that there are several domains in which perception and production and linked.

2.2 TT(Scientific)

I have chosen the term TT(Scientific) to refer to the scientific version of TT because the authors tend to use the term 'theory-theory' alone. Using that would be confusing since there now are multiple types of TT. As mentioned above, there are two major variants of TT. The two variants of TT are the 'scientific' view —TT(Scientific) —and the 'innate view' —TT(Innate). Both variants of TT hold that there is a body of theoretical knowledge underpinning the abilities of S to predict and explain the behaviour of O which could be expressed as a set of rules or generalisations —even though S himself need not necessarily be able to do that. Similarly, more people can apply the rules of grammar correctly than can state them. On TT(Scientific), the body of knowledge that underlies ToM is learned while on TT(Innate) it is not learned. TT(Scientific) holds that the body of knowledge is akin to scientific knowledge, with children developing by improving the body of knowledge in a quasi-scientific way. They would form and test hypotheses, discarding those disconfirmed by data. The data in question would come from observing the behaviour of other individuals.

Below I set out the claims that define TT(Scientific), as set out by Davies and Stone (1995, p. 4).[9] They begin their discussion by stipulating the definition of the thought T as follows: T = '[O] believes that P.' With that in hand, TT(Scientific) is defined by the following set of claims.

- TTa: In order to predict and explain the behaviour of O, S must be able to entertain thoughts of the form (T)

- TTb: To entertain those thoughts, S must have the concept of belief

[9]Davies and Stone (1995) discuss the TT(Scientific) definitional claims in terms of the False Belief Task. I will discuss this task later, but for now we need not restrict ourselves to one form of ToM test.

- TTc: To have the concept of belief, S must have a body of psychological knowledge
- TTd: Development of folk psychological ability is expansion of this body of knowledge
- TTe: This development may be understood as analogous to "development [...] of bodies of professional scientific knowledge" (Davies and Stone 1995, p. 4)
- TTf: "Information processing mechanisms" (Davies and Stone 1995, p. 4) are needed to use the body of knowledge

It might be questioned whether all of these claims are essential either to TT(Scientific) or other possible forms of TT. TTa to TTc appear to be necessary to TT(Scientific). TTc also appears to entail that all forms of TT which assert it mean that use of generalisations will be how ToM works, since it seems that this is how the body of knowledge will actually be constructed and used. TTd appears to be optional for TT(Scientific), though it is certainly asserted by TT(Scientific) proponents in the literature. It would be possible to construct a version of TT(Scientific) which made a claim approximately along the lines of 'development of folk psychological ability is improvement in the ability to use this body of knowledge.' It would be important to avoid collapsing into TT(Innate) of course, if the task were to improve TT(Scientific). Avoiding such collapse might be difficult if the expansion of the body of knowledge claim were replaced by an improved ability to access a body of knowledge that would be static because it would be innately specified. TTe is essential to TT(Scientific) as set out by its proponents, since it is definitional of their project that there is a useful analogy between scientific progress and ToM development. As I mentioned in the Introduction, I will not consider at any length in this thesis other possible types of TT, including types which do not

assert TTe. TTf appears to be an optional extra which in any case does not add much to the body of knowledge other than insisting that it be accessible via some mechanism.

A major proponent of TT(Scientific) is Gopnik. Gopnik and Wellman (1992, p. 145) summarise TT(Scientific) well when they write that it is: "the view that the child's early understanding of mind is an implicit theory analogous to scientific theories, and changes in that understanding may be understood as theory changes." We have here then the claim that even young children are using a theory that they have developed themselves. Any explanation of ToM must apply to young children because they have ToM capacities by the age of five. It is important that the theory postulated not be one requiring explicit reasoning, since there seems to be little phenomenology in either children or adults that is consistent with explicit theory use. That is, it seems rare for anyone to explicitly consider pedestrian sequences of deductions like 'Peter believes the ball is in the yard, Peter desires the ball, I conclude that Peter will go into the yard in order to find the ball.'[10] We also have the explicit claim that theory is analogous to scientific ones, meaning that there is hypothesis selection and confirmation. Gopnik and Wellman (1992) explain development in children's ToM abilities on the basis of changes to the theory: viz., improvements in that theory.

TT(Scientific) also looks analogous to science in what it understand a theory to be. Gopnik and Wellman (1992, p. 146) explain that TT(Scientific) involves theoretical constructs which "are abstract entities postulated, or recruited from elsewhere, to provide a separate causal-explanatory level of analysis that accounts for evidential phenomena." The abstract entities involved are the mental states of others. They must be abstract, because they cannot be observed directly. Postulating them though allows S to explain evidential

[10] I will later posit a carve-out for such explicit reasoning; this carve-out will distinguish Weak S/T Hybridism from pure ST.

phenomena of the sort generated by the behaviour of others.

The way these theoretical entities should interact with each other and the items to be explained should be 'law-like.' As Gopnik and Wellman (1992, p. 148) put it, theories "should invoke characteristic explanations phrased in terms of these abstract entities and laws." This means that there ought to be a law-like relation between a postulated mental state and the behaviour that it always or sometimes results in, because it is behaviour that ToM aims to explain. Like scientific theories, under TT(Scientific), the child's ToM allows "extensions to new types of evidence and false predictions" (Gopnik and Wellman 1992, p. 148). The extension to new evidence is analogous to the way that Kepler's laws of planetary gravitation predicted moons before they were observed. The reference to 'false predictions' means that an incorrect theoretical law will result in ToM errors, a topic that will loom large in this thesis.

TT(Scientific) is naturally developmental, in that theories in science and in children may be expected to change as they are confronted with new data. The development of children's ToM is naturally explained on TT(Scientific) as reflecting improvements in the specification of the abstract entities postulated in the theory or calibration of the psychological laws that ToM assumes are true. As an example of the former improvement, Gopnik and Wellman (1992, p. 150) suggest that two-year-olds "have an early theory that is incorrect in that it does not posit the existence of mental representational states, prototypically beliefs." There will be stages of development as the child matures. Later on, the child will be working with a mature adult concept of belief.

Theories must have laws or generalisations. The starting point for suggesting some folk psychological laws ought to be those that the ordinary person would recognise, since we are seeking to axiomatise or provide generalisations for folk psychology. Such a type of 'common sense belief/desire psychology'

is sketched by Fodor (1987, p. 13) in a form which allows the generation of laws. S relies, he writes, on "causal generalisations" of the form "If [O] wants that P, and [O] believes that not P unless Q, and [O] believes that it is within his power to bring it about that Q, then ceteris paribus [O] tries to bring it about that Q." Fodor's primary argument for this claim is that it explains the widespread success of ToM abilities. The picture here of how reasons for action lead to action is the 'standard' account due to Davidson (1963) in which a reason for action is a combination of a desire and a belief. The belief is that the action will satisfy the desire. This is generally how ordinary people think actions are caused; so a generalisation of Davidsonianism seems to be among the laws of folk psychology.

The power of belief/desire psychology is demonstrated by Fodor by showing how it can correctly track through various complexities and background assumptions in the case of the Shakespearean character Hermia. Hermia sees that her lover Lysander is missing while Lysander's rival Demetrius is present and grim-visaged. Hermia uses the generalisation above with Demetrius as O. P is Demetrius's desire to woo Hermia. Demetrius's belief "that a live Lysander is an impediment to the success of his (Demetrius's) wooing" (Fodor 1987, p. 2) gives Hermia Q to the effect that Demetrius has killed Lysander. The generalisation is indeed powerful here because it explains all of Hermia's mental states together with the facts that Lysander is uncharacteristically absent and Demetrius is grim-visaged.

Gopnik (1993, p. 99) confirms that one of the "structural characteristics of theories [is] the fact that they involve coherent law-like generalisations." Gopnik and Wellman (1992, pp. 150-151) propose a couple of such generalisations of ToM: "[g]iven that an agent desires an object, an agent will act to obtain it. Given that an object is within an viewer's line of sight, the viewer will see it." Another typical statement of the central thrust of TT is given by Apperly

(2008, p. 268), who writes that TT accounts: "propose that theory of mind abilities are constituted by a set of concepts (belief, desire, etc.) and governing principles about how these concepts interact (e.g., people act to satisfy their desires according to their beliefs)." Other examples are given by Baron-Cohen (1993, p. 30) who writes that four-year-olds "make clear, theory-like assertions ('[i]f you haven't seen what it is, then you won't know what it is;' or, '[i]f you want an x, and you think what you're getting is an x, then you'll feel happy,' etc.)." It is clear then that the laws or theoretical generalisations connecting mental states to behaviour are central to ToM capacities on the TT(Scientific) account.

Saxe's TT account also constructs ToM on the basis of laws or rules. On the topics of folk physics and folk psychology, Saxe (2005a, p. 174) writes: "[i]n each case, we could construct a theory (or a body of beliefs) about the entities involved, and the rules governing their interactions." Although this is somewhat tentative, it is clear from an overall consideration of Saxe (2005a), including indicatively her citation solely of Gopnik and Wellman (1992) combined with the absence of any citations of TT(Innate) proponents, that her preferred account of ToM is TT(Scientific).

2.3 TT(Innate)

I will use the term TT(Innate) for the non-scientific variant of TT. The term used by many authors promoting such an account of TT is 'ToMM,' standing for Theory of Mind Mechanism. The major proponents of TT(Innate), Scholl and Leslie (1999, p. 133), set out TT(Innate) or ToMM as holding that "the capacity to acquire ToM has a specific innate basis [...and ...] the specific innate capacity takes the form of an architectural module." It is worth emphasising at the outset that their starting position is a single module. So the modularity aspect of the proposal is one way of explaining the innate

capacity but a TT proponent could presumably be nativist without being Modularist. As Scholl and Leslie (1999, p. 134) admit, their "claim is not that the entirety of ToM is modular, but only that ToM has a specific innate basis." Nichols and Stich (2003, p. 117) set out the distinction between TT(Scientific) and TT(Innate) when they write "In contrast with scientific-theory theorists, who think that the information used in mindreading is acquired, modified, and stored in much the same way that scientific and common-sense theories are, modularity theorists maintain that crucial parts of the information that guides mindreading is stored in one or more modules." The idea is that the modularity view of TT(Innate) allows for the body of knowledge employed in ToM to be 'located' in an innate module. One claim of TT(Innate) is that "certain core concepts used in mindreading, including 'BELIEF, PRETENCE and DESIRE' " (Nichols and Stich 2003, p. 125)[11] are contained within the innate module for ToM.

Many elements of TT(Innate) will be shared with TT(Scientific), beyond the obvious one that both postulate theories and bodies of knowledge that underly ToM capacities. Both will have ToM espouse the Fodorian belief/desire psychology.[12] Both will allow that ToM includes laws or generalisations of the type proposed above. Both accounts will see S's postulate abstract theoretical entities —mental states. In sum however, our discussion of TT(Innate) can be more brief than that of TT(Scientific) since there is much common ground.

It might be asked in relation to the last point above whether the entities postulated by naive physics (e.g. momentum) are abstract or theoretical. The point of the question is that one might take the line that the folk do

[11] I will employ the standard practice of capitalising the names of concepts.

[12] One might in principle construct a new form of TT while having a view quite different from Fodor and Davidson about the folk mental states and their explanatory role. For example, one could combine TT with the view that facts rather than beliefs typically explain actions. As I have mentioned previously, I will not consider in this thesis such putative further types of TT.

not really have any theories. This would be a 'thin' or 'hard' definition of theory such that only scientists or other professionals are, strictly speaking, in possession of theories. On such a line, folk physics and folk psychology would not involve theories while academic physics and psychology would. Similarly, the folk versions of both disciplines would not, strictly speaking, be postulating abstract theoretical entities. I will not take this line. One question which is 'in the wings' throughout this thesis is 'what is a theory?' The line above will be very strict; indeed, it could be interpreted as falsifying all forms of TT because nothing the folk have counts as a theory. That is one reason not to take the line. Another is that the thin account of theories will involve one in a difficult type of line-drawing exercise of the type used to great effect by Chalmers (1997). Most people learn some physics and perhaps some philosophy at school. Accounts of physical or theoretical phenomena held by the folk which were adjacent in terms of complexity, sophistication, predictive power and accuracy would lie either side of a theory/not theory boundary. So a better response is to say that the folk possess theories to some extent. One might then be tempted to say that they also postulate abstract theoretical entities to some extent, but postulation seems less liable to admit of degrees than theory possession. If it does, so be it. If it does not, then being in partial possession of a theory means making (complete) postulations which may be partly inaccurate about a proper subset of the full conceptual contents of the total theory. All of these postulations will involve making generalisations. If ToM includes a postulate including DESIRE along the lines of 'if A DESIRES X and BELIEVES that action ϕ will obtain X for A, then ceteris paribus A will do ϕ,' it ipso facto includes a generalisation. The fact that the postulate is expressed in general terms about person A suffices to make it a generalisation; it generally applies to all people with desire X and not just to some single person A.

The differences between TT(Innate) and TT(Scientific) lie most promi-

nently in differences in the description of how the body of theoretical knowledge is obtained. TT(Innate), in contrast to TT(Scientific), holds that the body of knowledge underlying ToM is more like the knowledge that underpins the ability to speak and read a language. This is distinct to scientific knowledge for several reasons including that the development of language knowledge does not seem to proceed via the formation and confirmation of hypothesis. The idea is more along the lines that would be termed Chomskian in theoretical linguistics. The languages themselves are not innate, but the ability to learn them may be. This approach has the advantages in linguistics that it explains the fact that children are able to learn languages but also that the one they learn is the one they hear. Similarly, on the TT(Innate) view, ToM abilities develop quickly not because the abilities themselves are innate, but because the ability to acquire the abilities is innate. This account has the same advantages as the linguistic one in terms of explaining the speed with which children acquire ToM abilities and also that they do so in such a way that their ToM predictions will be similar to those of the adults around them.

Only one of the TT(Scientific) claims set out above needs to be changed to arrive at TT(Innate): TTe. This is the one that encapsulates the 'scientific analogy' nature which is characteristic of TT(Scientific). If we change TTe to read as below, we have arrived at a set of claims outlining TT(Innate).

TTeI: This development may be understood as analogous to development of bodies of linguistic knowledge

Scholl and Leslie note that the fact that development occurs in ToM capacities has been taken to favour TT(Scientific) over their preferred TT(Innate) account, because modules are taken to be static. They oppose this argument by noting that modules may 'come online' at various times. Scholl and Leslie (1999, p. 132) employ a distinction due to Segal, noting that "Segal distinguishes between synchronic modules (which reflect a static capacity), and

diachronic modules (which attain their character from the environment via parameters, as in the case of 'Universal Grammar')." The reference to Universal Grammar here means that the TT(Innate) picture is Chomskian in that innate capacities to develop capacities are postulated. The capacities that develop are not themselves innate. In Universal Grammar, the capacities that develop are the ability to use languages. In ToM, the capacities that develop are the abilities to predict and explain the behaviour of others.

The parameters are an adaptation of another Chomskian idea. The idea is that while there is a very large number of logically possible languages, only a small subset of them are actually used by people. It is logically possible but in practice extremely unhelpful to have a language in which the words for common items change on a daily basis, or in which the surface grammar were not constant. Actually used languages are more sensibly constructed, and might be so because of the way their parameters are set. For example, in the German language, verbs come at the end of sentences. This location could be set by a parameter: a child that learned German would be one that had switched its 'verb location at end of sentence' parameter to TRUE. Other values of that parameter would be possible, but no useful human language would have a parameter like 'nouns change their referent daily' set to TRUE. Mapping these ideas across to ToM, we might find that TT(Innate) postulates a similar set of parameters which define which of several innate capacities to form capacities become operative.[13] Observation by children of the behaviour of others around them would set the parameters in appropriate ways. They might set their parameter 'people who say X believe X' to TRUE. Or, to employ Segal's example, they might have a switch "labelled prelief/belief," (Segal 1996, p. 151) with the improvement in the child's ToM reflected by

[13]Scholl and Leslie (1999) in fact oppose Segal's use of parameters to set children's ToM for their ToMM conception of TT(Innate), which is another reason to prefer the label TT(Innate) to their ToMM in the context of this discussion.

the switch changing its value. This particular improvement is the change from a deficient PRELIEF concept that does not distinguish PRETENCE and BELIEF to a mature concept of BELIEF shared with adults. This would go some way towards explaining how children quickly generate ToM capacities, and how they tend to make similar ToM predictions as do adults in their culture.

2.4 ST(Replication)

The definitional ST claims are set out below; it can be seen that they largely oppose the matching TT claims.

- STa: In order to predict and explain the behaviour of O, S "does not need to entertain thoughts of the form (T), but only thoughts of the form 'I believe that P' " (Davies and Stone 1995, p. 6)
- STb: "To entertain thoughts of just that first-person form, [S] does not need to have the full-blown concept of belief. In fact, the 'I believe that P' could just as well be deleted" (Davies and Stone 1995, p. 6)
- STc: S does not need the concept BELIEF to have beliefs
- STd: Development of folk psychological ability is a case of "the child gradually becoming more adept at imaginatively identifying with other people" (Davies and Stone 1995, p. 6)
- STe: This development is a gain in skill not knowledge
- STf: Information processing mechanisms are needed "to engage in these imaginative tasks" (Davies and Stone 1995, p. 6)

The central thrust of the ST approach can thus be seen to be anti-theoretical, as would be expected. S does not need the concept BELIEF, but

just to be able to believe things. S need not be able to think representational thoughts like 'O believes P' but merely note his beliefs.

STe provides a distinction between ST and TT. Recall that TTe insisted that there is a body of knowledge and increases in the scope and quality of that knowledge is what explains children's improvement in ToM abilities. ST denies that there is a body of knowledge and explains the improvement by appealing to improved skill at 'imaginatively identifying' with other people. Provided that we can maintain a clean distinction between skill and knowledge, it can be seen that STe denies both TTe and TTeI, so that on this view, ST is distinct from TT.

Each ST claim is not the exact negation of the corresponding TT claim, though the ST claims are in each case generally opposed to the corresponding TT one. The way Davies and Stone couch STa and STb is initially perhaps a little confusing. They seem to start in STa by insisting that S needs the concept of belief, because while we can accept that S can have a belief without having the concept BELIEF, that does not entail that S has the meta-ability to form the belief 'I believe that P' even when S does in fact believe that P. To see this distinction, observe that we may be prepared to allow that non-human animals believe that P —although this is controversial —but the ascription of 'I believe that P' to non-human animals is absurd. However, in STb we learn that the concept of belief is not needed by S and in a slightly throwaway fashion, Davies and Stone concede that the 'I believe that P' can be dispensed with. I submit that the two approaches are significantly different and the version without 'I believe that P' is both more plausible and expresses the main idea promulgated by ST proponents viz. mind-reading can be performed by those who can form beliefs and no mental state concepts including BELIEF are required.

The major proponent of ST(Replication) is Heal. Replication is set out as follows. Heal (2003, pp. 13—14) asserts that if S wishes to predict the action

of O, then S will "endeavour to [...] replicate or re-create [O's] thinking. [S will] place [himself] in what [S] takes to be [O's] initial state by imagining the world as it would appear from [O's] point of view and [S will] then deliberate, reason and reflect to see what decision emerges." This gives us success conditions for replication. If the function of a thought in S's simulated and contained Replication is the same as the function played in O's un-simulated and unconfined cognition, then pro tanto, the simulation has been successful.

We can obtain more insight into ST(Replication) by considering Heal's responses to three objections that have been raised to it. To my mind, she is successful in all three cases in defusing the objections, and in the third case, she raises an important issue which I will be discussing further.

The first objection aims to disarm the argument in favour of ST that claims it is less complex and demanding than TT. The objection does this by considering the need on ST for S's to perform 'initial state matching.' This means that for replication to be successful, S must be able to do two things: "know what psychological state [O] is in" from external observation and "put [himself] in the same state" (Heal 2003, p. 14). If this is difficult, we will struggle to understand how replication could often be successful. Heal's response is to claim that the objection misdescribes the target of replication. S is not examining O but rather the situation around O as seen from O's perspective. As Heal (2003, p. 15) notes, "[i[t is what the world makes [S] think which is the basis for the beliefs [S] attributes to [O]." The objector cannot here continue to urge that it is difficult for S to contemplate the world around O, because S contemplates the world all the time. Moreover, any common errors that S makes in contemplating the world around O will presumably also be made by O, and thus not impede simulation.

The second objection, ascribed by Heal (2003, p. 15) to Dennett, is that ST(Replication) lacks parsimony. The objection urges that replication must

involve special beliefs about beliefs and those require more complex mental machinery than merely having beliefs. This objection is met by noting that nothing is required here for simulation purposes that is not already required in S's own case, to enable S to chart actions now, allowing for the fact that S's own beliefs and desires may change in the future. Heal (2003, p. 16) also argues that "[m]ake-believe belief is imagining," and that people already have the capacity to imagine. Heal (2003, p. 16) agrees that it would be absurd to claim that replication can only be successful if S "believes what [O] believes," so there must be some way of preventing O's beliefs becoming S's beliefs in a way that causes S to act on the beliefs as opposed to ascribing them to O. This leads a further distinction known as the on-line vs off-line ST distinction which I will discuss below. If S ascribes a belief to O, that belief must be off-line for S and not cause behaviour of S in the normal way.

The 'make-believe beliefs' or off-line beliefs approach could also work in another way. We could adopt the view on which beliefs are seen as items in the 'belief box.' Firstly, there could be multiple subscripts in the belief box, to speak metaphorically, which tag various beliefs as those of S or those of O.[14] As long as the beliefs tagged 'O' are kept off-line, we would have a mechanism that performed as needed. Secondly, there could be a contained simulation environment in which beliefs are as effective as they are outside that environment in fecundating other consonant beliefs, but the outputs of which are translated into contained or simulated desires rather than actual desires. The outputs do not leave the contained simulation environment in the form of action proposals. In either case, the simulated beliefs of O stay off-line and do not directly issue in desires or actions of S, as required. The simulated beliefs of O do have effects in S: they are ascribed to O and used to predict his behaviour.

[14]Such a subscript approach is proposed by Pratt (1993, p. 72).

The third objection holds that ST(Replication) requires theoretical elements and thus collapses back into TT. The objection holds that under ST(Replication) there is in S a sequence of thoughts and mental state transitions that replicate those of O in order to explain O's behaviour. The objection suggests that S can only do this by using some theory of how mental state transitions in O are likely to follow from S's view of what O believes. This would mean that replication was less analogous to becoming like O and more like S applying a theory of O to supply the links in a chain of simulation of O. Such an account would have allowed a theoretical element to corrupt the pure simulationist ST(Replication) account, if the objection is successful. Heal's response involves questioning the nature of the links in a chain of simulation. One does not use a theory to get from 'I see p' to 'p' —it is merely a rational transition. This objection is a version of the 'setting the bar too low' error which I will outline below.

2.5 ST(Transformation)

The major proponent of ST(Transformation) is Gordon. Intuitively, the distinction between ST(Transformation) and ST(Replication) is that on the former transformation variant of ST, S simulates O by becoming O, while on the latter replication view of simulation, S simulates O by becoming like O. The first idea is clearly only metaphorical, since no-one can in reality become someone else. Gordon (1995a, p. 53) sets out three claims, all of which are asserted by ST(Replication) and denied by ST(Transformation). The three claims are that simulation involves:

1. "an analogical inference from oneself to others
2. premised on introspectively based ascriptions of mental states to oneself,
3. requiring prior possession of the mental states ascribed."

ST(Transformation) then is anti-Inferentialist, anti-Introspectionist and to coin a phrase, anti-Possessionist. ST(Replication) is Inferentialist, Introspectionist and Possessionist. By the term Possessionist, I mean views claiming that S must possess a mental state before being able to ascribe it to O. Gordon (1992, p. 32) confirms that on his account, S "is not using one individual, himself, as a model of [O], and there is no implicit inference of any sort from [S] to [O]." Instead, the idea is much more to make action predictions without an intervening ascription of mental states. This is done by 'becoming' the other person, or putting oneself in their position, and seeing what one will do. That of course is again somewhat metaphoric, since one rarely finds out what one is going to do by external observation: one merely acts. As Gordon (1992, pp. 31—32) puts it, "[w]hat is relevant is to ask, within a pretend context in which I simulate being [O], [is] 'What shall I do now?' [...] Thus, within the context of the simulation, the realisation that now is the time to ϕ spurs me to action." Here I employ ϕ to stand in for the action proposed in the simulation context. Under normal non-simulation circumstances, the realisation that 'now is the time to ϕ' will cause S to ϕ. Within the simulation context however, the realisation that 'now is the time to ϕ' will cause S to predict that O will ϕ. Gordon's view does not involve any ascription of mental states to S or to O. As Gordon (1995a, p. 53) writes, "people often predict what another will do in a given situation by imagining being in such a situation and then deciding what to do." The idea that the ability to pretend is related to the ability to predict behaviour is supported by many studies of autism. As one example, Baron-Cohen (2001, p. 7) notes the well-known finding that autistic subjects have impaired ToM and also that "[m]any studies have reported a lower frequency of pretend play in the spontaneous play of children with autism."

On the ST(Transformation) view set out by Gordon (1995a, p. 57), per-

sons "transform [themselves] in imagination" whereby they "modify [their] regular stock of mental states with a complement of artificially induced pretend states, keeping the resulting adjusted stock of mental states off-line." On the ST(Transformation) account, there is no need for an inference from S to O by O because S has become O —or successfully placed himself in O's position —and thus now can decide directly what O is likely to do because it can be read off from what S would do. S now knows what O would do because S is in the same situation, except 'off-line.'

As noted, Gordon's ST(Transformation) view is anti-Introspectionist, while on the alternative ST(Replication) account, the way S predicts the mental states of O is by introspecting S's own mental states within the pretend off-line environment within which the simulation of O takes place. S then would have direct introspective access to the mental states of O, or rather the mental states that S ascribes to O as a result of the simulation. So one potential advantage of ST(Transformation) is that it sidesteps questions about mental states and whether they can be introspected. People act; they do not form a mental state which has action as a consequence. ST can still be true even if Introspectionism is false.[15]

It is not an objection here to allow that Gordon (1995a) does not need Introspectionism when predicting action, but does need it when ascribing mental states, because such 'ascription' in effect comes for free. The output of simulation on the account of Gordon (1995a) is an action ϕ, and the mental states are as it were set on one side. They are whatever they need to be to produce the action ϕ. Observers might be led to make this mistake by thinking that Gordon (1995a) structures his argument about facts rather than mental states explaining behaviour precisely because it is difficult to understand men-

[15] In any case, phenomenology suggests and it has been argued, (Rey 2013) that Introspectionism is true. The truth of ST is consistent with either the truth or falsity of Introspectionism.

tal state ascription on his account; rather he is concerned to avoid ST being committed to Introspectionism which he sees as controversial.

2.6 Further Possible Types Of ST

As we have seen, ST(Replication) and ST(Transformation) differ from one another in that they assert or deny all of three claims. Yet other positions are possible, and may also be defined in terms of assertion or denial of those three claims. In other words, since the three claims may be asserted or denied independently from one another, there are other possible positions within the logical space available to ST proponents. Some of these positions may even be interesting. I show in Table 2.1 the possibilities in terms of the three claims (Inferentialist, Introspectionist, Possessionist) listed above. Since we have two options (assert or deny) across three options, there are eight possible positions in this logical space as so far analysed.

One such independent position is outlined by Goldman who examines Gordon's motivations for denying each of the three claims. Goldman (2006, p. 186) notes that Gordon thinks that "[t]he analogical inference element [. . .] threatens to make ST collapse into TT." This is the collapse risk already noted, which will be examined further below. Goldman also observes that Gordon wishes to avoid Introspectionism because that doctrine is "philosophically controversial" (Goldman 2006, p. 186). However, it is unclear, as Goldman (2006, p. 186) points out, why Gordon denies Possessionism, observing that Gordon's "rationale for denying the concept-possession element is elusive." So we can see that Goldman's account, ST(5), is a position of interest which is viable, distinct from ST(Replication) and ST(Transformation) and ably defended at length by Goldman.[16]

[16]The charge of collapse risk is brought against Goldman's position. For a persuasive riposte, see Goldman (2009), which also responds convincingly to the anti-Introspectionist critique of Carruthers (2009).

Position	Inferentialist	Introspectionist	Possessionist
ST(Replication)	✓	✓	✓
ST(Transformation)	✗	✗	✗
ST(3)	✓	✗	✗
ST(4)	✗	✓	✗
ST(5)	✗	✗	✓
ST(6)	✓	✓	✗
ST(7)	✓	✗	✓
ST(8)	✗	✓	✓

Table 2.1: Possible Variants Of ST

There are also at least two possible types of Possessionism. Recall that the original definition with which we were working was that under ST(Transformation), ToM use required "prior possession of the mental states ascribed" (Gordon 1995a, p. 53). Contrast this with the following claim that using ToM under ST(Transformation) means "requiring prior possession of the concepts of the mental states ascribed" (Goldman 2006, p. 186). There seems to be scope then to distinguish between two positions here, one requiring possession of mental states and one requiring possession of the concepts of mental states. This distinction is picked up by Gordon (2009, §2) who notes that "[f]or Goldman, but not Gordon, it is essential that the simulating system recognise its own mental states. This recognition generally requires, according to Goldman, that [S] possess the relevant mental state concept [while] Gordon takes the position [...] that simulation does not require the application of mental concepts."

Moreover, one might insist that S needs to possess the mental states or the mental state concepts in order to ascribe them to O. One might in addition insist that S needs to possess the mental states or the mental state concepts in

Position	Mental State Required	Mental State Concept Required	For Recognition	For Ascription
Possessionism_1	✓	✗	✓	✗
Possessionism_2	✓	✓	✓	✗
Possessionism_3	✓	✗	✓	✓
Possessionism_4	✓	✓	✓	✓

Table 2.2: Possible Types Of Possessionism

order first to recognise them in S before ascribing them to O. We can see that at an early stage, Gordon (1995a, p. 53) denies the latter duplex view when he opposes claims that "to recognise and ascribe one's own mental states and to mentally transfer these states over to [O], [S] would need to be equipped with the concepts of the various mental states." It is worth noting that Gordon discusses Introspectionism and Inferentialism in his canonical statement of his position —"Simulation Without Introspection Or Inference From Me To You" (Gordon 1995a) —but says very little about Possessionism, consistent with Goldman's claim that Gordon's reasons for denying Possessionism remain opaque. Gordon (2009) provides a fuller discussion of this precise point, probably in response.

I show in Table 2.2 the possible options for Possessionism as described above. I assume that one can possess a mental state without possessing the concept of that mental state but not vice versa. I also assume that one can be required to possess a mental state or concept for recognition but not for ascription, but that one cannot be required to possess it for ascription but not for recognition. Relaxation of these assumptions would create further options.

If we agree that these four types of Possessionism are distinct and that all can be held independently of the eight positions specified in table 2.1, then

we now have 32 possible variants of ST. As said at the outset, I will remain neutral in this thesis between these many possible variants of ST. While they are interesting, which one of them exactly is the best variant and whether it is certain that they do not collapse into each other is not critical to my project. By contrast, collapse of ST into TT would be serious for my project since there would then not be a separate theory from TT to defend. If it transpires that my claims elsewhere commit me to one or other version of ST so be it; all my project requires is that there be at least one viable variant of ST which is distinct from TT.

2.6.1 On-line Vs Off-line

Stich and Nichols name the ST account the 'off-line ST,' but they also note that for Gordon, the "off-line picture" is "only an 'ancillary hypothesis' [...] albeit a very plausible one" (Stich and Nichols 1995a, p. 91). Heal tells us the off-line hypothesis is widely-held by supporters of ST, but not by herself. The 'off-line' claim holds that when S simulates O's decision making, S does something similar but not identical to what S does when S makes a similar decision on his own behalf. S knows for example, that if S desires coffee, and S is outside a coffee shop with available sufficient resources of time and finance, S may well go in. S can extrapolate from this. If S sees O behave in a particular way —viz. entering the coffee shop —S may by analogy with himself as a model use simulation to decide that O has entered the coffee shop because O desired coffee.

Note that we need to be careful with the use of the verb 'to know' here. If knowing that people who want coffee and are outside a coffee shop may go in and buy coffee constitutes a law or generalisation or element of a body of knowledge, then ST may have collapsed back into TT. So we need to think of talk of 'knowing' here as meaning 'able to simulate beliefs and desires such

that we can find a combination of the key ones which result in the prediction 'enter the coffee shop.' This does not appear to be much more difficult than the following sequence: 'O wants coffee, O is outside the coffee shop, S(off-line) wants coffee, S(off-line) is outside the coffee shop, S(off-line) action: enter coffee shop, predict: O enters coffee shop.

When S runs the simulation, S does not want to be prompted actually to enter the coffee shop himself. This is what is meant by the simulation or the beliefs simulated being off-line. The behaviour of entering the shop should be prompted only by S's own desire for coffee, unless we begin constructing some more complex story in which perhaps S enters because O has and S wants to talk to O. Thus, S's decision-making system is off-line in that S arranges for it to output a specified behaviour, but not actually to execute that behaviour, because S is interested in simulation and not performing that action at this point. The on-line version of ST then sees S as having his practical reasoning system working on actual occurrent beliefs of S. So the postulation is that the system runs as normal but there is no translation into action of the output.

We would also need a mechanism whereby the beliefs S has for standard reasons —e.g. perceptual input —are not contaminated with beliefs S has merely because S needs to have them because O does and S wants to simulate O. This is specially relevant if S needs to predict O's behaviour based on a belief of O's that S knows is a false belief —we must avoid any account where S is required knowingly to have a false belief because that seems impossible. Where on-line simulation does look plausible would be in the grammar-type question discussed by Harris (1992, p. 124). The task is for S to decide which of a set of sentences O will think are grammatical in the common language of S and O. It is very likely that S decides what O will say by forming true beliefs about the grammaticality of the sentences and ascribing them to O as well. Similarly, to adapt Gordon's example, if S and O both form the belief

that a tiger is confronting them, S will have no difficulty deciding to run and predicting that O will do so also. Harris (1992, p. 124) concludes his argument by noting plausibly that the TT proposal to explain this, on which the subject has a first order representation of grammar for his own use and a second order meta-representation of other people's grammar, "strains both credulity and parsimony." It is simpler to postulate that S forms on-line beliefs about the grammaticality of the sentences —i.e. S actually has those beliefs —and then ascribes those beliefs to O, together with the corresponding behaviour.

The point where it is critical for the system to be off-line is immediately before behaviour. S must be able to infer what desire it was that caused O to enter the coffee shop —or alternatively predict that O will enter the coffee shop if S knows that O has the desire for coffee —without either circumstance causing S to enter the coffee shop or want to do so, unless we again add extraneous factors such as that now S sees O enjoying all the coffee, S wants one as well. But that leaves open all of the stages where S could inhibit behaviour by applying an off-line status. S could inhibit behaviour by any of the following. Firstly, S might have a pretend belief that does not issue in behaviour because it is not S's actual belief.[17] Secondly, S might in some subsystem allow for a real belief that S desires coffee to operate, but prevent it from having the usual effects of a belief. S would need to restrain it from exiting the subsystem to the extent that it causes other propositions such as 'S desires fluid' or 'S should enter the coffee shop' to become assertible. Thirdly, S might have a real belief, which has real effects in the inference mechanism or a special contained subset of it, but intervene at the last moment before behaviour with an inhibition of action.

[17]One might see this as not inhibition because one might hold that strictly speaking, a pretend belief does not require inhibition to avoid being acted upon. It essentially 'comes with' its own inhibition since it is a pretend belief and not a belief. On this line, one would need to take account of this when discussing ToM mechanisms.

Stich and Nichols (1992, p. 247) suggest that off-line ST is unparsimonious. They assert that under off-line ST, we would need to postulate a control mechanism to "take that system 'off-line,' feed it 'pretend' inputs and interpret its outputs as predictions." They claim that this task would be "very non-trivial" (Stich and Nichols 1992, p. 247) but do not provide any evidence for this claim. We also have an everyday example of the decision-making system being taken at least partially off-line: sleep. As O'Shaughnessy (1991, p. 160) observes: "[t]ypically in sleep either simple bodily act-inclinations or sensation-caused act-inclinations [immediately] generate basic bodily willings; and these primitive transactions make no demand upon belief or concept system." What this means is that when dreaming, we seem to perceive the external world; these apparent percepts sometimes cause us to wish to move in response, but such desires to move do not result in the usual changes in beliefs. If I dream that I walk into the sea, I nevertheless do not believe that I am wet, at least not in the same way as I do if I actually walk into the sea. Thus we have an example of some control mechanism existing or taking effect in that the decision making system caused me to 'want to walk into the sea;' this then uses 'I am in the sea' as a 'pretend input' and generates the output 'I am wet' as a prediction for my dreaming self in the dream case. There seems to be no special difficulty about doing the same when awake in respect of another person.

In the same vein, as Heal (1998, p. 89) notes, "we can reason with representations which we do not believe" because otherwise we could not, for example, "explain what we are doing in arguing by reductio ad absurdum or reasoning hypothetically." Reductio ad absurdum means to reason by assuming a proposition that is probably false for the sake of argument, and finding that it has absurd consequences. That is taken as proof that the proposition initially assumed is indeed false. On such occasions, it is indeed the case that persons

have reasoned using propositions that they do not believe and so there must be some mechanism whereby they can hold such propositions off-line. Thus, nothing extra need be postulated for simulation and ST continues to be a parsimonious and elegant explanation of ToM.

My position on the on-line vs off-line debate will be similar to the one I took on the other different types of ST. What is required for my project in this thesis to succeed is that one of the options be correct; I do not need at this stage to select a champion. Both options look viable; it is fair to suggest that the off-line version has received more support in the literature.

2.7 Avoiding Collapse Between ST And TT

As mentioned, collapse between simulationist and theoretical accounts would be a problem for my project, because I cannot defend ST against TT if they are not separate. TT proponents have sometimes tried to collapse ST back into TT. In this section, I will argue in three ways that ST and TT are separate. These three ways are as follows. I will first note some plausible distinctions that have been drawn in the literature between ST and TT. Then, I will argue that TT proponents have often been guilty of 'setting the bar too low' in their claims that ST collapses back into TT. They have done this, I suggest, by regarding the employment in any way of any theoretical items like beliefs as sufficient to render an account of ToM theoretical. Since the main challenge to ST nowadays is the denial that it can handle systematic ToM error rather than that it is not distinct from TT, I will treat the collapse challenge only briefly here. Responding to the systematic error challenge is my main project in this thesis. We will nevertheless learn more about our two competing theories by seeing how obvious or not it is that they are distinct.[18]

[18] Davies and Stone (2001) provide an extended discussion of collapse risk between ST and TT.

2.7.1 Distinctions Between ST And TT

I will here briefly outline some plausible distinctions that have been observed between ST and TT. I will suggest in turn that ST and TT employ different data sources; that the Rylean distinction between knowing-how and knowing that will also divide them; that they predict different answers as to whether folk psychology and scientific psychology are continuous and that they differ in the way they handle some real-life examples. In the latter case, I will also be responding to an objection to ST raised by TT proponents.

Different Data And Processes

One broad-brush and intuitive but clear way of distinguishing ST and TT is given by Arkway when she discusses Heal's position. Arkway (2000, p. 128) writes that "[S] looks not at the [O] to be understood but at the world around [O]." This provides a clear contrast with the process under TT accounts in which S will look at O as well as at O's environment, and where we might expect the bulk of the theory to be about people, albeit perhaps people in situations.

Arkway also sees the importance of the off-line nature of some ST accounts, whereby beliefs held by S for simulation purposes of O must be held in quarantine and not result in decisions or behaviour of S. Arkway (2000, p. 128) notes the view of Stich and Nichols that all accounts which posit the off-line use of the decision-making system count as versions of ST. While ST accounts may be on-line as well, there do not seem to be any off-line TT accounts for the simple reason that none are needed. Theoretical reasoning processes normally culminate in some new theoretical beliefs which may become candidates for motivating behaviour, but will not automatically do so.

Knowing How Vs Knowing That

Freeman (1995, p. 68) writes: "[TT] is intellectualist in that emphasis is put on the child's own [ToM], a theory through which children are held to filter psychological evidence. [ST] is grounded in a consideration of pre-theoretical practical intelligence plus a competence at imagining." There is a clear division here between the activity postulated in ToM use. Children are either filtering evidence through a theory, or using their imagination. These activities are very different.

Also, since the divide here is between 'pre-theoretical' and 'theoretical;' we may see it as closely analogous to the distinction due to Ryle (2009, p. 68 et seq.) between knowing-how and knowing-that. TT would involve theoretical knowledge, which would be knowing-that —the possession of propositional knowledge, expressed in the readiness to affirm propositions. The pre-theoretical knowing-how of ST would not involve any propositional knowledge: children can therefore be deemed to be able to perform simulation successfully without thereby necessarily being able to affirm any propositions.[19]

Continuity Of Folk And Scientific Psychology

It has been claimed that folk psychology will be superseded by scientific psychology. This would involve ordinary people learning more psychology over perhaps many decades and eventually abandoning their current inaccurate theories of how people think.

The prevalence of error seen in ToM capacities has implications for the accuracy of folk psychology and scientific psychology and whether there is some form of continuity between the two. Gopnik seeks to distinguish TT

[19] Note though an extended argument (Stanley 2011) to the effect that knowing how to ϕ is the same as knowing the fact that 'w is a way to ϕ.' This would collapse Ryle's distinction, and mean that simulating children know some propositions like 'w is a way to simulate ϕ.'

views from ST views based on these accuracy and continuity points. Gopnik (1996, p. 180) writes that both modular and innate TT views suggest "folk psychology could be, indeed is likely to be, wrong in important ways." Also, the "theory-formation view [...] proposes a deep continuity between folk psychology and scientific psychology" (Gopnik 1996, p. 180) since the latter is a formalisation of the former. Since, as Gopnik (1996, p. 180) states, her views "stand in contrast to other accounts, such as simulation theory," we can see that she claims there are two additional distinctions between TT and ST as set out below.

- TTg: folk psychology is wrong in important ways
- STg: not TTg: it is not the case that folk psychology is wrong in important ways
- TTh: there is a continuity between folk psychology and scientific psychology
- STh: not TTh: it is not the case that there is a continuity between folk psychology and scientific psychology

Thus, ST and TT make different predictions as to whether folk psychology will eventually be superseded by scientific psychology or not. This already distinguishes them, and will do so more clearly to the extent that data or argument makes it look more or less plausible that scientific psychology is superseding folk psychology.

2.7.2 Theory Driven Vs Process Driven

This useful distinction arises from an objection to ST which claims that simulation cannot be done without the surreptitious importation of some theoretical elements. If this is true, then ST requires TT and the distinction collapses.

This particular objection is ascribed to Dennett by Davies and Stone (1995, p. 18), who also ascribe the response to Goldman. Dennett asks how simulation can work "without being a kind of theorising in the end" since there is no difference in principle between when S "makes believe [S] has [O's] beliefs" and when "[S] makes believes [S] is a suspension bridge." There can be no simulation by humans of what it is like to be a suspension bridge; the only knowledge available of how suspension bridges behave is theoretical. This is true, but S may not need to have O's beliefs in the same way that O does.

Responding to the objection, Goldman appeals to standard belief/desire folk psychology, where if S simulates O as having a desire for coffee and a belief that there is coffee in the cup, S will predict that O will drink from the cup. Employing such a form of belief/desire psychology does not commit a position to either ST or TT. As Strijbos and De Bruin (2013, p. 760) show with copious references, the "assumption that folk psychology is rooted in belief-desire psychology is taken for granted by almost all participants in the debate" whether those participants favour TT, ST or hybrid views. Goldman outlines a 'process-driven' simulation where S "simulates a sequence of mental states" of O so that S "wind[s] up in the same (or isomorphic) final states" as long as S and O had i) the same initial states and ii) "both sequences are driven by the same cognitive process" (Davies and Stone 1995, p. 18). 'Isomorphism' means that S must pass through similar states as O does in reaching his conclusions if S is to simulate O successfully. S need not make believe S has O's beliefs in order to go through an isomorphic process.

2.8 Setting The Bar Too Low

There are many examples in the literature of TT proponents making their task too easy by making almost anything count as use of theory. I will set out some examples below, and then consider some useful remarks by P. Mitchell,

Currie, and Ziegler (2009b). This is an important topic for the argument of this thesis since failure to appreciate it fully allows TT proponents to claim that more ToM activity is theoretical than is entailed by observations. The philosophical underpinning of this point is the distinction between conforming to a rule and following it (Wittgenstein 2001, §201 et seq.). TT proponents may show regularities in ToM but this does not show theory-use in ToM. The general idea is that the bar for the truth of TT is set too low if it suffices for TT to be the true account of ToM that ToM propositions can be expressed as generalisations. It might just be that the simulations always produce the same outputs when they have the same inputs.[20]

Daniel (1993, p. 39) writes that "simply [...] resorting to a simulation presupposes that some theory or other is in place," by which he means that people cannot run a simulation without a theory to explain why the simulation has performed as it has. I deny the force of this objection on the grounds that everything is a theory if it counts as being a theory to say 'my simulation will work' or 'this theory tells me how my simulation works.'

Riggs and D. M. Peterson (2000) set the bar too low twice when they discuss the False Belief Task. Since this is the first of several mentions in this thesis of the important False Belief Task, I will outline it here. The canonical example of the False Belief Task is given by Wimmer and Perner (1983, p. 106) thus: "[a] story character, Maxi, puts chocolate into a cupboard x. In his absence his mother displaces the chocolate from x into cupboard y. Subjects have to indicate the box where Maxi will look for the chocolate when he returns." The point is that subjects must avoid being 'seduced by

[20]There is a reluctance to assent to the assimilation of potentially imprecise simulation to precise generalisations. The reluctance is paralleled by the reluctance Soteriou (2013, p. 174) discusses to accepting the above-mentioned Stanley (2011) assimilation of possession of rather precise propositional knowledge to the less precise notion of knowing how to do something.

reality': younger children tend to be impressed by their own knowledge of where the chocolate is actually located —in cupboard y —and thereby fail to take account of the fact that Maxi was not present when the chocolate was moved and therefore fail to predict that Maxi will have a false belief that the chocolate is still in cupboard x. These errors are known as 'realist errors' in ToM. Roughly, it was found that most normal children under four years old would fail the False Belief Task while most would pass at five.

Riggs and D. M. Peterson (2000, p. 91) claim that identifying what Maxi is ignorant of " requires the theoretical understanding that people can be ignorant of things that we are aware of, and also, that if a person is absent when a change takes place, that person may be unaware of that change." However, such understanding need not be theoretical. Consider the following simulation alternative. Imagine that you are in a room without windows. It was sunny when you entered the room, but now it is raining. Do you know that it is raining? You can answer this question in the negative very easily by simulating your position in the room not knowing anything about the change in the weather. In addition, you can know by the same simulation that you will continue to remain ignorant about the rain while in the room. This provides you with both of the points listed above.

Similarly, Perner makes the 'setting the bar too low' error when he argues that the Maxi results favour TT over ST. Success on the False Belief Task requires omitting the information that Maxi's mother moves the chocolate while Maxi is out playing when assessing what Maxi knows. Perner (2000, p. 396) states that this is because "it is far from clear that the critical [information about the movement of the chocolate] is omitted just because one is imagining to be Maxi;" rather what is needed is "some theoretical knowledge that events that are not perceived are to be omitted." It is clear though. Simulate being Maxi playing in the field. Can you see the kitchen? Can you see the

chocolate being moved remotely? No, and so you know that Maxi cannot see the chocolate being moved. A rule about knowledge of remote events can be derived from the above simulation. Every time a decision about whether someone knows about a remote event is required, the simulation can be run and the result will always be that people do not know about remote events, at least if perception is their mode of access to the remote event in question.

Davies and Stone (2001, p. 146) claim that there is a "minimal theoretical background for mental simulation" which is the adoption of an assumption that O is like S. They give several examples of such an assumption, but they all suggest that S must assume that O is relevantly similar to S, or O's processes are relevantly similar to those of S, if S is to simulate O. This is false. While it must *in fact* be the case that the claims made in these assumptions are true for simulation to be successful, S need make no assumptions at all. S merely needs to simulate. There is no such minimal theoretical background.

Wilkinson, Ball, and Cooper (2010) set the bar too low in an experiment purporting to examine whether simulation or theory use was involved in particular cases of ToM use. They have observers making the decision, which illustrates one problem: if we do not already know the answer as to whether theory or simulation was involved, asking someone else will not help. In any case, the authors give an example of what counted for them as theory use. Their example is of an occasion when S comments as follows: "I think Mike's gonna feel the more regret in the short term coz he's actually chan- he actually made a bad decision whereas Timmy" chose not to make a decision (Wilkinson, Ball, and Cooper 2010, p. 1011). The claim is that S has applied the rule that O's who make a decision will feel more regret about a consequent negative outcome than O's who do not make a decision. This is setting the bar too low because such an outcome can be obtained from simulation.

Similarly, Jackson (1999) claims that any method which allows us to make

predictions about an object or person K is in virtue of that fact a theory of K. Jackson employs the question from the literature about two travellers who are late arriving at the airport. Which is more annoyed, the one who misses his flight by an hour or the one who misses it by five minutes? Jackson (1999, p. 88) believes people need to apply a theoretical generalisation to the effect that a mental exercise can "reveal how you would feel in some given situation." In fact, S just needs to perform the mental exercise and have it *be* right without S needing to know that it *is* right.[21]

Consideration of the useful commentary of P. Mitchell, Currie, and Ziegler (2009b) also throws light on the issue of setting the bar too low. They are Strong S/T Hybrid theorists with a simulationist bent; P. Mitchell, Currie, and Ziegler (2009b, p. 513) propose that "although simulation is primary, rule-based approaches develop as a shortcut."[22] Within their Strong S/T Hybrid approach, the authors present a candidate rule to be used in situations where they believe theory is more likely to be used than simulation.[23]

The allegedly rule-based scenario involves chocolate in the displaced item test. Subjects are asked the standard False Belief Task question as to where Maxi will look for his chocolate when it has been moved in his absence. P.

[21]Further, Garson (2003, p. 511) describes a further example. TT proponents set the bar too low when they claim that any use of general knowledge about people constitutes theory use.

[22]The primary argument of P. Mitchell, Currie, and Ziegler (2009b) that favours ST over TT is also valuable. They note that children gradually develop ToM competence. ST predicts this as perspective-taking capacities develop while TT predicts sharp transitions in competence as better rules are acquired.

[23]For a useful commentary on P. Mitchell, Currie, and Ziegler (2009b) and also the criticism that it is empirically unclear whether children develop ToM gradually or with sharp transitions, see Apperly (2009). See also Harris (2009) for further valuable commentary. For an enlightening response from the original authors to these two commentaries, which includes pressing the important claim that TT proponents ought to specify their generalisations, see P. Mitchell, Currie, and Ziegler (2009a).

Mitchell, Currie, and Ziegler (2009b, p. 513) suggest that this is done by those who provide the correct answer by use of a rule to the effect that O will believe that items are where they were when O last saw them. This may be true, but it is of course equally possible to obtain the correct answer by simulation. S can run through an imaginary scenario in which S's chocolate is moved from location A to location B in his absence and predict that S will not have any reason to update his belief set in relation to the chocolate: S predicts that S or O in this scenario continues to believe the chocolate is where it was before he left the scene. The suggestion of P. Mitchell, Currie, and Ziegler (2009b, p. 513) is that rule use in this scenario might be the "best method for mentalising" because it is "quick, relatively effortless and tolerably accurate." Indeed, but this assumes that ToM invariably proceeds on the most efficient basis, which would make it an unusual element of human cognition.

More importantly, running a simulation in these circumstances would always produce the same result: viz., O does not know to where the chocolate was moved in his absence because S in the simulation also does not know for the same reason. This would mimic rule-based ToM. Perhaps S remembers in some sense the output of previous relevant simulations to answer this question. There seem to be two potential lines to take here. The first one would allow that ToM sometimes relies on generalisations, but would note that these are derived from simulations so are derivative. I would not take this line, because I do not think it gives an appropriate definition to the term 'generalisation.' When we say that ToM use has involved using a generalisation, we should mean just that, and not that it has appeared to do so. So I would prefer the second line to take, which holds that ToM does not rely on generalisations, but it can appear to do so ('mimic') because simulations reliably produce the same results. My line here assumes that memory use does not constitute theory use, as seems plausible; nor does memory use constitute use of a generalisation.

Chapter 3

Objections To Pure TT Accounts

3.1 Introduction

Broadly, there are three sorts of accounts of ToM which can be constructed from TT and ST. I will not consider other putative alternatives such as the intentional stance (Dennett 1981) or the intersubjectivity account (Gallagher and Hutto 2008). These accounts are not mainstream and it is unclear to what extent they involve theory and simulation. The target of the objections to TT presented in this chapter is a particular class of TT theories on which the theories of the mental consist in more-or-less precise generalisations yielding insight into how processes such as abduction work. Both of the two major types of TT so far proposed in the literature fall into this category. As I mentioned in the Introduction, I will not be considering in this thesis any putative further types of TT which do not involve such generalisations.

Setting this aside then, the three sorts of accounts of ToM are: pure ST, pure TT and S/T hybrid accounts involving both simulation and theory. I will further divide the S/T hybrid accounts into Strong S/T Hybridism and

Weak S/T Hybridism – to be defined below. Since as we saw earlier there are also multiple accounts of TT and of ST, we can also have hybrids within each of those accounts. For example, there could be a hybrid of TT(S) and TT(I). I will term these accounts 'Theoretical Hybrids.' There could also be Simulational Hybrids involving ST(R) and ST(T) or combinations of any of the possible ST accounts listed in table 2.1. We will not need to discuss Simulational Hybrids further here since, also as mentioned earlier, I do not select a preferred variant of ST in this thesis.

I will in this chapter and the next consider all of these possibilities and conclude that all of them except Weak S/T Hybridism succumb to severe objections. I will show this by outlining objections to each type of account in turn.

I will cover three objections to TT(Scientific) in §3.2. The first objection, which I think is the most important one, is that an excess of complexity leads to a lack of parsimony. The account requires additional machinery which represents a substantial theoretical cost, and it is implausible that children could develop such complex machinery at young ages. The second objection to TT(Scientific) suggests that under it, S's must solve the Frame Problem and that finding such a solution is impossible. The third objection asks how it can be that all children converge on the same ToM even though their evidential bases are different.[1]

I will cover three objections to TT(Innate) in §3.3. The first objection claims that ToM development cannot be accounted for by TT(Innate). This arises because the TT(Innate) account, as so far proposed, has been essentially modular (and it does not appear as though the rather ad hoc non-modular add-on Scholl and Leslie (1999) employ can be coherent or solve the developmental objection). It is hard to see how innate, informationally encapsulated

[1]For further objections see Stich and Nichols (1998), Stich and Nichols (2002), Scholl and Leslie (2001), Bishop and Downes (2002), Fuller (2013).

modules could develop. The second objection claims that TT(Innate) cannot account for default belief attribution. As mentioned above, it is useful for S to start from the assumption that O has the same beliefs as S. The modular nature of the TT(Innate) account means that it is informationally encapsulated. Informational encapsulation rules out access to the entire set of beliefs of S. The third objection claims that TT(Innate) lacks a parsimonious explanation for ToM deficits in autism. Autistic subjects show deficits in pretend play as well as ToM; these two deficits ought parsimoniously be explained together.

3.2 Objections To TT(Scientific)

3.2.1 Too Complex And Too Difficult

The most important objection to TT(Scientific) claims that it is too complex. This complexity makes TT(Scientific) an unappealing account of ToM for two linked reasons. Firstly, it makes the account unparsimonious. Secondly, it is implausible to ascribe such complex and difficult abilities to young children.[2] The second form of the objection may be set out as below.

- P1: Children can use ToM by around the age of five or earlier
- P2: Children cannot learn to use complex capacities by around the age of five or earlier
- C: Using ToM cannot require learning to use complex capacities

The problem for TT(Scientific) then is that it postulates just such complex capacities as are ruled out by the conclusion of this simple argument. The objection claims that the significant theoretical apparatus that TT(Scientific) postulates is too theoretically expensive and cumbersome. Admittedly, the

[2]This variant of the objection is forcefully pressed by Gustafson (1995) in his article aptly entitled "Eighteen Months On The Planet And Already A Psychological Theorist."

ToM capacities that are explained are themselves complex and sophisticated but that does not mean that a significant theoretical cost need be borne in explaining those capacities. Such costs are only acceptable when no cheaper theory is available. ST is exactly such a less theoretically costly theory, since it explains ToM capacities by using only machinery that is already present: people's own minds. No additional significant body of knowledge need be postulated, since that 'body of knowledge' can be generated on the fly by the mind of S.

It is accepted that children acquire ToM abilities at the latest by four or five years old, becoming able to pass the False Belief Task (Wimmer and Perner 1983). The complexity objection urges that there is not enough time for children to complete an extensive programme of hypothesis formation, confirmation and disconfirmation and theory building, as is envisaged under TT(Scientific). Moreover, almost all children must complete the programme eventually, even if they are cognitively disadvantaged. As Baron-Cohen, Leslie, and Frith (1985, p. 44) show, "severely retarded Down's syndrome children performed close to ceiling" on a ToM task.

Gopnik and Wellman (1992, p. 167) attempt to respond to this complexity/difficulty objection on behalf of TT. They argue that the objection requires that we have "some a priori way of measuring the temporal course of conceptual change, of saying what is slow or fast or easy or difficult." The key idea is that saying that a theory is 'too complex and too difficult' requires a measure of complexity and difficulty. Gopnik and Wellman (1992, p. 167) do not propose such a measure themselves, but suggest that "the three-year-old child may be working on the theory of mind virtually all his waking hours" and ask "who knows what adults could accomplish in three years of similarly concentrated intellectual labour?" so we can see that they mean something approximately like "three years of full-time work" as a yardstick for how long

it takes to complete a complex and difficult process. The problem that they will continue to insist upon is that one cannot exactly know how complex and difficult the task was, merely that it took three years.

Gopnik and Wellman (1992) attempt to shore up this response with an analogy from the history of science. Their response relies on the claim that we cannot assess the complexity and difficulty of Kepler's heliocentric theory. The 'a priori' element seems to be requiring that the difficulty of theory change is measured by some mechanism that does not just look at how long it takes in practice to perform theory change, whether one is Kepler or a three-year-old. Measured 'culturally,' developing the heliocentric theory may have taken centuries, as the necessary developments in observational technology and mathematical underpinnings were put in place. However, Gopnik and Wellman (1992, p. 167) argue that Kepler did not take a long time to formulate it and also a modern student learns it in "days, weeks or months." If we allow Gopnik and Wellman (1992) the reasonable assumption that easy tasks can be performed more quickly than harder ones, we would arrive at an argument suggesting that without an objective measure of the difficulty of a piece of theory change, we cannot arrive at a prediction of how long it ought to take to perform such a piece of theory change. And without that, we cannot deny that children are not doing something similarly easy or difficult or quick or slow than what Kepler did.

So Gopnik and Wellman (1992, p. 156) suggest that ToM change in children is like theory change in science when they write that: "during the period from three to four many children are in a state of transition between the two theories, similar, say to the fifty years between the publication of *De Revolutionibus* and Kepler's discovery of elliptical orbits." Their argument for this relies on developmental data, with specific focus on whether children have the idea of misrepresentation, or false belief. Experiments – in fact, one can

readily replicate this with any young children that are to hand – show that children first deal with the difficulties involved in passing the False Belief Task by denying the evidence. If one shows them a smarties tube and asks them what it contains, they will answer 'smarties.' If one then shows them that it in fact unexpectedly contains pencils, and then asks them what they believed was in the tube just previously, they answer 'pencils.' It is quite striking that one can obtain the false answer 'pencils' even when the gap between first saying 'smarties' and then subsequently responding 'pencils' is a matter of seconds (Cassidy et al. 2005, p. 105). The problem the younger children have is diagnosed as not having the concept of false belief, or misrepresentation. They simply cannot ascribe a false belief to anyone – whether themselves at an earlier juncture or other people. If X is the case, then everyone believes X. So the theory change, on the Gopnik and Wellman (1992) account, is that the five-year-old children have acquired the facility to ascribe false belief, and can now pass the False Belief Task.

This response on behalf of TT is inadequate for four reasons. Firstly, it seems unmotivated to exclude the underpinnings to Kepler's work that were prerequisites to his breakthrough. Even if we exclude work done by others —which is dramatically different to how science actually makes progress —it is hard to believe that Kepler started and finished his work on the heliocentric theory in a short period, even if it culminated in a breakthrough moment. This weakens the analogy and the evidence that both scientific theory change and children's ToM development are similarly quick and easy.

Secondly, the roles of the protagonists seem grossly dissimilar, weakening the analogy. There are two scenarios between which Gopnik and Wellman (1992) wish to draw an analogy. In one scenario, we have Kepler using the entire development of science, observational technology and mathematics developed by experts that was available to him to develop the heliocentric theory.

In the second scenario, children are developing ToM using only what is innate or observable or by the pre-five-year-old, who moreover is not born with observational or social capacities fully developed. A group of pre-fives and their mothers may well be highly social, but does not seem highly similar to the scientific community. In particular, Kepler seems to be a leading agent rather than a passive recipient.

There is some evidence that ToM development is driven by social interactions; a category I would include under the category of data observable by pre-fives. I will now discuss several experiments showing this, and will be suggesting a). that the results are consistent with ST and b). there is nothing here to support the science analogy.

For example, it has been found that children with a variety of siblings (but not twins) perform better on ToM tasks at a younger age than otherwise similar children. Cassidy et al. (2005, p. 103) suggest that "the sibling effect is associated not with mere exposure to another mind but specifically with exposure to a mind or minds different from one's own." This has been questioned though: C. Peterson and Slaughter (2003, p. 419) "found no significant correlation, in either study, between false belief understanding and number of siblings." Ruffman et al. (1998) found that having older siblings enhances false belief understanding, but also noted that their data was consistent with both TT and ST views; this supports my claim that ToM development through social interaction can equally well be accommodated on simulational lines. It is also worth noting that the data are mixed here; Ruffman et al. (1998, p. 170) found no "support at all [for] the idea that older siblings enhance source understanding." Source understanding means being able to answer 'how do you know?' questions; it is seen as one of the battery of standard ToM tasks. Hughes et al. (2006, p. 55) argue that "it is the quality (rather than the simple presence) of the sibling relationship that matters." They find that two-

year-olds with older siblings exhibit a higher frequency of talk about internal (mental) states; this correlated with ToM performance but the correlation was entirely explained by differences in verbal ability. So it looks as though general verbal facility enhances ToM; this result is entirely consistent with ST. There is little here however which is illuminated by the TT(Scientific) analogy. It might well be that being a professional scientist requires a high level of verbal ability, but it seems unlikely that the ability to create new theories in the subset of scientists who do it will be correlated with verbal ability. Overall, while theory change in science may be at a stretch called social, reading and writing journal articles seems to bear little similarity to two-year-olds playing with five-year-olds. Finally, there may not even be an effect here to explain, Hughes et al. (2006, p. 55) note that "other researchers working with more diverse samples have not found any significant association with sib-ship size."

Likewise, some have claimed that children with mothers who more frequently give explanations involving mental states or more frequently talk about minds tend to pass the False Belief Task at a younger age. C. Peterson and Slaughter (2003, p. 419) found that a "mother's preferences for [elaborate explanations involving mental states] options were predictive of their children's false belief understanding, but not of their understanding of the other ToM concepts tested, namely emotion understanding and gaze-reading." So the data are somewhat mixed. TT proponents would not be well served to claim that there is something similar occurring in the scientific community and in pre-fives which aids in passing the False Belief Task but not in the other items of ToM. They might be able to say that something beyond generalisations is required in gaze-reading; perhaps detecting the focus of the eyes is a difficult activity which is not based on generalisations. But emotional understanding seems to be clearly based on generalisations if ToM is, as TT(Scientific) holds. So why do children pick one type of generalisation up from relevant discus-

sions and not the other? Admittedly, these data are potentially tricky for ST to explain also. ST may say that children's emotional understanding of others will be correlated with their capacity to identify and indeed produce their own emotions. If they struggle to do so, perhaps because the process is introspection and this is hard, then ST would have an explanation but would of course have made a testable empirical prediction.

It is even possible to get children to pass the False Belief Task earlier than they would otherwise by giving them training relevant to understanding and reasoning about mental states. Wellman and C. Peterson (2013, p. 2358) found though "considerable individual variation," raising the question as to why "some children gain but others, exposed to the same conditions, do not." While this sits well with the TT(Scientific) line that ToM learning means conducting a very complex activity, but can something complex be learned by looking at cartoons for six half-hour sessions over six weeks, or by training over two weeks (Slaughter and Gopnik 1996)? Slaughter and Gopnik (1996, p. 2977) themselves admit that under TT(Scientific), "knowledge is represented in a complex, coherent system of concepts that are interrelated." These data also further weaken the analogy. To what in the Kepler scenario is this range of abilities analogous? ST can say that the 'thought-bubble' training provided gives the children practice simulating. It can also say that the complexity of the task is much reduced; the task is not to build up a body of knowledge by testing generalisations, but to gain skill in accessing what is already there: the child's own mind.

Lohmann and Tomasello (2003) use deceptive objects like a candle that appears to be an apple. They find that "perspective-shifting discourse using contentful linguistic symbols (not necessarily mental state language) aids in developing false belief understanding" (Lohmann and Tomasello 2003, p. 1141). These results are congenial to ST, as is indicated by their initial ci-

tation of ST proponent Harris as proposing one of four hypotheses relating language development to ToM development (Lohmann and Tomasello 2003, pp. 1131–1132). The data show that children exposed to the deceptive apple and discourse involving questions like 'what did you think it was?' improve on ToM tests. ST will suggest that this is because the children are now brought face-to-face with their own false belief, and are then invited to apply the new skill to a third party (a puppet dog). They are in effect being taught to simulate false belief.

Moeller and Schick (2006) study deaf children, who have been widely found to exhibit delays of several years in ToM development. Moeller and Schick (2006, p. 751) found that "[m]aternal sign proficiency was correlated with child language, false belief, and mothers' talk about the mind," suggesting that the link between language development and ToM development is still observed here. Notably, Moeller and Schick (2006, p. 752) cite the simulationist argument of Harris (1992) to the effect that "it is not just the mention of mental states in families that fosters children's ToM development; instead it is the back-and-forth shuttling from one viewpoint to another that makes a difference" in false belief understanding. Moeller and Schick (2006, p. 761) find that "the presence of siblings in the home who could sign was significantly correlated with the deaf children's false belief understanding, but not with their language scores" of which one interpretation is "that families where mothers and siblings are able to sign are providing more exposure to the back-and-forth shuttling of viewpoints through triadic conversations" postulated by Harris (1992) as the ToM improvement mechanism under ST.

Similarly, Slaughter and Gopnik (1996, p. 2984) think that TT(Scientific) is strengthened because the "training in this study not only affected children's performance on the false belief posttest, but transferred to other theory of mind posttest tasks." This they argue is positive for TT(Scientific) because

it outlines how concepts therein are coherently interrelated. But it could equally well be that false belief training enhances simulational capacities which also cross-over to other ToM tasks. Contrary to the claim of Slaughter and Gopnik (1996, p. 2986) this it is "not so clear how improvement in the skill of simulation could influence children's abilities to distinguish "guess" and "know," " it seems quite clear that making that distinction is already an improvement in simulation capacity. If S can improve his understanding of the difference by training involving puppets that use the terms, then S can pro tanto apply the new understanding within simulation.

If there is indeed a link between social interaction and ToM, the result seems equally congenial to ST and to TT. Both can say that such exposure improves ToM performance by, in the former case, developing simulational capacities and, in the latter case, by improving theoretical capacities. However, Cassidy et al. (2005, p. 111) also found that twin S's "did markedly better when the false belief in question was that of their twin [O] instead of that of a friend [O]." This one might say is a consequence of ST. However, explaining it commits TT to the claim that twin S's have a better theory of their own twin O's than others, even when the subject matter is the same in both cases viz. selection of the cupboard in which Maxi will look for the chocolate. Why would that be? If the answer is that the theory is the same, but the twin S's found it easier to apply to their twin O's, we may equally ask why that would be.

Thirdly, Kepler was a very special and talented individual while ToM is acquired by almost all children, including the less special, the less talented, and as mentioned above, the severely retarded. So it looks wrong to respond that we cannot say that TT(Scientific) is not too complex and difficult for children since we cannot say that Kepler's theory is too difficult and complex.[3]

[3]Nichols and Stich (2003, p. 108) also plausibly press this complexity objection, complaining that TT(Scientific) needs a lot of machinery to explain complex and detailed behaviour

Fourthly, what is observed is that three-year-old children do not pass the False Belief Task and five-year-old children do. There is no direct evidence of conceptual change in the children. The Gopnik and Wellman (1992) account whereby children undergo theory change – specifically, they acquire the concept of false belief – is just one possible account. Simulationists might alternatively point to the dramatic expansion of cognitive capacities that children undergo between three and five,[4] or note that five-year-old children have had two years of extra observation.

The severity of the complexity objection to TT(Scientific) is greatly increased by a range of converging data supporting the Onishi and Baillargeon (2005) breakthrough results. Ruffman et al. (1998, p. 171) note 1994 evidence "that children as young as 2 years 11 months to 3 years 2 months show signs of implicit but not explicit understanding" of false belief in that 80% of them looked at the correct location while only 20% could answer the question about false belief correctly. Onishi and Baillargeon (2005) themselves showed that even 15 month infants can succeed on non-linguistic variants of the False Belief Task. Luo (2011, p. 289) notes several reports that "children in their second year of life have been found to hold false-belief understanding, using non-verbal tasks;" Luo (2011, p. 295) found that "10-month-old infants may consider an agent's beliefs, true or false, when predicting and interpreting her actions." Strijbos and De Bruin (2013, p. 755) cite several replications of the Onishi and Baillargeon (2005) results including "13-month-olds." Heyes (2014, p. 647) includes "more than 20 experiments" favouring the claim that infants understand false belief in her review article. Against all this, it should be noted that Moeller and Schick (2006)[p. 757] had to eliminate non-verbal versions of the False Belief Task from their study since they found it confused

prediction in children.

[4]Gopnik and Wellman (1992, p. 167) themselves note the "general cognitive achievements of young children."

children who were around five and thus able to pass the verbal version.

One possible response here for TT would rely on the speculation made by Ruffman et al. (1998, p. 172) that "older siblings shorten the gap between the first sign of implicit understanding and the emergence of correct explicit answers." This however would do nothing to assist TT(Scientific) in relation to children with no older siblings, who are still completing a task which under TT(Scientific) is highly complex. In addition, Ruffman et al. (1998, p. 161) find "no such effect for children younger than 3 years 2 months" while the children for which TT(Scientific) lacks an explanation are much younger.

A second potential response would be to divide ToM into multiple stages (Butterfill and Apperly 2013), say that the simpler easier stage is what infants are using and then to say that children have the full five years to pass the verbal False Belief Task, because that needs to second stage as well. This could work, but means that TT(Scientific) is still saying that infants have learned some generalisations – those sufficient to 'pass' in the looking-time violation versions of the False Belief Task – with less than 10 months (Luo 2011) – or less than 7 months (Heyes 2014, p. 651) – to learn them.

A final way out which is initially more promising could be to argue that the infants are not in fact showing false belief understanding in the various experiments, but are responding to novelty with increased looking time (Heyes 2014). This would allow TT(Scientific) to at least retain the five year period of observation and generalisation formation which it could assume prior to Onishi and Baillargeon (2005) and replications. Her line can of course be questioned, as can be seen from the response following it in the journal. The central idea is that "infants look longer at test events that, when compared with events encoded earlier in the experiment, display new spatiotemporal relations among colours, shapes and movements" (Heyes 2014, p. 648) which is a testable empirical prediction. It also means that TT(Scientific) is committed

to the claim that the surprise of the infants is generated by novel combinations of "colours, shapes and movements" (Heyes 2014, p. 648), which is possible but somewhat unappealing. Why should red/square/up be surprising after blue/triangle/down? The account of Heyes (2014) also requires a rather convenient memory disruption effect.

Overall, it is still the case that if TT(Scinetific) cannot explain away these results, then the complexity objection is made more severe for TT(Scientific), since it means that TT(Scientific) is now postulating that the ability to handle great complexity and difficulty arrives at a very early age. As the ages at which children appear to develop implicit false belief understanding is reduced, TT(Scientific) is left claiming that children have not only completed whatever learning and generalisation formation is required, but have done so using a less developed set of cognitive capacities. This state of affairs should lead TT(Scientific) proponents either in the direction of TT(Innate)[5] or ST. ST is also somewhat challenged by these results, but at least it is only postulating that 10-month-olds (Luo 2011) can act as if they do not know about what they do not see and expect O's to be similar, while TT(Scientific) is postulating that they have learned a generalisation to that effect and can apply it to O's.

I conclude that this objection raises severe problems for TT(Scientific) which have not found an adequate response.

3.2.2 Requires Solving The Frame Problem

In this section, I will expand on an objection to TT(Scientific) which has been touched on in the literature but which has not in my view received attention commensurate with its gravity in the ToM arena. It is mentioned briefly by Heal (1996, pp. 81–84) in a section which deserves much wider attention, and a response is attempted by Glymour (2000), which I will outline below. I propose

[5]Fodor (1987, p. 132) uses this objection to argue for TT(Innate) against TT(Scientific); and for the cultural universality of belief/desire psychology, see §3.2.3.

that the objection is a major differentiator between TT and ST because we know that humans can solve the Frame Problem but we do not know how they do it. TT assumes, implausibly, that we have a set of generalisations that embody a solution. Heal (1996, p. 83), writing of the solution to the Frame Problem, notes the "oddness of supposing that we have it tacitly while at the same time possessing no inkling of how to set it out explicitly." I contend that ST may note, by contrast, that the mystery of how we solve the Frame Problem in relation to others is not distinct to or more mysterious than how we solve it in relation to ourselves.

The syllogistic form of the objection is as below.

- P1: If TT(Scientific) is the correct account of ToM, then S's must possess generalisations that solve the Frame Problem[6]
- P2: There are no generalisations that solve the Frame Problem
- C: TT(Scientific) is not the correct account of ToM

This is a serious objection since no ways of solving the Frame Problem are at hand, either in human psychology or artificial intelligence. Worse still, some authors have argued plausibly that the Frame Problem is insoluble.[7] In one sense, humans solve the Frame Problem all the time: whenever I decide to raise my arm, I do not in fact consider whether the gravitational field of Mars is relevant to how I will make the arm-raising happen. It is in the sense of providing a formal solution to the Frame Problem that it appears insoluble. This dichotomy is why the Frame Problem raises difficulties for TT but not for ST.

What is the Frame Problem? Every time we make a decision or form a belief, we must consider relevant facts before doing so if we are to do so

[6]This depends on the assumption that theories consist in generalisations only.

[7]Fodor (2008, pp. 116-121) sets out the difficulty of the Frame Problem in general; D. M. Peterson and Riggs (1999, p. 82) mention in passing its difficulty in their ToM context.

appropriately. The relevant facts form the 'frame' of a question. For example, if I want to decide whether to take an umbrella, I will learn a relevant fact from the weather forecast: whether it is expected that it will rain. There are other relevant facts which fall into the frame; potentially this number is quite large. The total set of known facts constitutes a 'model of the world;' the frame will be whatever subset of facts are relevant to the question at hand. Remaining with the example, my decision about the umbrella is defeasible by facts in certain other scenarios. For instance, I may abandon my previous decision to take an umbrella even if I learn from the weather forecast that it is expected to rain if it is also true that I do not expect to be outside for very long during the day. My model of the world is updated to include the new information that it is expected to rain, but may not result in changed behaviour when conjoined with other elements in the model of the world and my own expected behaviour. If my model of the world is updated to include new information about the weather in a remote location, this will not be in the frame as far as my decision to take an umbrella is concerned.

This leads to the Frame Problem. I must consider all of the actually relevant facts, but the number of potentially relevant facts is too large for them all to be considered. But how can I decide whether a potentially relevant fact is an actually relevant fact without considering it?[8] Thus it seems I need to examine every fact I know to see if it is in the frame for a particular question. That task is impossible. On top of this, I need an updated model of the world to reflect the consequences of my actions, which means I need to know what

[8]This formulation of the frame problem is quite approximate; I have adapted it to the specific context. More precisely, Shanahan (2009, §3) writes that the "epistemological problem is this: How is it possible for holistic, open-ended, context-sensitive relevance to be captured by a set of propositional, language-like representations" and how computationally "could an inference process tractably be confined to just what is relevant, given that relevance is holistic, open-ended, and context-sensitive?"

facts to change in the model. The Frame Problem occurs again because I need to work out what potentially changed facts in the world as a result of my actions are actually changed facts as a result of my actions. The two versions of the Frame Problem can therefore exacerbate each other.

The Frame Problem translates directly into problems for TT(Scientific) as an account of ToM. If S is to predict and explain the behaviour of O, how does S decide which of O's beliefs and desires are relevant, and which generalisations of ToM to apply on any given occasion of prediction and explanation?[9] S must somehow know which of an infinite array of beliefs are in the frame for an action without considering them all.

Since on TT(Scientific), ToM just is the application of generalisations[10] to beliefs and desires, there is no scope to avoid the Frame Problem. By contrast, on ST, S can employ whatever mechanism people use generally to avoid the Frame Problem when they make decisions. It may be that what that mechanism is exactly will remain forever beyond human knowledge, but there must be an answer, because we can make decisions. The answer will not be to use algorithmic mechanics like those employed in ToM on the TT(Scientific) account.

One response here might be to ask whether this objection shows that a theory theory view is incorrect in every case. That would be unappealing since it would entail that even scientists do not have theories. I am in effect suggesting that having and using a theory requires being able to solve the frame problem in the theory. The correct line here I believe is to note that

[9] Fodor (1974, p. 102) argues persuasively that the notions of 'law' and 'theory' are "equally murky." This raises problems that are side-stepped by ST but not by TT.

[10] As has now come up several times, it may be that some forms of TT do not involve generalisations, but TT(Scientific) certainly does. An example to consider might be a biological theory hinging on DNA, where it might seem that a structure is central rather than some generalisations about it. But the structure could in principle be replaced by a different one while the generalisations about evolution etc. are what makes the structure interesting.

scientific theories are *explicitly* theories i.e. they make generalisations that are written down and discussed. No-one can in this way write down a solution to the Frame Problem and I suggest that they never will be able to. But ToM nevertheless solves the Frame Problem. So TT(Scientific) needs to avoid being a theory of this sort. It can certainly avoid being explicit at no cost, but can it reasonably include generalisations that cannot even in principle be specified? I suggest not.[11]

A further response we might construct here on behalf of TT might be to urge that the Frame Problem need not be faced by S's under TT because only the salient generalisations will be of relevance. This risks circularity in that it says something like 'only the salient generalisations are salient,' which is true but unhelpful for TT because it does not explain how the salient generalisations become salient. In any case, my view is that this response will not work because TT has no recourse to any method other than additional generalisations to specify which generalisations are salient under particular circumstances, and that method threatens to require an infinite number of generalisations. It looks as though if tagging a particular generalisation as salient can only be done by considering that generalisation, explicitly or using implicit rules, which means that the Frame Problem arises in connection with the task of selecting which of the potentially salient generalisations are in fact salient.

Glymour (2000) attempts a response by restating the Frame Problem in terms of causation. In this form, the problem is knowing what facts to change in the model of the world as a result of a potential action: what effects will be caused by my action? This I need to know in order to decide whether it is a good idea to take the action or not. Glymour (2000, p. 65) writes, contra Hume (2000), that causation can be observed and learned. The child

[11] Again, being a theory without being constructed on generalisations might enable future TT which is not TT(Scientific) a way out here.

learning TT(Scientific) "notes associations either produced by its actions or otherwise, and the time order of associated events. From that information it infers that some associated features are not causally connected [...] or are more or less directly causally connected." For example, if the child observes that the arrival of parents occurs after it cries, it will conclude that crying causes the arrival of parents.

As Glymour (2000, p. 65) concedes, however, "the procedure is reliable only so long as a form of 'closed world' assumption holds, namely that the associations the baby [...] observes are not produced by unobserved or unnoticed common causes." Imagine that every Sunday, father plays football. This causes two things: he sleeps in the afternoon and runs the washing machine, causing vibrations. The child might falsely conclude that sleeping in the afternoon on a Sunday caused vibrations. It would form all kinds of inaccurate ToM generalisations like 'people who get muddy in the morning sleep in the afternoon' or 'people who sleep in the day on Sundays vibrate.' This particular difficulty might eventually be soluble empirically for the child scientist of the TT(Scientific) account, though it is not clear how if Sundays are always the same up to the age when the child completes its ToM development. However, the TT(Scientific) account needs the child to disentangle correctly all causal chains involving actions and beliefs which are to form the child's data for developing its ToM. The 'closed world' assumption must generally hold, if the TT(Scientific) account is to avoid predicting that children perform badly on ToM tasks, which is the opposite of the truth. Since the closed world assumption does not in fact hold, I conclude that Glymour has not provided an adequate response to the Frame Problem objection to TT(Scientific).[12]

Here ST appears more plausible than TT. ST requires only that S can believe P as opposed to solve the Frame Problem in relation to believing P,

[12]Further difficulties for Glymour derive from the extended argument presented by Taleb (2007) to the effect that humans ascribe more causation than is justified.

and we already know S can believe P. Here I disagree with Wilkerson (2001), who attempts to suggest that the Frame Problem is also a problem for ST because it must be solved in order to generate generalisations. This is once again "setting the bar too low" (cf. §2.8) since ST can be describable by generalisations without it needing to apply those generalisations.[13]

3.2.3 Cannot Explain Convergence

This objection claims that TT(Scientific) cannot explain why children from different cultures develop the same ToM at the same time. The objection runs as below.

- P1: If TT(Scientific) is the correct account of ToM, children develop their ToM by observing relevant behaviour around them
- P2: The relevant behaviour around them is different in different cultures
- C1: If TT(Scientific) is the correct account of ToM, children will not all develop the same ToM
- P3: Children all develop the same ToM
- C2: TT(Scientific) is not the correct account of ToM

This objection can in fact be more general: different people in the same culture can come up with different ToM. This does not make the objection easier to handle for TT(Scientific) proponents. One question which arises here is whether ST is consistent with there being variation in ToM. I think it is certainly consistent with there being little fundamental variation, but the other

[13]See also Dreyfus (2006) for Heideggerian argument to the effect that "[o]nly if we stand back from our engaged situation in the world and represent things from a detached theoretical perspective do we confront the Frame Problem" i.e. TT faces the Frame Problem while ST does not.

possibility might be thought to be more problematic for ST. One immediate response which is now available as a result of the account I am proposing is to recall the major effects of biases. For all of the experiments I will later discuss, ST results in different predictions – which are often ToM errors. This I think should go a long way to make ST consistent with there being variation in ToM, especially when one recalls that I will be arguing that the bias mismatches are driven by inter alia affect mismatches and system mismatches. So this route is available to ST if it is needed. Whether it will be or not depends on the outcome of a lively debate in the literature, as I will briefly outline below.

Premise 2 in the first sub-argument appears highly plausible merely from experience. There are significant differences in behaviour even among developed nations, with Japanese culture being more collectivist and less individualist than the American one, for example. It is no objection here to suggest that if young children spend most of their time in a nursery, they will not observe culturally specific behaviour, because inculcation of culturally specific behaviours starts young. As Prinz (2011, p. 222) observes, data exists to show that "in contrast to Americans, Japanese parents [...] introduced toys into play as opportunities for sharing [while] American parenting practices foster independence." So even if the children were only to observe the behaviour of other children when forming their ToM, it would still be culturally specific behaviour with more sharing behaviour in Japan and more individualistic play in America. Moreover, moral judgments are culturally highly-specific (Prinz 2006, p. 40) and we may assume that the cultural differences in moral judgements will drive differing sorts of behaviour in adults. We would therefore also expect that different cultures would have different generalisations in their ToM or, what is an equivalent statement on TT, different ToM in fact.

Premise 3 in the second sub-argument of this objection holds that all children in all cultures develop approximately the same ToM. They make more-

or-less the same behavioural predictions under similar circumstances. As Carruthers (1996, pp. 31-32) puts it, "it remains remarkable that all normal children should end up with the same body of knowledge at about the same time." Anecdotally, we would expect to have noticed by now if people we met from different cultures had a different ToM, because ToM is such a ubiquitous underpinning of human interaction and conversation. Empirically, children from Western Europe (Wimmer and Perner 1983) and North America (Gopnik and Astington 1988) perform very similarly on the False Belief Task; as do children from a preliterate society in Cameroon (Avis and Harris 1991). Children from all of these disparate cultures begin to pass the verbal False Belief Task by the same age, around five. Avis and Harris (1991, p. 460) write that their Cameroon results "provide support for the claim that belief-desire reasoning is universally acquired in childhood." The strength of this objection to TT(Scientific) is increased by the similar ages at which children from different cultures pass the False Belief Task, since this entails that not only do children acquire the same ToM from apparently different data, but they do so at the same speed, meaning that all cultures appear to offer similar richness of relevant ToM data. Wellman and C. Peterson (2013, p. 2358) note that there is a five-stage sequence of ToM developments which children in the US, Australia, Germany, China and Iran all pass in the order at the same ages. So Premise 3 appears empirically to have a fair amount of empirical support; I will discuss the opportunities for TT(Scientific) to deny it below.

It might be a response here to ask whether there is a mismatch between my defence of Premise 2 and Premise 3. On Premise 2, I have pointed out that there is variety in moral judgements. On Premise 3, I pointed out that there is convergence on the False Belief Task. The question is: how do we know that there is not divergence in ToM appropriate to the divergence in data? My response is that we do not, but we do know that there is not

divergence in performance on False Belief Task, if one accepts that data I outlined in the previous paragraph. TT(Scientific) still needs to explain this, or explain it away, even if later work shows that cultural ethical differences result in cultural ethical differences in ToM. One might expect the latter; since often S will predict that O will behave ethically and will decide what is ethical by consulting his own ethics. Note how that is highly congenial to ST. TT(Scientific) presumably needs to make the whole of ethics available within ToM.

On TT(Scientific), children develop their ToM by forming hypotheses and confirming or disconfirming them based on the behaviour they see around them. The problem for TT(Scientific) is that this behavioural evidence base will be very different in different cultures, with more collectivist behaviour in Japan and more individualistic behaviour being observable in the US. So it is hard to explain why there is a cross-cultural convergence in ToM. The objection to TT(Scientific) is then that it is committed to the interim conclusion C1 and that conclusion is empirically false. Four potential responses are available to proponents of TT(Scientific). They may deny premise 1. They may deny premise 2. They may deny the interim conclusion C1. Finally, they may embrace the intermediate conclusion C1. If they do that, they are committed to denying premise 3 in order to avoid C2, which falsifies TT(Scientific). I will consider each of these potential responses.

It is hard to see how TT(Scientific) could deny premise 1 (the claim that children develop their ToM generalisations by observing relevant behaviour around them) without becoming TT(Innate). If there are no generalisations, then we have ST[14] If there are generalisations, we have TT. If they have gen-

[14]Unless an account can be constructed which embodies a theory that does not make generalisations. I do not think that the account of Hughes et al. (2006) linking internal state talk and ToM offers a way out here; interactions leading to improved ToM will still be coded using generalisations if there is a body of knowledge.

eralisations, then either children learn their ToM generalisations scientifically or they do not. If they do, we have TT(Scientific). If they do not learn them, then they are already there, and we have TT(Innate). It is unclear what other options to put them there, short of supernatural ones, exist.

TT(Scientific) could attempt to deny premise 2 (the claim that relevant behaviour around them is different in different cultures) by asserting that there is a hidden equivocation on 'relevant.' Such a response would claim that while there are many surface cultural differences in behaviour between different societies, what matters to the development of ToM is 'deeper.' So while the stockbroker in Manhattan may go to a hot dog stand while the Japanese doctor visits the sushi restaurant, both are acting on similar belief/desire pairs. Both desire food and believe that they will be able to obtain some at the stand or the restaurant, respectively. It is the way belief/desire pairs interact to produce behaviour that is important to the development of ToM; the exact content of the belief/desire pairs is unimportant. This response seems unlikely to succeed however since actual behaviour predictions are more fine-grained. The S is in fact able to predict the eating of hot dogs or sushi as opposed to the eating of food. Also, the S can predict that the Japanese doctor might well eat a hot dog when in Manhattan.

Gopnik and Wellman's response to this objection takes this form of denying premise 2, as Segal points out. Gopnik and Wellman aim to make the denial of premise 2 plausible by explaining why it is that children converge on the the same, culturally non-specific ToM. As Segal (1996, p. 153) writes, Gopnik and Wellman suggest "that if adults converge on the same [ToM] in different cultures, then we would not expect much cross-cultural variation in children. But this is not a good response. First, we might ask how the adults happened to converge. The obvious answer is that they converged as children." So the response claims that the children converged on the same ToM because the

adults did, even though both children and adults grew up in different cultures.

Segal provides a powerful criticism here of the Gopnik and Wellman response. That response I think is like avoiding solving the problem by pushing it back a level. There is something of a pair of analogies here between the response, and its criticism, and theistic explanations of the creation of the universe, and criticism thereof. Those who postulate a divine creator of the universe owe an explanation of the creation of the divine being; failing that, they have merely pushed the problem back a level. Because of this, in my view, Segal has raised a criticism of the response which carries the day.

This leaves only the fourth option for TT(Scientific) proponents: to accept the interim conclusion C1 (that children will not all develop the same ToM) and deny premise 3 (the claim that children all develop the same ToM). TT(Scientific) proponents could cite additional empirical data and claim that it is at odds with that supporting the claim that ToM is universal. For instance, Strijbos and De Bruin (2013, pp. 746–747) claim that "there are large differences between the mature folk psychologies of various cultures." Strijbos and De Bruin (2013, p. 746) cite empirical studies which "have confirmed that there are several cultures without concepts analogous to BELIEF, DESIRE [...]" and note a further paper which argued that even " "mind" is a unique English-specific construct without precise equivalents, even in European folk psychologies such as that of the French, German, Russian or Dutch." Data to support these rather contentious claims is provided by asking subjects questions; in certain cultures respondents frame their responses along the lines of 'what will he say?' rather than 'what does he believe?' It seems as though there is room to suspect that this reflects mere superficial linguistic differences. It also seems as though if it were true that European cultures as close to that of native English speakers as France and Germany lack the same construct "mind," then this would be readily apparent. It is worth noting that

Strijbos and De Bruin (2013) think that their argument is problematic only for accounts of ToM which embrace standard belief/desire psychology. That seems to include all variants of TT, though arguably Gordon's version of ST would escape the difficulty in virtue of only requiring the ability to have beliefs and desires rather than the concepts of belief and desire. ST can also claim that ToM is not culturally specific since children's minds are not very culturally specific to begin with; Segal is in fact using this objection to argue for TT(Innate) as opposed to TT(Scientific). It looks though as though TT proponents would have to embrace the rather unappealing line set out above, which would represent a substantial theoretical cost for them.

I conclude that TT(Scientific) has no adequate responses to the three objections canvassed and now we may consider TT(Innate).

3.3 Objections To TT(Innate)

3.3.1 Cannot Explain Development

This objection claims that TT(Innate) cannot explain the observed development of ToM abilities in children. The objection runs as follows.

- P1: ToM capacities develop via data-driven learning
- P2: A correct account of ToM must explain how ToM capacities develop via data-driven learning
- P3: TT(Innate) does not explain how ToM capacities develop via data-driven learning
- C: TT(Innate) is not the correct account of ToM

The idea behind this objection is the claim that it is difficult to provide a mechanism for development of modules when those modules are informationally encapsulated. TT(Innate) proponents have responded to the objection

in two ways. Firstly, they have sought to deny P1, which means that they have no data-driven development to explain. Secondly, they have attempted to expand TT(Innate) so as to provide an explanation for how ToM capacities can develop. This response takes the form of denying P3. I will consider both responses in turn.

Denial of P1 at first seems unappealing, given the weight of empirical evidence supporting it (Wimmer and Perner 1983), (Gopnik and Astington 1988), (Avis and Harris 1991). It has nevertheless been attempted, by denying that passing the False Belief Task requires only ToM capacities. If that were so, then younger children failing it and older ones passing it would not be conclusive evidence for data-driven development in ToM capacities. Their ToM capacities could remain constant over the period at issue but the non-ToM capacities also required to pass the False Belief Task might only become available later. Bloom and German (2000, B27) argue that the verbal False Belief Task "is too hard for 1- and 2-year-olds, as they lack sufficient attentional and linguistic resources to cope." The response relies on the range of empirical evidence discussed above (§3.2.1) that shows younger children passing the task if it is made simple enough for them to understand. As Bloom and German (2000, B27) admit, though, this argument has not convinced supporters of the developmental change view such as Gopnik. Such developmentalists note that the empirical data merely shifts the development to a younger age but does not eliminate it. This leaves TT(Scientific) relying on what appears to be an ever-shorter window for data-driven learning, which may now be less than 7 months (Heyes 2014, p. 651).

One experiment sought to investigate whether language training could assist with ToM capacities, as the Bloom and German (2000) line seems to require. Partial support for the line was found, in that training on sentential complements such as 'A thought that B did X' produced improved performance

on the False Belief Task. However, Hale and Tager-Flusberg (2003, p. 355) also write that the "children who were trained on false belief showed equivalent developmental changes in theory of mind as did the children trained on sentential complements," so the improved sentential component abilities are sufficient for ToM improvements but not necessary. This seems difficult to explain on the Bloom and German (2000) line.

Note also that pursuing this non-developmental approach requires assuming that the medium-sized army of psychologists who have written on ToM using the False Belief Task have all been conducting experiments which do not measure ToM capacities. It is safe to say that this area of debate is extremely controversial and so denial of P1 is not a low-cost option for TT(Innate) proponents. Finally, I will note the discussion I have provided elsewhere (Short 2015, Ch. 10) on schizophrenic subjects. These individuals exhibit widely replicated ToM deficits which a) come and go as their symptoms appear and remit and b) are independent of cognitive ability. One schizophrenic subject with an IQ of 125 failed eight of nine simple irony comprehension tasks. Anti-developmentalists who suggest that apparent ToM deficits are caused by cognitive deficits need to explain this. They will also need to decide whether they accept that the irony comprehension task is a measure of ToM performance, as is widely accepted in the literature.

I now turn to the attempts to add a developmental explanation to TT(Innate). I will observe in passing that denial of P3 is incompatible with denial of P1, so defenders of TT(Innate) will have to choose. TT(Innate) proponents have so far been Modularists, claiming that ToM is subserved by a Fodorian module. This Modularism has shaped the two ways in which TT(Innate) proponents have attempted to deny P3. As discussed above, the originators of TT(Innate), Scholl and Leslie, suggest that the Modularist can account for development by postulating that different modules may

come online at different points. Scholl and Leslie (1999, p. 131) write that "[m]odules must 'come on-line,' and even fully developed modules may still develop internally, based on their constrained input."

The account immediately loses some of its appeal and parsimony[15] in virtue of being forced to postulate multiple modules subserving ToM. It will be difficult for the account to escape charges of being ad hoc and unmotivated. The most serious problem with this response on behalf of TT(Innate) is that one of the central elements of what it is to be a module is informational encapsulation. Information cannot pass across the boundary of a module in either direction. This seems to mean that the multiple modules postulated cannot communicate with each other. The exception to this is in relation to the inputs and outputs to a module: that information and only that information can cross module boundaries. However, we would then be left with an account on which the several modules subserving ToM on the modified TT(Innate) account could only communicate via these inputs and outputs. They could in other words only be connected in a chain, with the output of one being the input to another. It is hard to see how the TT(Innate) account as modified could square the circle of there being multiple modules which can only feed into each other and that still meaningfully counting as being multiple modules. The interaction picture could perhaps be more complex, involving for example feedback loops between modules. This would add flexibility to the account but threaten to make it unmanageable and unparsimonious.

TT(Innate) proponents may demand to know: what evidence is there that ToM abilities are a consequence of developmental changes that cannot be explained by invoking innate modules which might come online at a particular point in development? I concede that it is difficult to imagine such evidence. The criticism would have to be, rather, that a very large number of modules

[15] Here again the original p. 32 definition of minimal number of moving parts.

would have to be postulated to explain the data. There would need to be a set of modules to explain the non-verbal False Belief Task data which come online in the period from perhaps 7 months to 3 years, and a second set of modules to explain the verbal False Belief Task data, which come online in the period from perhaps 4 years to five years. The first set would already include a set of complicated generalisations about 'O does not know about what O does not see'; 'O reaches for what O desires' and 'O cannot see through visors.' Perhaps the latter is implausibly innate, and so the account would have some learned generalisations as well. The second set would either include a further set of generalisations or it would say that the first set was sufficient. I am not sure what they would be in option one. In option two, the account would be taking the Bloom and German (2000) line that what changes in the verbal version of the False Belief Task is verbal abilities not ToM ones. It will then still need to deal with the Hale and Tager-Flusberg (2003, p. 355) results. Overall, this account would be doubly unparsimonious on the original p. 32 definition of moving parts, because it has a lot of modules and a lot of generalisations.

It is also hard to see how these ToM modules could develop internally, since "ToM has a specific innate basis in that the essential character of ToM is given as part of our genetic endowment." (Scholl and Leslie 1999, p. 134) This appears to be a claim that the development of the modules is genetically specified. That is prima facie implausible, because it would not appear to allow much scope for differential development speeds in different populations, including some clinical populations. The TT(Innate) account also becomes committed by this line to a genetic explanation of all ToM differences. That means for example that autism has a genetic basis, which may be true, but is an empirical question. It is of course true that TT(Innate) proponents are not required to sign up to all of the commitments of Scholl and Leslie (1999). It does seem though that given the assumption that everyone has a broadly

similar ToM, modular accounts of ToM will be committed to everyone having broadly similar ToM modules and connections between them. Otherwise the account would be claiming that broadly similar ToM performance can arise from several different modular constructions, which is possible but lacks independent motivation. So it looks like such accounts would need to explain why everyone has a broadly similar construction of ToM modules, which looks in turn like it will rely on a genetic explanation. A further response here for TT(Innate) might involve noting that to say that a module is genetically specified does not preclude the possibility that the expression of the genes specifying it is affected by environmental factors. That seems to allow the account to explain the ToM deficits of, for example, autistic subjects, by suggesting that their ToM deficits result from atypical development of their ToM modules which in turn results from environmental factors affecting how the relevant genes are expressed. This seems possible, but is an empirical question and also an assumption of some magnitude.

More troubling still for TT(Innate), the ToM deficits of all clinical groups which exhibit them will have a genetic basis. Schizophrenic subjects, for example, exhibit ToM deficits when they are suffering from the effects of schizophrenia and not when they are not so suffering (Koelkebeck et al. 2010, p. 115), (Bora and Pantelis 2012, p. S142). About a third of schizophrenic subjects have a relapsing/remitting form of the disease whereby they go through periods when they are not suffering the effects and periods when they are. A genetic explanation of such a pattern of relapsing/remitting deficits in ToM seems difficult to provide, even if TT(Innate) can avail itself again here of the potential escape routes canvassed above. Again, there are significant empirical predictions here which would need testing. I conclude that pending such empirical data, the attempt of Scholl and Leslie (1999) to deny premise 2 fails.

The second attempt to deny P3 (the claim TT(Innate) lacks a mechanism

to explain development) is due to Segal (1996). Segal offers a response to the developmental objection involving the use of parameters. The response claims that the modules making up ToM are innate but can develop according to the switching of parameters. As the the child learns, it does not create new modules, but tunes innate modules by the setting of such parameters. As an example, we may consider the way that the child's ToM improves in relation to belief and pretence. Younger children are unable to distinguish between belief and pretence. This means that they cannot distinguish between a situation in which someone asserts in the context of a pretend teddy-bear's picnic that teddy is drinking tea and a genuine belief that teddy is drinking tea. Formally, the lack of such a distinction is indicated by the child continuing to use the inaccurate concept of PRELIEF, which has not yet separated out into the two adult concepts of PRETENCE and BELIEF.

The way development would occur in such a scenario would be that initially a parameter – call it the PRELIEF parameter – is set to ON and the BELIEF and PRETENCE parameters are both set to OFF. As the child develops, it gains the ability to distinguish the two concepts and so the PRELIEF parameter is set to OFF and the BELIEF and PRETENCE parameters are both set to ON. By this, I simply understand the parameter as an on/off switch, which represents whether a particular module or a submodule is operative. As a way of denying premise 2, and allowing TT(Innate) to accommodate development in ToM, the parameter response appears more promising than the multiple module response of Scholl and Leslie (1999). It nevertheless fails in my view because of the rather large and cumbersome nature and number of parameters that would be needed. We already have three parameters here to accommodate a single development, and it appears that there is a vast amount of such development that takes place before the age of five. We would need further parameters to represent all developments in ToM, such as 'people who want an

X and think they are getting one will be happy' and 'people who see that X is the case know that X is the case' and in fact all of the generalisations foreseen in TT. So, maybe several thousand parameters would be needed to explain the enormous amount of development that occurs in the child's ToM between the ages of three and five. ToM(Innate) defenders may respond with a denial that there are an enormous amount of changes, but not necessarily by denying that there are any. This would equate to taking the position that a minor number of changes in the ToM modules could explain the very great changes in ToM performance, whose existence seems empirically unassailable. Such a line would therefore involve enormous changes hanging from few parameters, which seems implausible, especially when one considers the example given of PRELIEF. The parameter which controls the switch to PRELIEF would also have to trigger a large number of similar changes, which then appears to have the account saying that switching this one parameter produces many improvements in ToM simultaneously. This then looks less like a parametrised account of ToM and more like a general development picture of ToM. In addition, the account would need to explain how the parameters became switched around again or disturbed in schizophrenic subjects in a period where they are exhibiting symptoms, since a substantial minority of schizophrenic subjects exhibit episodic ToM impairments (Short 2015, Ch. 10).

A further problem with the large number of parameters is whether their use is even consistent with TT. As Scholl and Leslie (1999, p. 144) point out, Stich and Nichols "defend a modularity view against a 'theory' theory by repeatedly pointing out that a module with enough parameters effectively reduces to a theory." This is a slightly strange defence, in that it seems in fact to point to collapse risk between TT(Innate) —the 'modularity view' —and TT(Scientific) —the 'theory' theory view. If TT(Innate) is not separate from TT(Scientific), then it cannot be defended by attacking TT(Scientific).

However, for our purposes, collapse risk between different sorts of TT is not a concern. In any case, it does appear that both accounting for development of ToM by postulating either a large number of parameters, or a large number of stages of module initiation, represents a significant lack of parsimony[16] and is rather ad hoc. I conclude that TT(Innate) cannot explain the observed development of ToM without making large number of ad hoc changes which substantially destroy its parsimony as an account of ToM.[17]

ST can avoid this developmental objection by suggesting as a first approximation that the child can simulate in others what it can do itself. That account is also naturally developmental, but also allows for a less than total correlation between development of the child's own mental capacities and its abilities to simulate those same mental capacities in others because the latter task is more difficult.

3.3.2 Cannot Explain Default Belief Attribution

Default belief attribution is the process whereby S's starting point for using ToM in relation to O is to make predictions on the basis that O has the same beliefs as S. Default belief attribution seems to occur. This objection claims that TT(Innate) cannot explain such default belief attribution. The objection takes the following form.

- P1: If TT(Innate) is the correct account of ToM, there is no default belief attribution in ToM
- P2: Default belief attribution is a very valuable starting point in ToM capacities

[16] On p. 32, I defined unparsimonious as meaning having a large number of working parts. Here we have something slightly different, in that there are a large number of parameters or stages.

[17] Karmiloff-Smith (1998) also suggests that Williams Syndrome data are hard to explain on TT(Innate) for developmental reasons.

- P3: If TT(Innate) is the correct account of ToM, ToM capacities do not avail themselves of a very valuable starting point
- P4: ToM capacities avail themselves of a very valuable starting point
- C: TT(Innate) is not the correct account of ToM

Premise 2 seems very plausible. If S is to predict the beliefs of O, it is very valuable as a first approximation for S to ascribe to O all of S's beliefs that are relevant. It will be impossible for S to predict O's behaviour without ascribing any beliefs, while it would also be inefficient for S to explicitly consider what O's beliefs are on a case–by-case basis. Since S and O are roughly similar and inhabit roughly similar worlds, it is a valuable efficiency gain for S to start from the assumption that all of S's beliefs are shared by O.

ToM could succeed in many simple cases without going further. Imagine S is to use ToM to explain the behaviour of O who has just entered the coffee shop. S can ascribe a desire for coffee to O, since some desire has caused O to act and that seems like a good candidate. S can ascribe S's belief that there is coffee in the coffee shop to O without further ado; S now has a belief/desire pair to ascribe to O which explain O's behaviour. Leslie, Friedman, and German (2004, p. 528) accept premise 2 on behalf of TT(Innate) when they write: "because people's mundane beliefs are usually true, the best guess about another person's belief is that it is the same as one's own."

The difficulty for TT(Innate) will be seen when O has different beliefs to S. This is a central case in ToM since predicting the behaviour of O's who have false beliefs, which are perforce different to those of S in the False Belief Task paradigm, is the most common experimental test of ToM. S must here decide which of O's beliefs are different to S's and false. It still is an efficient start point in either the experimental or an everyday setting to ascribe in some sense all of S's other beliefs to O, such as 'I have enough money for coffee'

and 'I can safely enter the coffee shop,' in order to predict O's behaviour. Note though in passing how such an approach is much more congenial to ST than either variant of TT. Both variants of TT would need to solve the Frame Problem here in order to decide which of O's ancillary beliefs were relevant to O's behaviour whereas on ST these ancillary beliefs can just remain in the background unless needed, as it were. On ST, S does not need explicitly to consider whether O believes that it is safe to enter the coffee shop unless there is a specific reason to do so. ST avoids this default belief attribution objection because the whole approach under ST is to simulate O as being like S, meaning that the starting point for S is default belief attribution to O of S's beliefs.

Premise 4 (the claim that ToM capacities avail themselves of a very valuable starting point) seems hard to deny, since its denial entails that ToM performance is generally poor, which is empirically false. A denial of premise 4 also entails that there is no default belief attribution in ToM, which again seems contradicted by experience.

Premise 3 can be denied, but one of premise 1 or premise 2 must also be denied on that route. As discussed above, denying premise 2 is unpromising and TT(Innate) proponents have not gone down that route. This leaves the option of denying premises 1 and 3, which means finding a way of allowing TT(Innate) to accommodate default belief attribution. Nichols and Stich (2003, p. 120) note that Leslie attempts to do this when he writes that the ToM mechanism "always makes the current situation available as a possible and even preferred content because (a) the current situation is a truer picture of the world, and (b) beliefs tend to be true' ." Note that 'the current situation' is shorthand for a vast amount of data which is to represent a 'picture of the world' as seen by O. All of these data must be available inside the ToM module; moreover, this picture of the entire world must also be adjusted for O's false beliefs or errors in his picture of the world.

Some observers attempting to deny Premise 1 on behalf of TT(Innate) might respond to this difficulty by urging that in fact, S only needs the content of the belief which O would have if O had a true rather than a false belief on the matter concerning which O has a false belief. This is certainly a more parsimonious[18] line, if it can be made out. One issue with the line will relate to the individuation criteria for beliefs; does someone who believes a single proposition also ipso facto believe all of its entailments? Perhaps TT(Innate) defenders can avoid this potential pitfall by saying that if the entailments are indeed required, they in some sense come for free. It will still be the case though that there is a very large number of beliefs inside the ToM module, which does not sit well with encapsulation. As I will now outline, TT(Innate) defenders have not in fact taken this line, which I believe is telling.

In sum, not only is this account immensely unparsimonious as an explanation of ToM, but there is also a fundamental conflict between this idea and the advertised Fodorian nature of the modules proposed in TT(Innate). The conflict flows from the aforementioned fact that "an essential characteristic of modules is that they are informationally encapsulated" (Nichols and Stich 2003, p. 120). So how can 'the current situation' be made available to an encapsulated module? Indeed, "a cognitive system that has unrestricted access to all of the mindreader's beliefs would be a paradigm case of a non-modular system" (Nichols and Stich 2003, p. 121). The TT(Innate) proponents attempt to respond to this serious objection by introducing a new item, called a 'Selection Processor' (SP) which is to handle inhibition of some of the default beliefs ascribed to O by S. This is unmotivated, but worse, Leslie, Friedman, and German (2004, p. 532) ask the speculative question as to whether "SP, and not ToMM, [could] be the source of the true-belief default? SP is a non-

[18]Here, again in slight contrast with the definition of parsimony I gave on p. 32, lack of parsimony means fewer beliefs rather than fewer moving parts.

modular, penetrable mechanism."[19] This introduction of a non-modular item to TT(Innate) loses parsimony[20] and makes the TT(Innate) account incoherent. If a closed module must communicate with a non-closed module, in what way is the system still modular and encapsulated?

I conclude that TT(Innate) does not have an adequate response to the default belief attribution objection.

3.3.3 Cannot Parsimoniously Explain Autism

Some children exhibit a deficit in pretence, meaning that they engage in pretence less often and at a later age than other children. Such a pretence-deficit is a noted feature of autistic children. Autistic children also exhibit ToM deficits. This objection claims that TT(Innate) is not a parsimonious[21] explanation of autism because it does not explain both the pretence and ToM deficits seen in autistic subjects. More widely, as Ruffman et al. (1998, p. 161) point out, "[n]umerous researchers have suggested that false belief understanding may be assisted by pretend play." Such a unified explanation is a consequence of ST. Formally, the objection is as set out below.

- P1: A major symptom of autistic subjects is a deficit in pretend play
- P2: A major symptom of autistic subjects is a deficit in ToM
- P3: A parsimonious account of ToM must explain both deficits
- P4: TT(Innate) does not explain both deficits
- C: TT(Innate) is not a parsimonious account of ToM

[19] As mentioned previously, the authors use in their TT(Innate) approach a 'Theory of Mind Mechanism,' or 'ToMM.'

[20] Here exactly as specified on p. 32 i.e. too many moving parts.

[21] Here, the other aspect of the parsimony definition I gave on p. 32 comes to the fore: with a given number of moving parts, how much explanatory power does an account have?

This objection is philosophically slightly weaker than the foregoing two, in that it points to a demerit in TT(Innate) as opposed to a fatal flaw. Nevertheless, as Nichols and Stich (2003) correctly observe, the objection constitutes an embarrassment for TT(Innate) proponents since they focussed on autistic subjects to provide much of their supporting evidence. Nichols and Stich (2003) also correctly note that proponents of TT(Scientific) have not addressed in detail the topic of autism, and that this is a problem since autistic subjects have well-known ToM deficits which all plausible accounts of ToM should explain. ST parsimoniously explains the paired deficits since on ST, S is the model for O. TT(Innate) proponents could conceivably find another type of evidence to support their case, but a modular account naturally will find its best support from paired deficits, involving the claim that a single deficient module explains both deficits. I will start by laying out the original TT(Innate) account of autism and then consider the objection in more detail.

The original TT(Innate) account of autism begins from two suggestive facts about autistic subjects. They engage in pretend play much less and much later than non-autistic children (Baron-Cohen, Leslie, and Frith 1985). They have well-known ToM deficits, passing the False Belief Task much later than non-autistic children, even when matched for IQ (Baron-Cohen 2001). These two empirical claims are not disputed and mean that all sides of the debate accept premises 1 and 2.

Leslie (1987), who is one of the main proponents of TT(Innate) (Leslie, German, and Happe 1993), (Scholl and Leslie 1999), (Scholl and Leslie 2001), (Leslie, Friedman, and German 2004), agrees that these two deficits are related. Nichols and Stich (2003, pp. 128-129) set out the TT(Innate) view here as follows: "Leslie also maintains that mindreading is central to pretence and [...] ToMM plays a central role in the capacity for pretence [...] It is ToMM [...] that does not develop normally in people with autism." So the

TT(Innate) view is that the undeveloped ToMM in autistic subjects causes both the impaired mindreading and the lack of pretend play that is observed. This is consistent with acceptance of premise 3.

One merit of an account which accepts premise 3 is that it explains why children spend so much time engaged in pretend play. They pretend that they are, for example, at a tea party with teddy bears who drink pretend tea and, it is pretended, enjoy conversation. The children are in fact exercising their ToM, which brings important social advantages in childhood and later. The type of exercise involved would be in predicting what teddy might say about the tea he is enjoying, and how he might later say he has had enough tea and it is time to go back to the woods etc. All of this is good practice in ToM use for the children, who we know are in a way predicting the speech and action of teddy because they are supplying teddy with that pretended speech and action.

All accounts of pretence must explain quarantining, or the way some propositions that are held true only within the pretence must be separated from the general beliefs of the pretending subject. Leslie (1987) terms this 'decoupling' and postulates that it is a failure of decoupling that explains both the lack of pretend play and the ToM deficits of autistic subjects. It would not occur, for example, that an adult who had been pretending to be at the teddy bears' tea party would refuse an actual cup of tea later on because they had already pretended to drink a cup of tea. Leslie (1987) proposes that one of the mental representations underlying the tea party pretence might have the form: I Pretend 'this empty cup contains tea.' So there is a Pretend operator which operates on an actual object, the empty cup, and applies a special proposition to it: that it contains tea. The special proposition is special in that its entailments are not to be used to form further beliefs, as they would be normally. By contrast, if I had already in reality drunk a cup of tea, I would later be

disposed to assert the proposition 'I have had a hot drink' which I would not be if I had merely pretended to drink a cup of tea.

The central objection of Nichols and Stich (2003) is simulationist in spirit. It challenges the Pretend operator of Leslie (1987) which they argue plausibly is too sophisticated for young children to use. As Nichols and Stich (2003, p. 52) observe, "[f]or Leslie, all episodes of pretence are subserved by representations of the form: I PRETEND 'p.' Thus, while Leslie would agree that an agent can have desires and act on them without having the concept of desire, his theory entails that an agent cannot engage in pretence without having the concept of pretence." Recall again that children as young as four (Wimmer and Perner 1983) are competent ascribing false beliefs; and other evidence suggests that those as young as 15 months are competent as well (Onishi and Baillargeon 2005). Nichols and Stich (2003, p. 51) write that "the pretence could proceed perfectly well even if the subject did not have the concept of pretence." This is the crux of my argument against the account of Leslie (1987) of the relation between pretence and false belief in autism.

So the line of Nichols and Stich (2003) is similar to the simulationist one that S does not need the concept BELIEF in order to have beliefs and requiring the former as well as the latter to explain young children's ToM capacities is to require too much. There are lines on which concepts are innate (Fodor 2008), (Carey 2009) which allow TT(Innate) defenders to hold that 15-month olds have concepts like PRETEND. This in fact is the line taken by TT(Innate) defenders. Scholl and Leslie (1999, p. 147) write that their ToMM "incorporates innate notions/concepts of propositional attitudes such as BELIEF and PRETENCE, and makes them available to a child before general problem-solving resources have fully developed." It is safe to say that the line that concepts are innate is highly controversial.

An alternative formulation of the dialectic here would run as follows. Leslie

(1987) argues that both passing the False Belief Task and engaging in pretence both require a mental operation involving representing a non-true proposition. That mental operation is termed 'decoupling' by Leslie (1987). Therefore they could both be impaired by such an inability to represent a non-true proposition. Nichols and Stich (2003) then argue that pretence does not obviously involve representing a non-true proposition. Leslie (1987) can then say that this is not an objection to his view since he never said it was obvious. Observers taking this line will likely think that Leslie (1987) does have a candidate explanation of the link between pretence and passing the False Belief Task; although Nichols and Stich (2003) show that this candidate explanation relies on a non-obvious premise, this does not mean that the candidate explanation is incorrect or that TT(Innate) cannot explain links between pretence and the False Belief Task. One issue with this line is the difficulty explaining, as is admitted by Leslie (1987), the developmental lag. Two-year-old children can pretend but only four-year-old children can pass the conventional False Belief Task (this might be an occasion where the results of Onishi and Baillargeon (2005) are of assistance to TT, but again, there are theoretical costs associated with taking that nativist line).[22]

ST can explain links between pretend play and ToM. It can simply say that autistic subjects who are less able to supply pretend dialogue for teddy are by the same token less able to imaginatively project themselves into teddy's position or anyone else's. Since that is exactly what ToM requires on the simulationist account, it is unsurprising that autistic subjects exhibit ToM deficits.

[22] A final further problem for TT(Innate) is the lack of supporting data; in the assessment of an experimentalist, there is "presently little specific evidence" for modular accounts (Doherty 2008, p. 5).

3.4 Conclusion

There are severe objections to both TT(Scientific) and TT(Innate). TT proponents in the psychological literature do not escape these difficulties by the expedient of not spelling out to which account they cleave. Since both of the pure accounts have been found wanting, the next step is to consider whether hybrid accounts involving mixtures of capacities can improve the position for TT.

Chapter 4

Objections To Hybrid Accounts

4.1 Introduction

Given the severity of objections set out above to pure TT(Scientific) and pure TT(Innate), TT proponents may attempt to add resources to their position by considering hybrid accounts of two sorts. The hybridity may be within TT itself, or may involve the addition of some simulation capacities as well. The first option involves what I will term Theoretical Hybrid accounts. The idea would be that such a combined account could perhaps avoid some of the severe objections I raised against TT(Scientific) and then TT(Innate) separately. In §4.2, I will examine which of the objections raised against TT(Scientific) and TT(Innate) separately may be avoided in this way. I will conclude that while this does enable some progress to be made, severe objections remain unaddressed and the account has become considerably more unparsimonious – it has many more moving parts, as I specified in the original definition of parsimony on p. 32. These sorts of account should therefore not be preferred to a more parsimonious account of the sort I will later be proposing.

I understand Strong S/T Hybridism and Weak S/T Hybridism as below.

- **Strong S/T Hybridism**: ToM use involves significant application of both theory and simulation.
- **Weak S/T Hybridism**: ToM use involves application of both theory and simulation, but the role of theory is extremely limited.[1]

The Weak S/T Hybridist account is the one I will defend; the Strong Hybrid account is the mainstream one favoured by Saxe and many other commentators. They vary in the amount of significance they allow to either theory or simulation, but all allot major involvement to each. As I will argue below, this raises interaction problems which a Weak S/T Hybridist account avoids.

Since we now have rather a lot of terminology referring to various accounts of ToM, I provide fig. 4.1 below for illustrative purposes. For simplicity, I do not show all ST versions. I also show the one example of a Theoretical Hybrid which is possible with two variants of TT and one example of the several possible S/T Hybrids, which could be either Strong or Weak.

As is by now clear, my sympathies lie very much with ST. Yet there is reason to avoid a *completely* pure ST account. The reason is that it seems clear that there are occasions when S explicitly reasons about the mental states of O and do so in order to predict and explain O's behaviour. There are some occasions when I explicitly ask myself why O has gone into the coffee shop. Perhaps I know that O dislikes coffee and therefore infer that O is meeting someone in the coffee shop and will drink something else. It would be unappealing to deny that this type of reasoning occurs, and it also looks difficult to deny that it is theoretical. Even if it is not theoretical, assimilating it to simulation would require argument. There might be scope to deny that this

[1] One might also term an account which was almost entirely theoretical with minor simulation elements a Weak S/T Hybrid account, but I of course do not favour such an account and neither does anyone else in the literature.

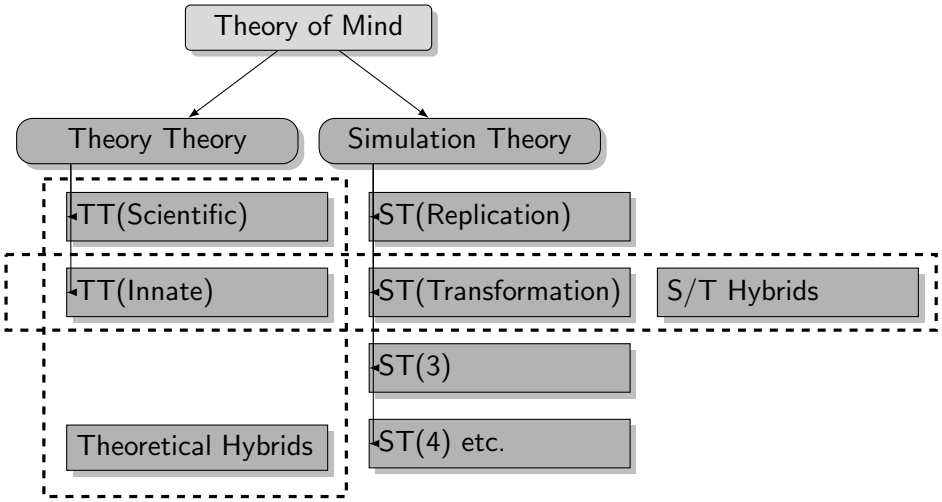

Figure 4.1: Types Of ToM Hybrids

type of activity forms part of ToM. One way to do that might be to stipulate that all ToM use is sub-personal and non-explicit. On reflection, this is also unappealing, being rather unmotivated and also excluding processes from ToM that look very much like ways to predict and explain the behaviour of others. For these reasons, I embrace Weak S/T Hybridism rather than the conceptually adjacent pure ST account. Weak S/T Hybridism allows that explicit reasoning processes of the type outlined above do in fact occur, they do form part of ToM and they are theoretical. This however is the only concession that Weak S/T Hybridism makes to TT; there are no other occasions when theoretical activity is a part of ToM other than these rare occurrences of explicit step-by-step reasoning. This bridgehead falls far short of the amount of theoretical activity envisaged by the mainstream Strong S/T Hybridist account and does not require an account of the interaction between the theoretical and simulation elements. It simply allows that sometimes S uses explicit reasoning purposes as described but does not see any interaction problem between the use of those processes and the vast bulk of ToM activity which is processed by simulation.

We have already seen that TT comes in two major variants: TT(Scientific) and TT(Innate). I will press the three major objections to each of those accounts. Saxe responds to the difficulties raised in the previous chapter for ST by urging the merits of the third option beyond pure TT and pure ST: a Strong S/T Hybrid account of ToM in which both theory and simulation play a major role. Such a Strong S/T Hybridist account is the mainstream view in psychology. Strong S/T Hybridist accounts, since they involve TT, inherit either the three objections to TT(Scientific) or the three objections to TT(Innate). Beyond that, Strong S/T Hybridist accounts are prone to further unique objections, because any successful Strong S/T Hybridist account must also describe how the two parts of ToM are to work together. Weak S/T Hybrid accounts need not do this. I will therefore close this chapter by considering a further three objections unique to Strong S/T Hybrid accounts.

In §4.3, I will cover two objections to the Strong S/T Hybridist accounts favoured by Saxe and others. We see that her position is Strong S/T Hybridism when she writes that "to conclude that a naïve theory of mind, and some capacity to simulate, interact" is a "better option" than the idea that "in some contexts, [S] is a pure simulator, whereas in other contexts [S] uses pure theory" (Saxe 2005a, p. 175).

All successful Strong S/T Hybridist accounts must specify *how* the two elements of ToM are to interact. To provide such specifications, Saxe (2005a) appeals to two Strong S/T Hybridist approaches in the literature, referred to by the terms 'which tool when?' and 'perspective taking.' I will be raising objections to both specifications, primarily on the grounds that setting out in detail how they would work would involve immense complexity. Indeed, as set out it seems that they are grossly underspecified; prospects of finding such a full specification are in my view slim. It will be clear I think from a description of how the interaction is supposed to work in even simple cases that a Strong

S/T Hybridist account of ToM would require a great deal of rule setting and caveating.

In view of all of the objections raised in this chapter and the previous one, I will conclude in §4.4 that the merits of the almost pure ST account I term Weak S/T Hybridism deserve reconsideration, since the array of objections to all of the alternatives appear insuperable.

4.2 Objections To Theoretical Hybrids

Faced with the above objections, TT proponents might aim to respond by considering whether a Theoretical Hybrid account is an improvement. Although the account would be less parsimonious, perhaps TT(Innate) can meet the objections that TT(Scientific) could not, and vice versa, leading to an account which overall handles all the objections. Saxe (2005a, p. 174) gestures tentatively at such a possibility when she writes that a "lay theory of psychology [...] could be constructed (possibly over a scaffold of innate concepts) from observation, inference and instruction." The "observation and inference" is TT(Scientific) while the mention of "possibly innate" concepts points in the direction of TT(Innate). It might be possible to construct an account of concepts and ToM such that concepts are innate but that ToM is not, but it seems more likely that they would both be innate, since at least mental state concepts such as BELIEF and DESIRE are so central to ToM on most views. Saxe's gesture to innate concepts is likely aimed at Carey (2009) who puts forward an account based on innate concepts which can be considered as a Theoretical Hybrid account of ToM.

I will in this section examine each of the six objections to TT(Scientific) and TT(Innate) mentioned above from this new Theoretical Hybrid perspective. I will conclude that three can be met by a Theoretical Hybrid account and three cannot. I will cover the ones that can be met first and the ones that

cannot be met second. It will of course still be the case that the account is less parsimonious than either of the pure TT accounts, since it has more moving parts and their interaction will need to be specified.

4.2.1 Objections Avoided By Theoretical Hybrids

Too Complex And Too Difficult (cf. §3.2.1)

Theoretical Hybrid accounts can reduce the force of this objection by pointing to the innate components of the account. Since these do not need to be learned, it is not a problem that the generalisations in that part of the ToM machinery are complex or difficult. Naturally the account would be committed to quite a heavy weighting of TT(Innate) in the mix in order to reduce the complexity significantly – which is not a problem for the account.

Entails Inexplicable Convergence (cf. §3.2.3)

Theoretical Hybrid accounts can reduce the force of this objection by appealing to some common genetic or other basis to TT(Innate). Proponents of the view could then make the reasonable assumption that the evolutionary environment has selected for fairly similar ToM capacities across cultures. As above, the account would again be committed to quite a heavy weighting of TT(Innate) in order to counter this objection.

Cannot Explain Development (cf. §3.3.1)

Theoretical Hybrid accounts can avoid this objection in one new way. TT(Scientific) is naturally developmental by definition, so adding it to TT(Innate) could in principle make the whole developmental. It would need to be explained how the two elements fit together. Secondly, while there are, as discussed above, some difficulties in providing a mechanism for modules to develop, TT(Innate) proponents can again take the viable albeit

controversial route of Bloom and German (2000) and deny that the False Belief Task measures ToM capacities. That second escape route is available to TT(Innate) alone, so it is not strictly speaking an additional benefit of theoretical hybridity.

4.2.2 Objections Not Avoided By Theoretical Hybrids

Cannot Explain Default Belief Attribution (cf. §3.3.2)

This objection which was raised to TT(Innate) seems to still apply in perhaps slightly less virulent form to TT(Scientific). The particular problem with default belief attribution for TT(Innate) was that it was difficult to square a modular and informationally encapsulated ToM with the apparent need to access the entirety of S's belief set as the default belief set for O. This objection continues to apply to the TT(Innate) component of a Theoretical Hybrid account. While TT(Scientific) is not described as modular and encapsulated by its proponents – which might lead us to ask how coherent it is to harness a modular and non-modular ToM system together – it also does not look highly parsimonious in this connection. S would need to apply a generalisation similar to 'If S believes X then O believes X unless there is reason to think otherwise.' This could scarcely be applied to the entire belief set of S within any reasonable efficiency constraint. So a selection of S's beliefs must be chosen to ascribe to O. They must be the relevant ones for the action prediction at issue, which runs straight into the Frame Problem again.

Requires Solving The Frame Problem (cf. §3.2.2)

Adding TT(Innate) to the TT(Scientific) element of the account does nothing to assist with this objection. Whether a generalisation arises from observation or is innate does not change the fact that it must key off relevant factors only. We might equivalently say that the Frame Problem objection applies equally

to TT(Innate) as to TT(Scientific). Practically speaking, humans solve the Frame Problem somehow but theorists do not know how humans do it and so cannot axiomatise the solution. Suggesting otherwise involves presupposing that there is a definable solution. Also, it is a very strong assumption that there is an innate solution to the Frame Problem, which is a much stronger assumption than that humans can solve the Frame Problem. Humans cannot have an innate theory of a solution that even today no cognitive scientist can specify.

ST accounts sidestep this objection simply because on the ST account, S can just use whatever method S uses in his own case to solve it in the case of O. Similarly, the Weak S/T Hybrid account which I favour is immune to the objection. As I specified on p. 114, the Weak S/T Hybrid account is pure ST with a carve-out to allow for some occasions when S explicitly reasons about the mental states of O in order to predict the behaviour of O. Similarly to the case with the response I briefly sketched in the previous paragraph, the Weak S/T Hybrid account can simply appeal to whatever mechanism allows us to perform deductive reasoning without considering irrelevant factors.

Cannot Parsimoniously Explain Autism (cf. §3.3.3)

As is suggested by the fact that Nichols and Stich (2003) employ this objection in order to favour ST as opposed to TT(Innate), it is unclear how adding TT(Scientific) to TT(Innate) would be of assistance to Theoretical Hybridists seeking to respond to this objection. If the Pretend operator is too sophisticated for young children to have specified innately, is it not even more likely to be too sophisticated for the same young children to have derived scientifically?

4.2.3 Interim Conclusion

I conclude that there remain objections that have not been solved by any version of TT. I will now move on to consider the question as to whether those problems can be dealt with by moving to Strong S/T Hybridist accounts i.e. whether adding simulation into a pure theoretical account is of assistance to TT proponents.

4.3 Objections To Strong S/T Hybrid Accounts

I will in this section raise further objections to Strong S/T Hybridist accounts. This does not imply that I believe that Strong S/T Hybrids are immune to the earlier objections. To take one example, Strong S/T Hybrids will still be in difficulty with the Frame Problem, since adding ST to TT(Innate) or TT(Scientific) does not mean that the generalisations in the TT elements of the ToM machinery can avoid the need to specify a solution to the Frame Problem within those generalisations that remain.

Saxe appeals to two Strong S/T Hybrid accounts. I lack space to consider triple hybrids not cited by Saxe (2005a). Nichols and Stich (2003, p. 60) propose a "highly eclectic account" which "includes processes that fit with [TT(Scientific)], [TT(Innate)], and [ST] as well as processes that do not have any clear parallel." Also, Slaughter and Gopnik (1996, p. 2986) write that "it seems likely that maturation, simulation, and theory formation all contribute to our intuitive psychological understanding." Such accounts it seems to me will be maximally complex and unparsimonious and inherit the bulk of the objections to the separate accounts as well as the interaction problems I am about to outline. Basically, I will object that the accounts are too complex and too unparsimonious; especially when one asks how the interactions between simulation and theory will be prescribed. These factors are exacerbated by the involvement of theory, since it seems that the only way of handling conflict

of generalisations in theoretical systems will be more generalisations. This threatens to result in a requirement for an unlimited number of generalisations and also raises the Frame Problem again at each new meta-level. The first Strong S/T Hybridist account to which Saxe appeals addresses the question of 'which tool when?' and the second deals with 'perspective taking.'[2] I will also be suggesting that ST avoids all of the complexity and lack of parsimony but can still explain the observed ToM performance.

The problem of 'which tool when?' arises in all Strong S/T Hybridist accounts which involve simulation and theory. The question is, when does ToM use involve simulation and when does it involve theory? Saxe cites Ames (2005) in order to make use of its account of 'which tool when?.' My primary objection here is similar to the excessive complexity and lack of parsimony one that I have previously levelled at pure TT accounts. The position in terms of complexity and parsimony is made much worse even than in the case of pure TT accounts because there need to be many more moving parts to handle the interaction. I will be suggesting that no reasonably parsimonious account including such interactions can be given. The problem cannot be avoided by prescribing a Strong Hybrid Non-Interactionist account because there would still have to be extra rules setting out how interaction was to be avoided; presumably by setting out specific areas of sole competence for simulation and theory. The Weak S/T Hybridism I favour must also do this, but can do so simply by adopting a minor carve-out for the explicit reasoning about mental states discussed above. Everything else is simulation.

On perspective taking, Saxe cites Epley et al. (2004). My objection here is once again the excessive complexity and lack of parsimony one, leading as above to difficulties in setting out how the different elements interact. I will suggest that the account is under-specified and replete with caveats. Removing

[2]For other objections and responses, see: Goldman and Sebanz (2005); Saxe (2005d); Gordon (2005); Saxe (2005c); J. P. Mitchell (2005); Saxe (2005b).

those caveats, were that possible, would leave one with a dramatically unparsimonious account of ToM, especially if the removal were done on a TT-basis. In both contexts, I will suggest that pure ST (or rather Weak S/T Hybridism) can handle the situation more straightforwardly. The problem throughout will be that all TT-approaches result in a metaphorical 'explosion of rules' with more and more rules needed to handle the rules. We would have an exponentially complex situation arising from such a need to have rules controlling the application of rules. Indeed, it seems as though there is no limit to the number of rules handling rules (and further rules handling those?) These will need to be specified by the account if it is to succeed in laying out a full and adequate set of generalisations which on the TT or Strong S/T Hybridist views underlie ToM.

4.3.1 Which Tool When?

On the Strong S/T Hybridist account proposed by Ames (2005), both theoretical and simulational activity form part of ToM capacity. Ames recognises four routes to mental state inference. I will set out the routes with a view to giving the account a fair hearing, but also of showing how it is extremely complex even as set out so far; it would not become simpler if extended to increase its explanatory power. Also, we want to know when the four routes are employed, whether they collaborate or compete, and whether such intra-account interaction is prescribed by further generalisations.

The four routes to mental state inference are as set out in Table 4.1. The first two routes fall into the category of 'evidence-based strategies,' which are the theoretical elements of the approach, though the second route appears to have simulational elements. The second two routes are called 'extra-target strategies.' 'Extra-target' means S is to use his own non-evidential resources to infer the mental states of O. The first of these, projection, is simulational.

ToM In Use	Source of Data	Use of Data
1. TT	behaviours in context	attribution of mental states
2. TT/ST	emotional displays	emotion perception
3. ST	S's own mental states	projection
4. TT	stereotypes	stereotyping

Table 4.1: Ames's Four Routes To Mental State Inference

The fourth route appears to be mostly theoretical.

We know that the first 'evidence-based strategy,' Route 1, is theoretical since "perceivers readily work from the visible evidence of human behaviour to posit invisible underlying mental states" (Ames 2005, p. 159). The mental states of O that are posited by S are theoretical entities because they cannot be directly observed. As an example of Route 1, Ames (2005, p. 159) gives "a grabbing hand entails wanting," which is the theoretical generalisation here. Whenever S observes the hand of O making a grabbing motion towards object X, then S is apt to ascribe to O the mental state of wanting X. Ames notes evidence that even six-month old infants seem to apply this generalisation.

The ST account replacing Route 1 would be that when S sees the grabbing hand of O, this may be combined with the simulational output akin to 'when my hand grabs, I want something' to ascribe the wanting to O. This would be another process analogous to that postulated in the Motor Theory of Speech Perception discussed above on p. 34.

The second 'evidence-based strategy,' Route 2, is also theoretical. A sample generalisation is "a person beams when proud of her work" (Ames 2005, p. 160). So when O beams, S uses 'emotional perception' to observe the beaming and interpret it, and S then attributes the emotion of pride to O based on the beaming. We may assume that 'emotional perception' involves simulation,

as outlined by some ST proponents (Gallese and Goldman 1998), (Goldman and Sebanz 2005), absent specification of generalisations keying off descriptions of facial expressions. The simulationist explanation is that S recruits his emotional display production mechanisms in the service of emotional recognition. Since simulation is now a component of Route 2, then Ames's account must provide a specification of whether there is interaction between theory and simulation and how the interaction works within Route 2. I contend that no parsimonious[3] specification thereof can be given, since it would involve generalisations adjudicating the application of generalisations, and a version of the Frame Problem would occur.

The ST account replacing Route 2 would be straightforward pure simulation. S simulates being proud of his work and finds that beaming is sometimes a consequence of such pride. Therefore if S observes O beaming, ascription of pride to O becomes one of the outputs. There are of course other reasons why O may be beaming, and so S may make a simulation error here. However, such an error does not add any traction in the TT vs ST debate since application of the generalisation is equally error-prone. If the generalisation is read as 'O beams if and only if O is proud of his work' then the account would, implausibly, be denying that O could beam for any other reason than pride. If the generalisation is weakened, to read 'sometimes O beams when O is proud of his work,' then it becomes true but loses predictive power. How will this account provide generalisations to handle cases when people are beaming because they are happy, or intoxicated, or enamoured? ST here merely accesses contextually plausible emotions which can result in beaming; these become candidates for ascription to O.

Route 3 is an 'extra-target strategy' involving projection. S "assumes [O] has the same mental states that he or she has or would have" (Ames 2005, p.

[3]Here, lack of parsimony means too many generalisations, which can I think be counted as moving parts.

163). The example Ames (2005, p. 159) gives here is "I'd be embarrassed if I were in your shoes." For example, if S sees O raise his hand in a seminar and then forget the question, S might think that had he done that, he would be embarrassed and therefore ascribes embarrassment to O. This looks as though it has some simulation elements, at least, though below I will raise the question as to whether the processes involved may be both simulation and theoretical. Naturally, the major objection to the entire Ames account continues to be how the simulation described in Route 3 interacts with the theoretical elements within itself, and with those in the first two routes, and what parsimonious description of these interactions can be given.

Route 4 is an 'extra-target strategy' involving stereotyping. This means the rather lazy prediction of the type that when S is using ToM in relation to an O who is Canadian, S assumes that this O "loves playing hockey" (Ames 2005, p. 163). There is no prospect that such generalisations are innate, so they can only be learned. This raises the difficulty that there will be a very large number of such generalisations and very often they will be wrong. So how are they learned? In any case, we can see that stereotyping is not regarded by the mainstream as a component of ToM since, as Ames (2005, p. 163) notes, "stereotyping has been almost entirely ignored" by ToM scholars.

Recall that Ames (2005, p. 159) offers "I'd be embarrassed if I were in your shoes" as a generalisation ascribing embarrassment. The question arises as to whether this ascription is theoretical or simulational. Such an 'embarrassment' generalisation is a good candidate for the sort of simulation-generated regularities that in ST replaces the body of generalisations needed by TT. Imagine that the embarrassing situation in question is raising one's hand to ask a question in a seminar and finding one has forgotten the question. S's will all predict that O's doing this will feel embarrassed. On TT, that prediction arises because S has a theoretical generalisation which states: 'O's who raise

their hand in a seminar to ask a question and then forget what it is will be embarrassed.' On ST, the account is simpler. S merely inputs the pretend belief 'I have just raised my hand to ask a question and forgotten what it is; how do I feel?' and produces the output 'embarrassed.'[4] The fact that the simulation produces the same result every time and could thus merely be described as generalisation-like does not mean that a generalisation was used: to make such a claim would once again set the bar too low for TT proponents. So we might be unable to know whether embarrassment prediction is simulation or theoretical on any given instance of prediction, because the observable inputs and outputs might be identical in different instances even though different mixes of simulation and theory were used. This makes providing a specification of interaction difficult for Ames, or at least makes it difficult for empirical results to shore up such a specification by showing whether simulation or theory was used.

Non-interactionist accounts like Weak S/T Hybridism avoid this difficulty by avoiding any need to specify how simulation and theory work together. I agree that when S explicitly runs through a sequence of arguments like "O wants X, O believes that O will get X if O does Y, O will do Y," this cannot be simulation. The reason though that such an account does not need to specify interactions is that this explicit standalone theoretical reasoning is as it were isolated and epiphenomenal. It does not communicate with any simulation outputs; if there are any produced subconsciously on the same question, they are superseded. So since there is no interaction, the account does not need to specify any interaction.

[4]It might be objected here that surely one has to actually ascribe the embarrassment and not merely feel it, which makes the ST account less simple. Firstly, as I mentioned above, at least the ST account of Gordon (1995a) does not really have to handle mental state ascriptions since it remains on the level of action prediction. Secondly, is it really a significant additional complexity to target the output on O?

The question arises here as to whether there is an implicit parallel of this explicit reasoning. If there is, the further question is whether this carve-out should extend to such an implicit parallel of the explicit reasoning. The argument for this is that there are many cases where it seems that explicit sequences of reasoning have implicit parallels. This could look something like arriving instantaneously at 'Socrates is mortal' without having any staged phenomenology representing each premise in the famous syllogism. The argument against is dialectical; I wish my account to be plausible without conceding so much ground to theory that interactions must be specified or worse, that my account becomes a Strong S/T Hybrid. I think the best way to thread this needle is as follows. My account can claim that the existence of implicit reasoning parallel to the explicit reasoning is consistent with simulation; this theoretical reasoning is not playing an explanatory role. So Weak S/T Hybridism claims that simulational capacities alone provide a sufficient understanding of other minds and behaviour predictions. The nature of the body of knowledge employed in ToM is decisive in the question as to whether ToM reasoning is theoretical or simulational. Weak S/T Hybridism denies that that body of knowledge is theoretical; the 'body of knowledge' is simulational even if there is theoretical reasoning going on. (Under ST, one might say that there is no 'body of knowledge' since the answers to ToM questions are generated on the fly.) Simulation always runs in the background so if there is in additional theoretical reasoning, it is epiphenomenal; it is not part of the answer as to how S arrives at ToM predictions in general.

Can the emotions be formalised? Consider again the generalisation to the effect that 'O's who blush are embarrassed.' Under TT, an implicit formalisation of the emotions and clear statements of when they are in play would be needed, in order to call the right generalisations into operation. However, problems for such an approach include, for example, serious difficulties in dis-

tinguishing shame from embarrassment (Zahavi 2010) and identifying what embarrassment is and when it occurs (Purshouse 2001). ST avoids this type of problem since S's know when they are likely to be embarrassed or ashamed even if they cannot give formal definitions of those emotions or state precisely in which circumstances they will likely be felt and which not. The vague delineation of S's dispositional emotional landscape is paralleled by that of O, giving S some chance of being able to make a prediction of O's emotions.

Further complexity arises for Ames in terms of the interaction of mental state ascription and affect ascription, and the time profile of this interaction. Ames notes that we may forgive someone who has, for example, spilt wine on a white carpet, if they exhibit appropriate remorse. The generalisation is "[a]ffect qualifies behaviour in the near term: perceived remorseful affect can lead to ascriptions of good intent to harm-doers in the short run, but repeated harm drives long run ascriptions of bad intent" (Ames 2005, p. 162). This makes clear the difficulties of providing simple generalisations, but the very name Ames gives his generalisation is telling: he calls it a 'contingency.' Now, a contingent event is one which is not certain but perhaps probable, and in this sense of contingency, Ames means to refer to something aiming to provide for the contingent event, should it occur. This again brings out the complexity of generalisations problem for TT discussed above. Further complexity may be seen in the description offered by Ames (2005, p. 162) of the inputs: they include "behaviour [and] arcs of behaviour over time and across situations" while "affective displays may augment or discount behaviours" and that last factor also has a changing profile of effects over time. The ST account here is rather more simple: S poses the question 'what were my intentions?' under the circumstances of having acted as O has. S can then ascribe those intentions to O and assess whether O is to be forgiven or not.

Another problem for generalisation-based accounts of ToM will be deciding

which generalisation to apply to whom without succumbing to circularity. Ames (2005, p. 159) gives a further example of stereotyping: "Jocks hate romantic comedies." We now have the question though: 'who is a jock?' i.e. to which O should S apply the generalisation in order to predict that O does not like romantic comedies? The answer had better not be 'anyone who does not like romantic comedies is a jock' or the generalisation has become circular. But there are similar risks involved in the other candidate characteristics. The chain 'jocks like beer;' 'who is a jock?;' 'everyone who likes beer is a jock' is vacuous and so are all the other candidate characteristics. These problems are familiar in philosophy in the form of 'how do we identify what falls under a concept?,' to which question none of the available candidate answers —including 'concepts as prototypes' or stereotypes —seems workable.[5] This type of problem is avoided by ST because it does not have any generalisations.

A similarity generalisation is proposed by Ames to decide whether a projection (Route 3) or stereotyping (Route 4) approach will be used in a particular use of ToM. Ames (2005, p. 160) writes that "perceptions of general similarity guide a trade-off between projection (ascribing [S's] own beliefs and desires to O's) and stereotyping." So the idea is that if S thinks O is like S, then S will simply simulate O on the model of S. That approach will not be used if S perceives a gross dissimilarity between S and O. If, for example, S perceives O as a jock and S himself as not a jock, S will use stereotyping to predict the behaviour of O. The similarity generalisation is again described as a second 'contingency' (Ames 2005, p. 164) so we have further multiple branches of conditionality. Ames (2005, p. 165) summarises the research as showing that often, "projection and stereotyping function as alternative strategies that displace each other" which brings out a question deriving from Saxe's citation of

[5]There is an enormous philosophical literature on concepts and what falls under them, which is a measure of the difficulty of the problem. See Wittgenstein (2001), Fodor (1994), Fodor and Lepore (1996), Crane (2003).

this paper as an illustration of a purportedly superior Strong S/T Hybridist theory. The superiority might come from the two elements working together —i.e. at the same time on the same question —to provide a superior answer to a ToM question, where a superior answer is presumably a more accurate one. The other possibility though is that the interaction takes the form of a division of labour. Some questions may be best answered by simulation and others may be best answered by theory. The superiority may devolve from an apposite selection of methods rather than the application of both. It may be a stretch to call this a Strong S/T Hybridist Interactionist theory, but it still seems possible. However —what decides which approach is used if S has both routes open?

Handling this division is complex; all four possible routes to mental state inferences proposed in Ames's diagram must be accommodated coherently in the interactionist account of ToM it proposed is to be successful. Ames (2005, p. 166) offers a third contingency to do this, as follows. "Cumulative behavioural evidence supersedes extra-target strategies: projection and stereotyping will drive mindreading when behavioural evidence is ambiguous, but as apparent evidence accumulates, inductive judgements will dominate." From our perspective, this means that before there is a sufficient weight of behavioural evidence, some mix of projection and stereotyping —i.e. some mix of simulation and theory —will prevail, but after that, the behavioural evidence will prevail. So we have a complex time development of interactions to handle as well.

We also need to know what mix of simulation/projection and stereotyping/theory operates in the initial stage. Empirically, it seems that stereotyping is "a default or initial stage of judgement" (Ames 2005, p. 166). This means that S will make stereotypical predictions about O until S has sufficient observations of O to make a less stereotyped prediction. Other accounts though,

take the opposite line. Ames (2005, p. 166) mentions one view on which "[w]hen the responses of [O's] are not known, [S's] project their own as a first bet." That account leads on to the Epley et al. (2004) perspective-taking account to be discussed in the next section, where S predicts O's behaviour by using S's own projected behaviour as a starting point to be adjusted for O to produce the prediction. That then looks like a simulation starting point with a theoretical adjustment. Ames (2005, p. 166) cites research intended to show that "time pressure may reduce these adjustments while accuracy incentives may increase them." This gives us a view of how simulation becomes more prominent in the mix of approaches used in ToM as time goes on. It looks like we have a simulation/theory mix with time pressure increasing the amount of simulation to be expected, or perhaps the probability that the prediction made by ToM will reflect the quick, simulation answer, while accuracy incentives will decrease the probability or weighting of simulation in the final answer. We will now want to know whether this works for some stereotypes or all of them, and whether an S who has overcome stereotyping in relation to one group of O's will be more likely to do so in respect of other O's for the ascription of the same and different mental states. This account is highly complex; the complexity is further increased by Ames's lengthy and diverse list of items that may boost or inhibit consideration of behavioural evidence viz. "interaction goals [...] self-relevancy [...] cognitive load [...] time constraints [...] social power" (Ames 2005, p. 168). Do these items also interact with each other? We might further think that stereotypes themselves can evolve, which would also need to be described by the account.

In conclusion, the Ames account of 'which tool when?' is extremely complex and specifying generalisations for its application will invariably result in a highly non-parsimonious account of ToM. The generalisations needed to specify the interaction between the various elements will add further complex-

ity. ST avoids all of this complexity and thus also avoids the need for any interactions to be specified.

4.3.2 Perspective Taking

I will again be suggesting that the generalisations-based approach involved in Strong S/T Hybridist accounts which include TT means the account of ToM succumbs to excessive complexity and lack of parsimony leading to interaction problems, and that simulational accounts explain the data more simply. Here I am using the original definition of parsimony given on p. 32, whereby a parsimonious account combines few moving parts with a significant amount of explanatory power. On this parameter, I will suggest that Weak S/T Hybridist accounts perform better than Strong S/T Hybridist accounts.[6]

Epley et al. (2004, p. 328) argue for an 'anchoring and adjustment' ToM paradigm which "simplifies the complicated assessment of another's perspective by substituting one's own perception and adjusting as needed." The first element, the substitution, is simulation because S uses S's own perspective as his starting point for predicting O's perspective. The adjustment is then added by theoretical means on this interactionist Strong S/T Hybridist account. This model "is therefore most likely to be engaged when one's own perspective is readily accessible but another's perspective must be inferred" (Epley et al. 2004, p. 328). On this account, the model is not invariably engaged, and thus we are entitled to ask when and why on the occasions when it is. Are there generalisations to specify when it is engaged and when not?

Epley et al. (2004) investigated understanding of ambiguous messages which could be interpreted as sarcastic or not. For example, one message

[6] While I also believe that this means that Strong S/T Hybridist accounts are more complex *overall*, and also that the Strong S/T Hybridist accounts is *excessively* complex, I do not here provide argument for these stronger claims. I concede that they do not follow from the earlier claim about Strong S/T Hybridist accounts being more complex *in certain respects*.

about a comedian was that "you have to see him yourself to believe how hilarious he really is" (Epley et al. 2004, p. 329). The variable that was adjusted was a description of the comedy show that was either positive or negative. S's were then asked to predict whether O's would understand the message as being sarcastic depending both on whether S had the positive or negative description and whether O did. Epley et al. (2004, p. 329) found that "people adopt others' perspectives by adjusting from their own." This is consistent with a simulational start point —'own perspective' —and a theoretical adjustment —O had a different description of the event than S did. However, it is also consistent with an entirely simulational account: 'what would I think if I had a different description of the event?'

Another type of adjustment investigated by Epley et al. (2004) is where people shift their estimates of the percentages of their peers who will hear something unclear when they themselves know the 'right' answer. For example, there are claims that certain songs contain secret messages when played backwards. The lyrics of a song sound meaningless backwards until one is told what the hidden message is supposed to be, whereupon that 'hidden message' becomes obvious. The results were that 88% of informed participants believed that they themselves heard the message while 0% of uninformed participants believed that they themselves heard the message. Epley et al. also expected "informed participants to estimate that a higher percentage of their peers would hear the phrase than participants who were uninformed" (Epley et al. 2004, p. 334) and this is indeed what was found. So the anchor here is whether S gets the message, which itself is basically controlled by whether S has been told the content of the message. S is then asked to estimate how many O's will get the message, and does this by starting from whether S did as an anchor. This is in essence a simulational account.

Remarkably, people agree with propositions more if they are nodding their

heads when they say them. Epley et al. (2004, p. 334) use this result to hypothesise that S's "who were nodding their heads should be more egocentric and give more extreme responses than participants who were shaking their heads." This hypothesis was confirmed; the nodders gave responses that if true would have had O more like S than was the case with controls or the shakers. "Extreme" here means that the nodding S's became more egocentric about the O's and the shakers less egocentric. It does not appear that TT can explain these results, because nothing about head movements should influence theory use. ST can offer an explanation which is once more based on claims analogous to those made by the Motor Theory of Speech Perception discussed above on p. 34. The central idea there is that motor capacities for speech production are also used in simulation mode to comprehend speech. Simulationists can argue that head nodding influences simulation, since head nodding is what S does when S favours a view. The original strange point that S will agree with a proposition more if S is nodding comes back into play, with the proposition in question being 'O has 'got the message.' ' Therefore if S is making a prediction about O based on S and S is currently nodding, the simulation process will start by modelling O's level of assent as adjusted upwards by S's nodding. It 'looks to S' more like O is favouring a view, or in this case, 'getting the message.'

Accuracy increases over time; Epley et al. find that the amount of adjustment increases from the egocentric anchor if time is not an issue while egocentric errors increase for harried S's. This might be explained as meaning that extra time is available to apply theoretical adjustments thus improving accuracy. However, part of the explanation for this effect may also be found from a simulationist perspective, since there is a very clear route elsewhere wherein additional simulation may improve accuracy. There is a technique known in mathematics and physics as Monte Carlo simulation. The idea is

to run many simulations with slightly shifted input conditions and consider the results of all of them. This provides a better estimation of the outcome when the exact initial conditions are unknown.[7] We may assume that multiple simulations will also allow S to make better ToM judgements if the time is available to perform them; also many simulations could be run in parallel. There is thus no need to retreat to a Strong S/T Hybridist theory, as Goldman (2006, p. 184) does when faced with this question. It is no objection here to say that we do not have phenomenology consistent with running multiple simulations, since we also do not have phenomenology consistent with running a single simulation or using a theory. Although Monte Carlo simulations for physics purposes use theoretical input, they need not do so. Running multiple simulations can explain why S becomes more accurate over time, though perhaps not why S always starts with the same egocentric bias. The explanation for that may rely on the simulationist case more broadly.

The conclusion of Epley et al. (2004, p. 338) is replete with caveats: "individuals' attempts at perspective taking are often something of an integration of theory and simulation. Adults' use of their own perspective as an anchor is similar to using one's self as a source model for predicting others. Additionally, adults' adjustment from that anchor is likely guided by their theories about how different perspectives and psychological states influence judgement and perception." Perspective taking is "often" "something of an integration" i.e. not always; we are not told what "something" means and we do not know whether to interpret 'integration' as more like 'summation' or 'selection.' The anchoring is "similar" to simulation. Adjustment is "likely" guided by theory. We are entitled to ask what evidence supports all of these hedges, what they are intended to carve out, and why, if not to explain inconvenient data. The ST perspective can naturally accept the anchoring side wholesale, so whether

[7] For an example of repeated Monte Carlo simulation being used to produce more accurate predictions, see Short (1992).

the adjustment process must be theoretical is a crucial point. There seems to be no reason at all why it could not be further application of simulation, but with shifted inputs: what would O believe if O was missing facts known to S is a different simulation that allows S to take O's perspective.

All of the theoretical elements in the Strong S/T Hybridist account here investigated depend on the adjustments in, for example, the lyric perception task, which can be equally well or better explained by additional simulation. In addition, the Epley et al. account is scarcely less complex than the foregoing Ames account. Finally, there is a recursive problem of the same nature as the under-specified nature of the interaction in a single account. Now we have two accounts: are the mechanisms they employ also to interact?

4.4 Conclusion

The options are pure TT, pure ST and Hybrids. I contend that the arguments presented in this chapter and the previous one provide strong motivation for a new examination of an option close to pure ST: Weak S/T Hybridism. That view appears more plausible than Strong S/T Hybridism or either of the totally pure accounts. Both pure TT accounts face at least three severe objections each, as I outlined in §3.2 and §3.3. These objections are not resolved by Theoretical Hybrid accounts as I showed in §4.2. Strong S/T Hybridist accounts suffer from all of these problems, since they include a major theoretical component, with moreover the additional complexity problems outlined in §4.3.

All Strong S/T Hybridist accounts face severe dialectical challenges. The Strong S/T Hybridist line seems to be forced on TT proponents by hard cases brought by the ST side. Apart from the obvious lack of parsimony, the claim would presumably be that while their preferred ST or TT account does the bulk of the work, some admixture of the other account must be admitted for

some questions. Alternatively, some questions may lie entirely in the domain of the other theory. This then entitles us to ask how much of the work is ascribed to the other theory, and what that claim even means. Commentators are here forced into vagueness. For example, Goldman (1993, p. 107) writes that he will "make no blanket rejection of 'theoretical' inference in self- or other-ascription. I just doubt that that's where all the action is, or even most of it."

We can understand what it would mean for less than all of of the action to be in simulation or theory. That is not much more than a restatement of the Strong S/T Hybridist position. However, we may legitimately require the Strong S/T Hybridist to say more about the mix. Goldman thinks that 'most of the action' is simulation. Does that mean that 80% of ToM activity is in simulation, and how would such a calculation be made? It might be done by dividing the number of questions resolved by simulation by the total, or the number of propositions, or the occasions of use. All of that would be complicated by any occasions of interactionist ToM use. Bach (2011, p. 28) describes the positions of the hybrid[8] theorists as involving the following calculation: "[i]f the majority of tasks are given to simulation, then simulation is termed the 'default' process (Goldman), and if the majority is given to theory, then theory is the default process (Nichols and Stich)." This seems unhelpful since not only is the question unanswerable, but it is not clear what non-circular value has been added by declaring one other of TT or ST the default process. We might allow commentators to define either ST or TT the default process while restricting that claim to meaning simply that one or the other is more frequently used, but that weaker claim is not very illuminating. Of course, Goldman's position is consistent with the Weak S/T Hybrid account for which I will argue as well as the Strong S/T Hybrid account which I oppose.

[8]He does not call them 'Strong S/T Hybridists' since that is my term; the definition nevertheless fits the theorists to whom he refers.

My criticism here is more of the vagueness which allows the view to extend over two very different accounts, one of which is I have argued is untenable.

One might ask about the possible view that says that both ST and TT mechanisms exist and we know relatively little about when each is used or how they interact. It might be thought that this is a view that says less but faces fewer objections. I accept this, but still insist that this is a less parsimonious account than Weak S/T Hybridist ones. It does not really reduce its number of moving parts by simply stating that the interaction cannot be explained, which is then not a positive move from the perspective of parsimony on the original p. 32 definition. It also fares badly from the explanatory power perspective, because it does not attempt to explain the interactions, which are a fundamental aspect of how ToM predictions are made if this view is correct.

One possible alternative to Saxe's Strong S/T Hybridist Interactionism would be Strong S/T Hybridist Anti-Interactionism. This position asserts that there are two major elements of ToM, theory and simulation, but denies that there is any interaction between the elements. They do not communicate with each other, or use each other's outputs as inputs. There is in addition no third master system combining the two theoretical and simulational systems. Such an Anti-Interactionist account would need to describe why it might come about that there is no interaction in order to be plausible. One option might be to specify separate domains of application. Some questions in ToM might always be resolved theoretically, and other questions might always be resolved simulationally. Or, particular questions might generally be solved theoretically and sometimes simulationally. Providing no episode of consideration of a question involved both simulation and theory at the same time, that would still count as an Anti-Interactionist account. There might even be ways of having a particular question considered both theoretically and simulationally

on a given occasion, but still qualifying as a non-interactionist account. But all of these options appear irremediably complex.

There is also the question as to where the charge of ad hoc domain specification may best be laid. Saxe (2005a, p. 177) claims that historically, "proposals for when [S's] use simulation tend to be somewhat ad hoc." The problem with this charge for Saxe derives from the fact that Saxe is a Strong S/T Hybridist, accepting a role for ST. Therefore her criticism about the ad hoc nature of the domain of application of simulation applies with equal force to her position. Indeed, it is even more virulent, because Saxe has not only the ad hoc domain for simulation, but additional ad hoc domains for theory and then for the interaction region where simulation and theory interact.

Chapter 5

The Systematic Error Challenge

5.1 Introduction

In this chapter I will set out the major challenge to ST as urged by Saxe (2005a).[1] Naturally any challenge to ST is ipso facto a challenge to the Weak S/T Hybridist view which I support. The challenge is that ST cannot account for the systematic errors observed when people perform ToM tasks. These errors exist; they are widely reproduced in the psychological literature. This in itself is not a problem for ST, because it can, as Saxe allows and ST proponents have proposed, avail itself of the Wrong Inputs Defence. I will discuss that defence more fully in Ch. 6, but in sum the Wrong Inputs Defence observes that it is not an objection to ST that a simulation is wrong when the inputs to the simulation were wrong. A wrong input could consist in a false belief held by S about the beliefs or desires held by O. Alternatively, it could consist

[1] She in fact brings two challenges to ST; I will not discuss the second one which relates to differential time development in children's ToM. Saxe (2005a, p. 176) notes that although children have their own beliefs and desires "all along," they can ascribe different desires to others a year before they can ascribe different beliefs.

in a false belief held by S about the factual environment around O which is of relevance to whatever behaviour of O's S is seeking to predict or explain. However, Saxe's particular challenge relates not to the errors but to their systematic nature. This means that all or a large proportion of S's will make the same mistaken predictions about O's behaviour in a particular context.

ST does not, according to its opponents, predict that ToM errors will be systematic. In fact, ST should predict a random spread of errors, according to TT proponents. I will outline the TT argument for this here by sketching the Two Colours task (Ruffman 1996), cited in support of TT (Saxe 2005a, p. 175). The task involves a child who sees a green bead being moved from a dish containing red and green beads into an opaque bag. An observer A is behind a screen so that A sees that a bead has been removed from the dish but not its colour. Also behind the screen —and thus visible to the child but not to A —is another dish containing yellow beads. The critical question is asked of the child 'what colour does A think the bead in the bag is?.' If the child is simulating, it should place itself in imagination behind the screen and conclude that it cannot see the colour of the bead. So it will give answers randomly spread across red and green (or conceivably also yellow, since there are yellow beads in the other dish that only the child can see) because the child has no reason 'from behind the screen' to pick one colour over another. This is not what is observed, as we will see: the child in fact will mostly say that A thinks the bead is red.[2]

TT, by contrast, has a ready explanation for the systematic nature of the errors. It can postulate a single item of theory, a generalisation, that is wrong. If everyone has the same incorrect generalisation, then everyone will make the same mistaken ToM prediction in all circumstances that activate that generalisation. Thus, Saxe can argue that the systematic ToM errors

[2]I will not offer in this thesis a specific ST defence against the Ruffman data; see Short (2015, Ch. 9) for such a defence.

that are observed are good evidence for TT and against ST.

Saxe (2005a) introduces a large variety of experimental evidence to support her claim. The evidence is broken down into several classes. I will consider two classes of data in this thesis. Later in this thesis, I will provide a chapter responding to each class of data on behalf of ST, but for now my task is solely to set out the problem. The two classes of data show the following types of systematic ToM error.

- In some experiments, ToM is systematically too 'rosy': S's are unwarrantedly optimistic about the rationality and logic employed in O's decision-making.
- In some experiments, ToM is systematically too 'cynical': S's are unwarrantedly pessimistic about the rationality and logic employed in O's decision-making.

In the next two sections of this chapter, I will introduce each class of data: first the too rosy data (§5.2) and then the too cynical data (§5.3). For now, I will only give one of the experiments cited by Saxe (2005a) in each class as an example. In later chapters, I will retain the separation into two classes but consider many more experiments cited by Saxe within each class.

5.2 The 'Too Rosy' Challenge

The class of 'too rosy' data supporting Saxe's systematic error challenge is introduced by her as below.

"Adults, too, have systematically inaccurate and over-simplified beliefs about beliefs that are often self-flattering. 'We are convinced of the rationality of [human] reasoning, highly adept at constructing plausible explanations for our decision behaviour, [...] and so on' (Evans 1990, p. 109). That is, we

share the conviction that, in general, beliefs follow from relatively dispassionate assessment of facts-of-the-matter and logical reasoning. As a consequence, people's expectations of how they and others should reason and behave correspond more closely to normative theories of logic, probability and utility, than to their actual subsequent behaviour (Gilovich 1993).

Historically, proposals for when observers use simulation tend to be somewhat ad hoc. In fact, if we could accurately simulate other minds, half a century of social psychology would lose much of its power to shock and thrill. The charisma of many famous experiments in social psychology and decision-making derives from the fact that they challenge our too-rosy theories of mind (Ross, Amabile, and Steinmetz 1977). The experiments of Milgram (1963), and Asch (1952), and Tversky and Kahneman (1974), are famous because there is a specific, and vivid, mismatch between what we confidently expect, and what the subjects actually do" (Saxe 2005a, p. 176).

As one example of Saxe's too rosy data, I will consider the Milgram experiment. I will cover many more in the chapter devoted to explaining this class of data, Ch. 8. This famous experiment involves some deception of the experimental subjects, which means that it has not been widely replicated, because it would not pass modern university ethics panels. It was conducted at Yale in 1961. The context continues to be that of the aftermath of World War Two, and the preliminaries to the experimental writeup mention that the Nazi regime is an explicit concern. How will ordinary people respond when asked to perform extraordinary acts that are beyond what they would claim are their moral limits? Should we understand the Nazi phenomenon as an aberration, or will ordinary people be generally be susceptible to persuasion beyond expectations when placed in extraordinary circumstances?

There are three protagonists in the Milgram (1963) experiment: the experimenter, the actual subject and the 'dummy subject.' The actual subject is an

innocent member of the public. The actual subject believes that the dummy subject is also an innocent member of the public, but this is not the case. In fact, the dummy subject is a collaborator of the experimenter. An apparently random but in fact rigged selection is run to allocate roles between the actual subject and the dummy subject. The two roles are 'learner' and 'teacher' in a word pair learning test. The selection is rigged such that the actual subject is always the teacher, and the dummy subject is always the learner.

The experimenter explains to the actual subject that the experiment is an investigation of how learning may be improved by mild punishment of error. In the standard version of the experiment, the dummy subject is placed in a different room to the actual subject and communicates the word pairs via a panel. Their performance is to be assessed by the actual subject, who is also tasked with applying punishment to them if they make a mistake. The situation is rigged so that the dummy subjects do in fact make many mistakes. The actual subject is now told to apply an electric shock to the dummy subject. They have a range of electric shocks available to apply, beginning from mild and increasing in voltage. In reality of course, no electric shocks are applied at all to the dummy subject. However, they do react as if they were being applied. The intensity of their reaction increases dramatically as the purported shock level increases. Bear in mind that since the dummy subjects are in a different room, their behaviour under the apparent shocks is not seen by the actual subjects. There is no verbal response from the dummy subjects, though the dummy subjects make audible sounds of protest as the experiment proceeds. At extreme levels in fact, they cease to respond to the requests for a new word pair, and "[w]hen the 300-volt shock is administered, the learner pounds on the wall of the room in which he is bound to the electric chair" (Milgram 1963, p. 374).

The actual subjects believe they are administering shocks ranging from

'Moderate' through 'Intense' to 'Danger: Severe Shock' and beyond to the mysterious 'XXX' category. If the dummy subject protests that this treatment is unreasonable or unethical or for any reason resists applying the shock, the experimenter encourages them. A fixed scale of four experimenter responses is set as actual subject resistance increases along with dummy subject distress. These "prods" were in order as follows: "[p]lease continue, or [p]lease go on; [t]he experiment requires that you continue; [i]t is absolutely essential that you continue; [y]ou have no other choice, you must go on" (Milgram 1963, p. 374). We immediately believe here that no-one will comply.

The surprising results though were that: "[o]f the 40 [O's], 26 obeyed the orders of the experimenter to the end, proceeding to punish the victim until they reached the most potent shock available on the shock generator" (Milgram 1963, p. 376). At this juncture, Saxe already has her point: we are amazed that any of the dummy subjects will go this far, and we are confident that we ourselves would not.

Crucially for Saxe's view though, there are a fourth group of players, who will provide us with evidence of systematic failure of ToM of the too rosy sort and indeed with hard numerical evidence thereof. Milgram later provides a group of psychology undergraduates with a description of the set-up. Milgram (1963, p. 375) writes: "[f]ourteen Yale seniors, all psychology majors, were provided with a detailed description of the experimental situation. They were asked to reflect carefully on it, and to predict the behaviour of 100 hypothetical subjects. [...] All respondents predicted that only an insignificant minority would go through to the end of the shock series. (The estimates ranged from 0 to 3%; i.e., the most 'pessimistic' member of the class predicted that of 100 persons, 3 would continue through to the most potent shock available on the shock generator —450 volts.)" This provides Saxe with a valuable data point. In the actual experiment, $26/40 = 65\%$ of O's set the dial to 450 Volts while

the psychology undergraduate S's estimated that that number would be 3% at most.

Since there are now four groups of protagonists in the experiment, there is room for confusion when we view it in our ToM framework. Recall that S is the subject in our terms who is using ToM to predict the behaviour of the object of ToM O. In this framework, the S's are the psychology majors who predicted the behaviour of the actual subjects or teachers, the O's. So the discrepancy between 3% and 65% represents the systematically too rosy ToM error which Saxe requires.

5.3 The 'Too Cynical' Challenge

Saxe (2005a) also cites a class of experimental data that tend in the opposite direction to those discussed in the previous section. While her challenge continues to be that there are systematic errors in ToM, the direction of those errors is opposite under different circumstances, and systematically so. As previously, defenders of ST must explain this directionality of error as well as the mere possibility of error. Once again, Saxe will appeal to a wrong theoretical generalisation being applied in the various cases, which gives TT an easy response to the data.

Saxe (2005a) introduces this class of supporting data as below.

"[L]ay epistemology is not universally charitable. Most adults believe that beliefs are sometimes false, that reasoning can sometimes be distorted —both inevitably, by the limitations of the mind, and wilfully, as in wishful thinking and self-deception —and that all of these are more likely to be true of other people's thinking than of their own (Pronin, Puccio, and Ross 2002, pp. 636-665). As a consequence, [S's] sometimes overestimate the prevalence of self-serving reasoning in [O's] (Kruger and Gilovich 1999), (Nisbett and Bellows 1977), (Miller and Ratner 1998).

In one study, Kruger and Gilovich (1999) asked each member of a married couple, separately, to rate how often he or she was responsible for common desirable and undesirable events in the marriage. Then, each was asked to predict how their spouse would assign responsibility on the same scale. Although everyone actually tended to take credit equally for good and bad events, each predicted that their spouse would be self-serving, that is, take more responsibility for good events, and less responsibility for bad ones. [...] Thus whereas reasoning about reasoning is usually characterised by overly optimistic expectations about people's rationality, in specific circumstances (e.g. the culturally acknowledged self-serving bias) observers are overly pessimistic, an effect dubbed 'naïve cynicism' [Kruger and Gilovich (1999)]" (Saxe 2005a, p. 177).

Note that there is a possible confusion in the last sentence. There are two 'self-serving biases' at play in this experiment. There is the self-serving bias(O) of O which would involve O making unrealistically positive claims about himself. The second self-serving bias(S) would be in S, where S predicts even more self-serving bias(O) in O than O exhibits. The self-serving bias(S) in S thus paradoxically allows S to predict that S is less self-serving than O and thus more virtuous. It is important to keep these different biases separate.

As before, I will provide here just one example of the sort of experimental data Saxe (2005a) appeals to in this class of too cynical data, while covering many more of her examples in my detailed response on behalf of ST in Ch. 9. Here, I will just expand on the marriage partners example.

Kruger and Gilovich (1999, p. 745) had married couples fill out a questionnaire about joint activities of either negative or positive relationship value. Here, 'joint activity' means something that either partner might do, not something that they necessarily both do together. For example, a negative activity would be "taking out frustrations on partner" while a positive one would be "resolving conflicts that occur between the two of you." Each partner was

asked to allocate responsibility for such activities by percentage between themselves and their partner. The idea was that the partners should think on a frequency basis. Imagine that there were 20 occurrences of an activity falling under the given description "taking out frustrations [...]" in the last month. So the total number of such occurrences for which the husband was responsible plus the total number of such occurrences for which the wife was responsible sum to 20. The same pattern should be visible across the board, with 100% of responsibility being allocated across partners and across activities.

The investigation of whether these allocations are biassed proceeds by comparing what partners say about themselves and comparing it with what their partners said about themselves on each task. This can then be compared with 100%. If the husband is responsible for 60% of a particular activity, then his wife can claim up to 40% of initiations of this activity for herself, and no bias has been measured. If however the total is more than 100%, then both partners have claimed more responsibility than is actually available and a positive bias has been measured in relation to that activity. Both parties want to claim credit for that activity. On the other hand, if the husband admits to only 30% of responsibility for a given action, and the wife also admits to only 30%, then a negative bias has been observed. Neither party wants to admit responsibility for that activity. As the authors write, "suppose a wife believes she initiates 60% of the discussions about the relationship and her husband believes he is responsible for 50%. Together, they have assigned 110% of the activity to themselves, yielding a bias score of + 10%" (Kruger and Gilovich 1999, p. 745). Initiating discussions about the relationship was a positive activity in the experimenters' paradigm.

By allocating responsibility to himself, the husband naturally also allocates the inverse responsibility to his wife. If he thinks he does 70% of the "spending time on appearance to please the other" (Kruger and Gilovich 1999, p. 745),

then he must also think his wife does 30%. The activities considered in the experiment were such that no-one else could do them other than the two spouses. Since the experimenters have the questionnaires from both spouses, they are now in a position to compare the data, and to cross-reference it with whether the activity is positive for the relationship or negative. But they took a crucial further step in this experiment, which is why Saxe (2005a) cites this particular experiment. Kruger and Gilovich (1999, p. 745) also asked each partner what they thought the other would say. Note that this allows for a sum greater than 100%. If the husband thinks that he does 70% of "spending time on appearance to please the other," he can consistently also think his wife will claim 70%, while he believes she actually does 30%. The husband can have a biased expectation of bias. Matching that in the other direction, the husband can think that he causes 30% of the arguments, and that his wife therefore actually causes 70% of the arguments, but that she will only admit to causing 30%. If this is the case, then there is a systematic error in ToM in a too cynical direction and Saxe has her data.

This is exactly what is observed; Kruger and Gilovich (1999, p. 745) report that "couples expected their spouses to claim more than their share of the credit for the desirable activities (M = +9.1%) —but less than their share of the blame for the undesirable activities (M = -16.1%)," where 'M' stands for mean bias. The systematic error in ToM here is then 'biased expectations of bias.' The S's expect their partner O's to be biased. The O's are indeed biased. But they are less biased than the S's predict; the quantum of how self-serving they are is less than predicted. Saxe has indeed provided data which help her in two ways. There is indeed a systematic error in ToM in that S's generally all make the same error. But secondly, these ToM errors are all in the too cynical direction when the previous ToM errors were all in the too rosy direction. Saxe may now demand that ST proponents explain this.

We have now seen two classes of data where Saxe (2005a) has shown systematic error in ToM. These errors pose an as yet unanswered problem for ST. That scenario has been a major factor leading to the consensus Strong S/T Hybrid view allotting major roles to both theory and simulation. I will therefore in the next chapter consider problems with Strong S/T Hybrid views.

Chapter 6

Bias Mismatch Defence: Background

6.1 Introduction

Given the drawbacks faced by Strong S/T Hybrid accounts outlined in Ch. 4, we should examine the feasibility of remaining close to a pure ST theory with a Weak S/T Hybrid account. The first problem is that, as discussed in Ch. 5, Saxe (2005a) has shown, pure ST (and so also Weak S/T Hybridism) is vulnerable to the systematic error challenge. Recall that the challenge is brought by TT proponents who note the existence of error in ToM performance which is systematically slanted depending on the circumstances. In some circumstances, S's are systematically too positive in their expectations of the rationality or morality of the behaviour of O's; in other circumstances, S's are systematically too negative. There is no reason, according to opponents of ST, why ST should predict such systematic errors. ST should on the contrary predict random errors, according to those same TT proponents.

Given the strong empirical backing for the existence of these errors, ST proponents have little prospect of challenging the data. Even were they to

succeed in doing so, the approach would resemble some kind of ad hoc patchwork which would lack simplicity and parsimony. As I will explain in §6.2, two previous attempts to defend ST against the systematic error challenge have been essayed. The first of these was the wrong inputs defence, which urges that simulation can produce errors if the simulation is fed with the wrong inputs. Also attempted was a translation defence, which suggests that even if the inputs are correct, simulation error can occur if O is not rationally translating the outputs of his practical decision making system into actions, while S simulates O as translating those outputs rationally. I will conclude in §6.2 that these two prior defences of ST against the systematic error challenge have proved inadequate; in dismissing the wrong inputs defence I am in accord with Saxe (2005a, p. 178). This is why we need a new defence, providing which is the central task of this thesis.

I will then go on in §6.3 to give an initial overview of my Bias Mismatch Defence which can be stated roughly but succinctly in the slogan 'simulation may not accurately model bias.' More precisely, cognitive biases are simulated; but the biases to which S and O are subject may differ because of the factors affect and system. If S and O are not is bias matched states, then there will be simulation error. The details of how this defence works in action will emerge more fully in discussion of its application to Saxe's specific challenges in Chs. 8 and 9. In some scenarios, the S's simulations failed because they failed to include a bias of the O's in their simulation of the O's. In turn, they failed to include that effect because they were not in the situation faced by the O's, who had an affective involvement resulting from being told something about their competencies which may have been pleasing or displeasing. There was an Affect Mismatch between S and O and a resulting Bias Mismatch leading to systematic ToM error.

The formal structure of the Bias Mismatch Defence is as below.

1. FACTOR X affects O but not S[1]

2. FACTOR X modulates the probability of being subject to BIAS

BIAS is a placeholder for any cognitive bias now known or discovered in the future. Roughly speaking, one might talk of employing the Bias Mismatch Defence in any scenario in which there has been a systematic ToM error as a result of a failure to simulate the BIAS of O. Also, one might loosely apply the defence when a BIAS in S which O does not have has caused the ToM error. Strictly speaking, the defence should have the exact structure outlined above, in which FACTOR has affected O but not S and FACTOR has *caused* the BIAS in O which has led to S's simulation error. Again, FACTOR is a placeholder for any grounds for O to have a BIAS. I will make two auxiliary hypotheses as to what FACTOR might be. FACTOR could be an affective distinction between S and O, or it could be a reasoning system mismatch between S and O; it could also be a combination of both. I would be happy to see any further values for FACTOR which result in BIAS leading to systematic ToM error being classified as an instance of the Bias Mismatch Defence.

I will then outline in §6.4 the various biases involved in Bias Mismatch. I only list the ones I will be employing; there are many more[2] which could doubtless explain many other cases of simulation error. Any other occasions of systematic error which can be explained by further applications of biases constitute further evidence for my position, whether those biases are listed here or not. Since many of these biases are familiar, I will not discuss them in detail, restricting myself to providing a sketch and citing literature that provides more detail.

[1]Here I am using capitals for emphasis, not as the names of concepts.

[2]I am aware of informal estimates of 150+ as to the number of biases to which S's are subject.

6.2 Why We Need A New Defence

Simulation Theory has been charged with failure to predict the robust systematic errors that are observed in ToM. Two types of defence have been suggested: a Translation Defence and a Wrong Inputs Defence. Greenwood (1999, p. 35), writes that in ST "failure can only arise in one of two ways: either the decision maker's practical reasoning system is different from the person whose behaviour is predicted, or the right pretend beliefs and desires are not fed into the system." This adds up to a concise statement of the Translation Defence and the Wrong Inputs Defence together with a claim that no other options are available.

I will explain these two defences, both of which are offered by Harris (1992). I will spend less time on the Translation Defence for a number of reasons, the most important of which is that in my judgement, it does not succeed in providing a wide-ranging and non-ad hoc defence of ST, as I will outline below. It is also true to say that the Translation Defence has not received much attention in the literature; in fact, I have been unable to find any references to it. In common then with other commentators, Saxe focusses her challenge on the Wrong Inputs Defence. Saxe (2005a, pp. 177-178) sets out her challenge to the Wrong Inputs Defence as follows.

"The pattern of errors described above is not consistent with this kind of Simulation. And, as we shall now see, the most common defence of Simulation Theory against the argument from error also fails: the claim that errors arise from inaccurate inputs to the simulation".

Here, Saxe uses the term "pattern of errors" to refer to the general problem she raises for ST, that of explaining the systematic ToM errors. Saxe is correct in saying that the Wrong Inputs Defence has been the one more frequently resorted to by ST proponents to explain ToM errors. I will agree that Saxe is right to claim that the Wrong Inputs Defence does not explain the systematic

nature of the errors. This is why we need a new defence which I will offer in the next chapters. I will now briefly outline in turn the Wrong Inputs Defence and the Translation Defence.

6.2.1 Wrong Inputs Defence

The best statement of the defence is given by Harris (1992, p. 132), who is responding to Stich and Nichols. He puts the Wrong Inputs Defence as follows: "it is necessary for [S] to feed in pretend inputs that match in the relevant particulars the situation facing the [O] whose actions are to be predicted or explained. Predictive errors will occur if inappropriate pretend inputs are fed in." The Wrong Inputs Defence is the obvious one for ST proponents to reach for when charged with failure to predict the observed systematic ToM errors, since it is seems prima facie straightforward to argue that the simulation failed because it was fed with the wrong inputs. This if successful would allow ST proponents to claim that ST is still the correct account of ToM. This unfortunately does not work, as has been shown by Saxe (2005a) and as I concede, because there is too much data to be explained. The ST account would be committed by its employment of the Wrong Inputs Defence to a prediction of widespread error in ToM which is empirically false. To be sure, there are plenty of errors, as the data show, but on many normal occasions outside the psychology laboratory, everyday ToM use seems to work pretty well. S's often think they can predict and explain the behaviour of O's, and often those S's are right about that. If their ToM were as error prone as it would be if it was so easy for the simulations to be fed with wrong inputs, then those S's would not be right about their often successful abilities to predict and explain the behaviour of O's.

Moreover, wide application of the Wrong Inputs Defence would make the ST account look rather ad hoc, because it would be postulating special sorts

of wrong inputs in various different experimental circumstances. Recall as well that those inputs would have to be wrong in that special way *systematically*; a particular experiment would have to involve the same sort of wrong inputs every time in order to explain the observed systematic ToM errors. TT proponents can rightly object that this is unwieldy and implausible; even if it works, it will have a lot of moving parts. The problem can be observed in its nascent state when Harris (1992) attempts to deal with three experiments that Stich and Nichols correctly argue are problematic for ST. The three situations deal with Suicide Note Assessors, Lottery Ticket Holders and Shoppers. The first two groups are handled using the Wrong Inputs Defence and the Shoppers are handled using the Translation Defence; I will postpone discussion of the Shoppers until §6.2.2. The problem becomes much more severe later when the basic approach of Saxe (2005a) is to introduce much more data. So the charge of being ad hoc becomes much worse for ST proponents because now, not just three experiments but dozens must be handled by postulating specific sets of systematic wrong inputs. It will be seen later than one merit of the Bias Mismatch Defence is that it handles all of those data without being ad hoc.

Suicide Note Assessors

In this subsection, I will briefly cover four topics. I will first explain the experiment in question. Then I will say why the experiment is held to be a challenge to ST. After that, I will explain Harris's response on behalf of ST. Finally, I will outline the TT objections to the response, and say why I agree that those objections are decisive. So I will conclude that ST is in need of a defence to this particular challenge and does not at present have one.

I will term the experiment in question the 'suicide note assessment task.' Since the point of interest for us is ToM errors made by us as S's in relation

to how the O's in the experiment perform, I will refer to the experimental participants as 'the O's' throughout. Ross, Lepper, and Hubbard (1975, p. 882) gave their O's "25 cards, each containing one real and one fictitious suicide note." The task was for the O to assess which one of the two notes was real and which one was fictitious. After each trial, "the experimenter said only 'correct' or 'incorrect' (Ross, Lepper, and Hubbard 1975, p. 882). On completion of the 25 trials, O's were given "feedback indicating that they had correctly identified the actual suicide note on either 24 (success), 17 (average), or 10 (failure) occasions" (Ross, Lepper, and Hubbard 1975, p. 882). After receiving this feedback, the O "was then left alone for a period of either 5 (short delay) or 25 (long delay) minutes" (Ross, Lepper, and Hubbard 1975, p. 883). That period having elapsed, the O was then told that the success, average or fail feedback had been false and that 'the O's "score had been determined randomly" (Ross, Lepper, and Hubbard 1975, p. 883). The surprising result is that O's continue to harbour some beliefs that they were good or bad at the suicide assessment note task. As Ross, Lepper, and Hubbard (1975, p. 884) summarise, "even after debriefing procedures that led [O's] to say that they understood the decisive invalidation of initial test results, the [O's] continued to assess their performances and abilities as if these test results still possessed some validity."

In sum, evidence was presented to suggest that beliefs are recalcitrant to later evidence. This includes not just beliefs about the self but also about others, because Ross, Lepper, and Hubbard (1975) also replicate the results with observers i.e. they found that S's also continue to attach some strength to the belief that the O's were good or bad at the task even after the evidence therefor had been discounted. This point will prove interesting later when I discuss this experiment again in the context of showing how the bias mismatch defence can handle it (see §8.2.5).

So much for the experiment. Why is this a problem for ST? The problem is that the 'belief persistence' observed in the O's by Ross, Lepper, and Hubbard (1975) is not predicted by S's. S's predict that immediately after O's find out that the evidence should be discounted, O's will abandon the belief founded on the now discounted evidence. The force of this may be illustrated by considering the following question: if you believed X solely because of fact Y and I show you that Y is not the case, would you continue to believe X? TT proponents may now bring their standard challenge to ST viz.: if ST were true, one would expect S's to predict the correct outcome. S's should avoid the error by simply putting themselves in the situation of the O's assessing the suicide notes. So we have here a systematic ToM error that ST must explain.

How can Harris respond on behalf of ST? Harris (1992, pp. 132-133) responds to this challenge by essaying what we might term a 'time-lag defence.' He notes that "[an S] reading about such experiments and attempting to simulate their outcome is presented with a single, integrated account of both the trait information and its disconfirmation [so S] will find it difficult to reproduce the naive, unsuspecting commitment to the initial information that is entertained by [O]." Harris's defence is then the Wrong Inputs Defence in that S's are held to be given both confirmatory and disconfirmatory evidence simultaneously, while the O's have a delay of five or 25 minutes between presentation of the confirmatory evidence and the disconfirmatory evidence. As Harris (1992, p. 133) remarks, S "feeds in the pretend inputs in a different way from a naive [O]." The S's wrong input results from the combination of confirmatory and disconfirmatory evidence, which may lead to no belief at all. By contrast, the O's hold the belief that they are good or bad at the task for a longer period; even five minutes is a lot longer than no time at all.

This immediately leads us to the question as to what an input is. Does the timing of an input change the content of that input, or is it that same input

content which may be treated differently by the ToM processing depending on its timing? I believe Harris is right to make the second assumption, that the content of an input is not affected by its timing. We see that Harris makes this assumption since he writes of "the inputs" being fed in in "a different way" (Harris 1992, p. 133) i.e. they are *the same* inputs but the timing differences allows them to be processed differently by S and O. A primary reason for following Harris here is that not making this assumption is tantamount to saying that *all* inputs are different; there would be few or no occasions when we could say that S and O had 'the same' inputs and so the Wrong Inputs Defence would over-generalise and predict almost complete ToM error.

This argument is not just restricted to temporal differences in context. So in fact, I will assume, both on behalf of Harris and myself, that the content of an input is unaffected by *any* of its contextual factors. This amounts to a decision to articulate the Wrong Inputs Defence by building in the assumption that inputs are context independent, so that S's affective and other states cannot change the nature and content of the inputs. It will still of course be possible for S's affective and other states to change the ToM processing; for example, if S is under extreme stress, it would be strange for an account to say that the stress will not affect the outputs of S's ToM at all. This difference will become crucial later, when I suggest that in fact it is just this possibility of S's affective state being systematically different to O's that allows ST to explain systematic ToM error without appealing to wrong inputs.

In any case, Stich and Nichols object to Harris's time-lag defence. They do not have additional properly conducted experimental data to cite, but they have tried a non-controlled version of an experiment that would distinguish between Harris's view and their own. They focus on Harris's point about the time-lag between the receipt of the confirmatory and disconfirmatory evidence, and reasonably ascribe to Harris the prediction that "if we presented the in-

formation in two distinct phases, separated by an hour or so, people would make the correct prediction;" nevertheless, they find that "[m]ost of [the S's] still got the wrong answer" (Stich and Nichols 1995a, p. 101). This objection is in my view fatal to the time-lag extension of the Wrong Inputs Defence.

I will not be raising methodological quibbles about Stich and Nichols not having run a fully controlled experiment, because I am satisfied that such an experiment would confirm their view that the time-lag is not the problem for the S's. I will though be suggesting that the difference is in affective engagement between the S's and the O's —however well the experiment is described to the S's, it will not be the same as being there as a *participating* subject. Being in the room with someone is more involving than reading about what happens to them, but it is still nothing like as engaging as being that someone.

Lottery Ticket Holders

This example relates to an experiment in which O's are much more reluctant to return some lottery tickets than they rationally should be. The O's demanded much more money to return tickets they had chosen than to return tickets they were given, even though the tickets had the same chance of winning. The two conditions were referred to as 'choice' and 'no choice' of tickets. S's did not predict this difference in the amount of money demanded by O's. The shape of the S's predictions of the O's behaviour and the reasons S's give make it look like people rely on simple belief/desire folk psychology, as seems independently plausible. The S's believe that the O's desire to win the lottery prize and believe that owning a ticket will make that a possibility. The S's do not believe however, is that the O's will behave as though they believe that a ticket they have chosen has more chance of winning than one they have not chosen. Since the O's all behave in this way and the S's uniformly do not

work on that basis, the S's make a systematic ToM error. In response, Harris (1992, p. 133) offers the defence that S "needs to simulate the vacillation and eventual commitment of the [O's]. Moreover, in making that simulation they must also set aside the tacit reminder [...] that any Lottery ticket whether selected or allocated, has the same likelihood of winning."

The same results are obtained by Nichols, Stich, Leslie, and Klein (1996) when they re-run the Lottery experiment. Nichols, Stich, Leslie, and Klein (1996, p. 50) write that Harris complains that "it would hardly be surprising if the [S's] used the wrong pretend-inputs in making their prediction" if the delay between buying the tickets and being asked to sell them back was several days for the O's and several minutes for the corresponding questions to the S's. So Harris is once again essaying a time-lag extension of the Wrong Inputs Defence. The problem though is that Nichols, Stich, Leslie and Klein reduce the viability of the time-lag defence offered by Harris by eliminating the time lag: they show their new S's a video of the actual lottery experiment. For our purposes, the most important element of the Nichols, Stich, and Leslie (1995) reply is to note that "simulation theory predicts that someone watching the videotape of that part will correctly predict (simulate) the outcome" whatever that outcome is. So the S's should simulate the O's more accurately since the video represents a closer approximation to the actual experiment than merely reading a description of it.

Stich and Nichols (1995b, p. 100) have again not employed the scientific methodology of experimental psychologists; they admit their evidence is "anecdotal." This quibble must be raised this time, since Kuehberger et al. (1995, p. 423) conducted a properly controlled experiment and "consistently failed to replicate the original difference between choice and no-choice under the conditions used by Nichols et al." so "it is difficult to use it as a yardstick against which the accuracy of simulation can be assessed." A reply to these

charges is offered by Nichols, Stich, and Leslie (1995, p. 437) who deny that the failure to replicate of Kuehberger et al. (1995) is a problem for their objection to ST —they introduce further empirical evidence such that the sum "still weighs heavily against simulation." I will discuss this further evidence in Ch. 8. It is a point in their favour that there is a great deal more experimental data that ST must explain.

Harris has twice attempted to provide a time lag extension of the Wrong Inputs Defence. In the first case, with the suicide note assessors, it looks as though the time-lag defence is committed to a prediction of empirical results not found by Stich and Nichols. It is not an appealing escape route here for Harris to point to the experimental methods of Stich and Nichols being less than rigorous because there is little doubt that a more rigorous experiment would produce the same results. It is just implausible that anyone would ever predict that there would be recalcitrant beliefs that survive the elimination of the only evidence for them, whatever the time-lag between events in actuality and in simulation was.

Similarly, the time-lag defence does not appear to help with the lottery ticket holders, because having S's watch a video of the experiment is a good way of ensuring that the time sequence of events for S's is the same as it is for O's. So Harris's second attempt to introduce a time-lag defence seems to have failed. The key distinction between S and O in the lottery ticket example is not the time-ordering of events but rather the fact that O owns the ticket and S does not. I conclude that the Wrong Inputs Defence, even with the time-lag extension, does not deal with the experimentally-based objections to ST raised by Nichols, Stich, Leslie, and Klein (1996) and by Stich and Nichols. Also, Saxe (2005a) introduces a host of additional data which the time-lag extension of the Wrong Inputs Defence would also have to deal with. Therefore, ST needs a new defence, which is what I will be providing after I

consider Harris's second attempt at a defence, the Translation Defence.

6.2.2 Translation Defence

Harris (1992, p. 132) suggests a second defence beyond the Wrong Inputs Defence when he writes that: "any simulation process assumes that [O's] behaviour is a faithful translation into action of a decision that is reached by the practical reasoning system. If that assumption is incorrect, the simulation will err." The simulation could also be wrong even without wrong inputs if there is an error in *translation* from decision to action. I have therefore chosen to call the defence, which is clearly distinct from the Wrong Inputs Defence, the Translation Defence. This seems to capture the essential element of what it suggests has gone wrong, without I hope causing confusion. Nothing necessarily linguistic is implied; there merely needs to be a translation of the outputs from the practical decision making system into action, by whatever mechanism that is accomplished.

Here, a translation error just means that the way S translates the decision into an action prediction is different from the way that O translates the same decision into actual action. S will therefore make a ToM error in relation to the prediction of O's action even if S had all the same inputs as O did. Stone and Davies (1996, p. 135) give another description of the Translation Defence when they note that "there may be purely mechanical influences on decision taking that are not captured by mental simulation." O's may, as we will see next in the shoppers example, bypass their decision-making system altogether. If so, S will not be able to use his decision-making system to simulate such an outcome. We can understand the Translation Defence more clearly by seeing the use to which Harris (1992) puts it, which is to explain the mysterious behaviour of some Shoppers. I turn to that experiment and Harris's explanation of what is happening next.

Shoppers

This example relates to an experiment in which Shoppers chose without a basis when there were no rational bases for making a selection. Shoppers were "asked to say which article of clothing was the best quality" (Nisbett and Wilson 1977, p. 243) from a selection of four identical pairs of stockings. It transpires that they choose the rightmost pair much more often than they would if they chose randomly across the four pairs. One might expect them to choose randomly across the four pairs since there were in fact no differences between the pairs of stockings in quality or otherwise. The systematic ToM error here is that S's do not predict this rightmost pair bias in O's. Harris (1992, p. 133) responds: "the shopping-mall experiment [...] I suspect, involves the second source of difficulty identified above: faulty assumptions about what causes the [O's] behaviour rather than an inappropriate choice of pretend inputs."

The mechanism that Harris proposes is as follows. He thinks that the "[O's] action of choosing the right-most item is not governed by the decision-making system at all" (Harris 1992, p. 133) which would mean that S's would err in simulation because they simulate the operation of the decision-making system which is not in this case operating. This does seem plausible because if the decision-making system is operating, it is at least not operating rationally when it makes a choice on a non-rational basis, as here. Here, calling the decision non-rational refers to the lack of a rational basis for making the particular decision made, which does not exclude the possibility that it is rational to make some decision, and therefore rational to choose one of the pairs of stockings even if there is no reason to choose a particular pair.

Harris's argument for this bypassing of the decision-making system rests on the post facto confabulation that is observed in the O's in the shopping experiment. They do not report having decided to take the right-most item

for no particular reason; instead they fabricate a reason based on a false claim about the distinctive qualities of the right-most item. This almost suggests that the decision-making system is called upon subsequently to manufacture a justification for the choice that was made. In any case, Harris seems not to be on solid ground when he argues that had O's used their decision-making system in the normal way, they would not need to fabricate a reason; neither would they have forgotten the reason they had if they had one, and so the decision-making system is bypassed. Many, perhaps all, of his opponents on the TT side would not accept that persons generally have good access to their reasons for acting, because TT proponents often deny Introspectionism, as described above.

We may be able to give Harris a possible response, involving an attempt to claim that the TT account here over-generalises. This would seek to make out the claim that the TT account basically involves a denial that there is a decision-making system at all, in the way one would normally understand the term. There is no decision-making system because we never or rarely have access to our reasons for acting. If there is something we refer to with the term 'decision-making system,' it might be more accurately named 'post-hoc decision justification system.' Harris could simply bite the bullet and assert Introspectionism; this would involve appealing to our phenomenology. People are sure they know why they act; we feel it 'from the inside': we all naturally talk in terms of belief/desire folk psychology. I think these approaches might work out for Harris, but I think it would leave him open in the wider context of Saxe's data to a fatal charge of being ad hoc, which I will outline now.

A more serious and wider problem for Harris here is that Saxe has introduced a great deal more data than just the suicide note assessors, the lottery ticket holders and the shoppers to support the systematic ToM error challenge to ST. All of the data she introduces would need some kind of special

treatment of this kind. There would be time-lag defences and shopping mall defences. One might wonder whether any responses are available to Harris here. Perhaps Harris can argue that the world is complicated and complicated explanations are therefore needed. There are two problems with this potential response. First, the complexity of the world, and the mind, both of which must be conceded, do not entail that all explanations are complex. The Mandelbrot set, while generated or explained by repeated application of a simple equation, is enormously complicated. Secondly, following on from that point, a complex explanation can only be the best route in the absence of a simpler one. My explanation of the data will be such a simpler one: I will say that bias mismatch between S and O is the simple common factor explaining ToM error in a wide range of experimental cases. Can Harris's position now be saved by denying that my explanation is a simpler or less ad hoc one? I believe not, and this will become clearer still once I have presented more of the experimental evidence in the next two chapters. It will be shown there that bias mismatch is an solution with a great deal of explanatory power across a wide range of circumstances. For now we may merely note that bias mismatch between S and O neatly explains the ToM errors seen in the three cases of the suicide note assessors, the lottery ticket holders and the shoppers. My defence also escapes the charge of being ad hoc in virtue of its auxiliary hypotheses that affect mismatch and system mismatch will often play a role in generating bias mismatch, as I will outline in Ch. 7, on motivating the bias mismatch defence.

What we have here is one version of the Translation Defence: S's do not simulate O's bypassing their decision-making system. I have preferred to term the account the Translation Defence rather than possible alternative names such as a 'bypass defence' since the central idea of the defence is that the output of the decision-making system of O is not accurately translated into action. This may occur because the decision-making of O is bypassed or because of

some other 'mechanical' influence. This approach means I am regarding a bypassing of the decision-making system as a form of translation error, which makes sense if we regard a bypassed output from the decision-making system as one which has not been correctly translated into action. We might think that simulating S's specifically engage their decision-making system because they have been asked, they believe, to simulate a decision. This could explain ToM errors, if there is in fact a great deal of mistranslation going on in O's. But it does not seem as though we can make much progress by assuming that O's bypass their decision making machinery on a widespread basis: that would entail an empirically false prediction of wholesale ToM failure.

It is interesting to note that all three of these examples relate to value judgments. In the case of the suicide note assessors, the value judged by S is the level of ability of O in assessing whether the suicide note is genuine or fake. In the case of the lottery ticket holders, the value judged by S is what economic value O will place on a chosen lottery ticket versus a non-chosen one. In the case of the shoppers, the value judged by S is which pair of a set of pairs of identical stockings O will say is the highest quality. It might then appear at this stage as though the bias mismatch defence is only going to apply for systematic ToM errors involving value judgements. While there will be more examples of this type in Ch. 8 and Ch. 9, there will be plenty of other types of data considered also. The bias mismatch defence will have wider application than solely to value judgements.

In summary, I will agree that the time-lag extension does not save the Wrong Inputs Defence; the Translation Defence lacks widespread applicability and pursing this route in any case will result in an ad hoc set of approaches, because of the wide array of data introduced to challenge ST. This necessitates a new defence, which I will provide. I will also be proposing that the Bias Mismatch Defence takes a unified perspective across the data and is thus not

exposed to the charge of being ad hoc, which a set of extensions of Harris's defence would be.

6.3 Bias Mismatch Defence: Outline

The Bias Mismatch Defence is the claim that S's simulation of O may fail because S and O do not apply the same cognitive biases. Simulation may fail because S operates with different cognitive biases to O, where the difference could be that a different bias is applied by S than by O, or the same bias is applied by S and by O but in different intensities, or the same bias is applied by S and by O with the same intensities but about different items. All of those eventualities would result in simulation error even absent wrong inputs.

I will defer the important motivation question – why should we expect there to be bias mismatches? – to the dedicated Ch. 6.

The question also arises as to whether or not this new defence I offer is a variant of one of the two previous defences or not. That is an important question, because I am arguing that both of those defences fail. Therefore, my position would become incoherent if I fail to show clear separation between my account and those previous two defences. It would still be possible for me to say that the efficacy of the defence is more important than its classification, but dialectically, that looks best retained as a fall-back position. In fact, clear separation is provided by making the assumption I outlined above on p. 161, that the content of an input is unaffected by its contextual factors. S's biases, affective and other states cannot change the nature and content of the inputs to S's simulation, on this assumption. This distinguishes my account from Wrong Inputs Defence since on my view, the inputs can be right and the simulation still fail. This is how I will account for the systematic ToM errors of which Saxe (2005a) complains without assuming wrong inputs.

The Translation Defence is also clearly distinct from the Wrong Inputs De-

fence since the former postulates that the difference between S and O lies in how the inputs are handled as opposed to what they are. Therefore I also need to show that my new defence is not a form of the Translation Defence. Fortunately this is straightforward. My defence is also clearly not to be classified with the Translation Defence since my account does not and need not postulate errors in translation of outputs of O's decision-making system into action; it will instead be postulating biases in O applying while O makes decisions.

In fact, attempting to classify my account as one of the two previous defences would involve saying where my account locates the source of the errors. Since my account locates that source in bias mismatch between S and O, that question devolves to 'where are the biases?.' And trying to decide where, for example, the failure of S to model a bias of O takes place could be seen as being an ill-formed question, since we cannot specify a location where something does not occur. If we had a specified functional location for where the biases are applied in O, then we might be able to say that the difference between those bias-applying locations in O and the same, but not bias-applying, locations in S are where the difference between S and O is found. However, it is possible that these biases are wide-spread throughout the isomorphic procedure of simulation; or that the question has no answer: as Apperly (2008, p. 281) writes, "there is no systematic basis for drawing a line between the inputs to a particular reasoning episode and the start of the reasoning itself." Since I have shown that my account is clearly distinct from the previous ones and therefore need not fail as they do, we may move on to the more pressing business of showing that my account can succeed in explaining systematic ToM error.

The idea behind my defence may be illustrated by considering an example from the book Asch (1952) cited by Saxe (2005a). Consider the following questions, all related to a scenario in which you are given a list of personal

characteristics and asked to assess the personality of the person to whom the list applies.

- Would you assess the characteristics fairly?
- Would you assess them regardless of irrelevant features?
- Would you assess them regardless of the order in which they were presented?

I submit that you will answer all of these questions in the affirmative. Moreover, if you were asked whether you would expect someone else to perform in the same way, you would also affirm that, short of any specific information suggesting malice or lack of competence in the other person.

Now look at the following two lists of characteristics from Asch (1952, p. 212).

A intelligent —industrious —impulsive —critical —stubborn —envious

B envious —stubborn —critical —impulsive —industrious —intelligent

Here I contend that, consistent with what Asch found, you will form a more positive impression of the person with the characteristics described in list A than in list B. In this, you will be representative of people generally. As Asch (1952, p. 212) puts it, list A describes "an able person who possesses certain shortcomings" while list B describes a "problem" person whose "abilities are hampered by his serious difficulties" (Asch 1952, p. 212). This means in your original assessment of yourself, you have committed a ToM error, because you failed to forecast that either you or the experimental sample will make such distinct judgements based on a list of characteristics which are the same in each case but merely in reverse order.

It might be objected here that it is in fact rational to apply a heavier weighting to the first-appearing characteristics, an approach equivalent to

making the assumption that the characteristics have been presented in order of significance or importance. I do not believe that the objection succeeds however, for two reasons. Firstly, no statement in relation to the importance of the ordering was given to the participants by the experimenters, so the objection assumes without other motivation that all of the participants took it upon themselves to accord importance to the ordering. This may not be conclusive, since it could and might have happened that many participants took the characteristics as having been ranked in order of importance, but that would be an assumption which such an account would be making which would be a theoretical cost and which would be empirically testable. Secondly, and crucially, one of Asch's more technical journal papers give us further examples. Asch (1946, p. 264) found that "a change in one character-quality has produced a widespread change in the entire impression." This means that changing a single characteristic in a list of six completely alters the participants' general impression of the person described. The switch in question, from 'warm' to 'cold,' is doubtless of some significance. However, it is not rational to weight it much more heavily than all of the other five combined, particularly since Asch (1946, pp. 267-268) also finds that changing the other five characteristics can greatly reduce the influence of warm/cold. Indeed, Asch (1946, p. 273) finds that the treatment of a characteristic can vary immensely, between its being ignored completely if it does not fit the general impression and outweighing all of the other characteristics. This is surprising. One response here might be that adding or removing one characteristic might rationally make a big difference to the overall perceived profile of a personality. This seems true for very significant, perhaps dominant characteristics. One might see 'violent' as being in a special category which can alone cause a complete revision of assessment of someone. But putting 'warm' into that category seems like a stretch. Since none of these character assessment data

seem to be predicted by our own ToM, we have an illustration of the sort of social psychology experiments which surprise us and indicate to Saxe (2005a) the presence of systematic ToM errors.

Another question here is to ask whether this is really a problem for ST. Naturally, if it is not a problem for ST, then my defence of ST need not deal with the problem. One might in similar vein think that the Conjunction Fallacy is not a problem for ST. That line would suggest that as soon as we see how Tversky and Kahneman (1983) have phrased their famous 'Linda' question, we know how people will answer. There is an immediate pull to answer the question wrongly and commit the fallacy. I nevertheless see value in including discussion of such questions as those raised by the Linda experiment, and the character assessment example above which is intended to illustrate and introduce Asch's work, because the data are consistent with the bias mismatch defence for which I argue. So even if ST is not strictly speaking caused problems by a particular experiment, it is still valuable to show that bias mismatch explains the data, as I will also be arguing.

It is interesting to note here that participants seem in parts of this experiment to succumb to the bias known as the Halo Effect, which outcome prefigures the type of explanation I will be presenting of systematic ToM error. I might finally point out that even if the objection is successful, that would just mean that this experiment does not illustrate a systematic ToM error, so it drops out of the category of data that ST must explain.

Now we come to the shape of the defence. The reason O's assess the characteristics 'unfairly' is that they fail prey to Confirmation Bias. The term Confirmation Bias refers to the "fundamental tendency to seek information consistent with current [...] beliefs, theories or hypotheses and to avoid the collection of potentially falsifying evidence" Evans (1990, p. 41). In other words, O's tend to look for data confirming what they already believe. Thus,

information arriving earlier is given more weight in assessments; the later information has to countervail the earlier information insofar as the later information goes against the earlier data. The reason S's fail to predict this is that simulation here does not model bias. The Bias Mismatch Defence is just this: it is the claim that simulation by S of O can be systematically inaccurate because there can be systematic bias in O which is not simulated by S. Note also here the clear distinction between being asked dispassionate, clinical, salient questions like the ones in the list about how you would do the job and actually being in the situation of assessing the characteristics. We will see this affective mismatch and its analogues on a great many occasions later.

6.4 Bias Mismatch Defence: Biases Involved

It is well-known that we exhibit many errors in our reasoning due to a large number of cognitive biases. We often use cognitive shortcuts or heuristics which are effectively biases, and as Tversky and Kahneman (1974, p. 1125) put it, "these heuristics are quite useful, but sometimes they lead to severe and systematic errors." I set out below a sketch of the biases I will employ in the mismatch defence. How they work will become clearer when I use them later to explain data on systematic ToM error introduced by Saxe (2005a). Naturally, any objections to the effect that I need to use further biases would count as a friendly amendment: I aim to prove that some combination of Bias Mismatches can explain the systematic ToM errors and that can be done using a variety of biases.

6.4.1 Representativeness Heuristic

Tversky and Kahneman (1974, p. 1124) define the Representativeness Heuristic as occurring when "probabilities are evaluated by the degree to which A is representative of B, that is, by the degree to which A resembles B." Intu-

itively, we may regard this as stereotyping, because a typical application of the heuristic will involve people deciding that someone is a librarian because they fit the stereotype of a librarian. The error is also known as 'base rate neglect.' Subjects fail to take account of what should be a much more significant factor in the probability estimate viz. the number of people in the population who are librarians.

The Representativeness Heuristic was investigated by giving subjects descriptions of the personalities of a group of persons. Tversky and Kahneman (1974, p. 1124) write that "subjects were told that the group from which the descriptions had been drawn consisted of 70 engineers and 30 lawyers" or vice versa. The subjects were then asked to assess the probability that a given person was an engineer or a lawyer. The descriptions were slanted to be engineer-like or lawyer-like. For example, a stereotypical engineer will enjoy fixing his car at weekends while a stereotypical lawyer will be tenaciously argumentative in personal situations.

Tversky and Kahneman (1974, p. 1125) found that subjects ignored the population probability data. If given an engineer-like profile, they said the person was probably an engineer, even when they had also been told that the sample consisted of 70% lawyers.

6.4.2 Availability Heuristic

Tversky and Kahneman (1973, p. 208) write that "[a] person is said to employ the availability heuristic whenever he estimates frequency or probability by the ease with which instances or associations could be brought to mind." For example, "one may assess the divorce rate in a given community by recalling divorces among one's acquaintances" (Tversky and Kahneman 1973, p. 208). This is reasonable as a first approximation, but will be subject to inaccuracy depending on the events of one's life. If one happens to know many divorced

people, one will likely over-estimate the prevalence of divorce in wider society.

Tversky and Kahneman (1973) measured the Availability Heuristic by asking subjects to rate the probabilities of certain syllables occurring in words. They found that subject's responses were driven by the ease with which they could think of examples, rather than the actual probabilities, even though subjects obviously had a great deal of experience of words in their native languages.

Tversky and Kahneman (1973, p. 212) found that subjects "erroneously judged words beginning with re to be more frequent than words ending with re." This came about because it is easier to think of words beginning with re than ending with re, because it is generally easier to think of words with a specified beginning than with a specified ending. This means the words beginning with re were much more available and this produced the faulty probability estimate.

Two further factors feed into availability: salience and vividness.

Highly salient events will warp probability judgments via their increased availability. Taleb (2008, p. 58) gives several examples including that of someone who heard of someone's relative who was mugged in Central Park. This is likely to be much more salient for them than the statistics relating to muggings in Central Park and therefore much more available. They will likely greatly overestimate the probability of being mugged in Central Park. Such a story is also highly vivid, which leads us to the second factor.

In outlining vividness, Evans (1990, p. 27) credits Nisbett and Ross with the observation that in our reasoning, we "overweight vivid, concrete information and underweight dull, pallid and abstract information." This is intuitively plausible, just from considering that we prefer the vivid to the dull. More vivid information is more available. Evans (1990) again relies on Nisbett and Ross to supply three characteristics of vividness, which are "(1) emotional interest;

(2) concreteness and imageability and (3) temporal and spatial proximity."

Salient and vivid items are more available and receive higher probability estimates.

6.4.3 Conjunction Fallacy

The probability of two events A and B is given by multiplying the probability of event A by event B. For example, if the chance of a coin toss coming up tails is 50%, then the probability of getting two tails in a row is 25%. The maximum probability of an event is 1, or 100%, for events which are certain to occur. A consequence of this is a law of statistics called the conjunction rule which holds that the probability of both events A and B occurring must be no greater than the probability of event B occurring alone. This is because the probability of A and B occurring will have a maximum value when A is certain and that maximum value will be the same as the probability of B occurring alone. As Tversky and Kahneman (1983, p. 298) state, "[t]he violation of the conjunction rule in a direct comparison of B to A&B is called the conjunction fallacy." In other words, the Conjunction Fallacy occurs whenever we assess the probability of two events as higher than one of them alone.

The canonical illustration of the Conjunction Fallacy is the famous 'Linda' experiment. Subjects are told that Linda majored in philosophy, is very bright and as a student "was deeply concerned with issues of discrimination and social justice" (Tversky and Kahneman 1983, p. 297). Subjects are then asked whether it is more likely that a) Linda works as a bank teller or b) Linda works as a bank teller and is active in the feminist movement. Subjects consistently state that b) is more probable, even though it is impossible that b) could be more probable than a) alone, since b) includes a).

The Conjunction Fallacy is closely related to the Representativeness and Availability Heuristics, since what is happening is that a reduction in extension

is being combined with an increase in representativeness and availability. Thus it becomes easier to think of examples of a category even when the number of members of that category has decreased. This is what leads us to make the errors in probability estimation. There are also links to what Taleb (2008, Ch. 6) calls the Narrative Fallacy, which combines our tendencies to remember facts linked by a story and over-attribute causation. It is much easier to construct a story about Linda being a committed social activist at college and continuing with those interests later. This is why Tversky and Kahneman (1983, p. 299) found that 85% of subjects rated b) more likely than a).

6.4.4 Fundamental Attribution Error

The Fundamental Attribution Error is defined by Ross, Amabile, and Steinmetz (1977, p. 491) as "the tendency to underestimate the role of situational determinants and overestimate the degree to which social actions and outcomes reflect the dispositions of relevant actors."[3] The error reflects our false belief in stable personality: we ascribe the behaviour of others more to their 'characteristics' than to the situation they were in. Darley and Batson (1973, p. 108) found that "personality variables were not useful in predicting whether a person helped or not:" that was explained by whether or not the person was in a hurry. Also, Kamtekar (2004, p. 465) reports on many experiments including honesty studies which showed no "correlation across behaviour types" e.g. that someone who cheats in a test is not more likely to take money from a box. There seems to be nothing like a character trait of dishonesty. Overall, we often commit the Fundamental Attribution Error, including whenever we say something like 'of course he would do that, that's what he is like' —but there is little evidence supporting the existence of stable character traits and

[3]See Andrews (2008, [p. 13) for argument to the effect that "folk psychology includes the notion that some behaviour is explained by personality traits," as is consistent with the Fundamental Attribution Error.

plenty against.

Saxe herself at one point employs the Fundamental Attribution Error in a way that could be seen as a version of the Bias Mismatch Defence. Saxe (2009, p. 263) suggests that "other people's actions are ascribed to stable traits, whereas one's own actions are generally seen as variable and situation-dependent" and this leads to ToM error.

6.4.5 Conformity Bias

I will term this particular bias Conformity Bias, following Plotkin (2011), who does not however give a brief definition of the term.[4] Although the pioneer, Asch (1952, p. 467), does not use the term Conformity Bias, he writes that he has observed "a great desire to be in agreement with the group;" the thwarting of this desire leads to fear, longing and uncertainty. The reference group might be those physically present or a group that the subject identifies with. The bias is often called "the Asch effect" in the literature, but I would prefer a more descriptive term.

The most significant chapter of Asch (1952, pp. 450-501) from the perspective of conformity is Ch. 16, on "Group Forces in the Modification and Distortion of Judgements." Asch describes experiments where small groups of individuals are asked to judge which of three test lines are identical in length to a given standard line. All participants call out their answers. A deception is involved, because all but one of the participants are in fact in confederation with the experimenter. They have been instructed to call out obviously false answers. The key question is what will the non-confederated participant —the 'critical subject' —say in the face of such a perplexingly obtuse majority.

The results are that the error rate of the critical subject is 33.2% if the majority is wrong but only 7.4% if the majority is correct. This means that

[4] Prentice (2007, p. 18) does use the term in the way I do: "conformity bias strongly pushes people to conform their judgments to the judgments of their reference group."

the critical subject is induced to abandon his correct choice in favour of an obviously false group choice with a much higher frequency than can be explained by genuine error. This majority influence meant that "erroneous announcements contaminated one-third of the estimates of the critical subjects" (Asch 1952, p. 457). This observation forms a clear illustration of the Asch Effect or Conformity Bias. This bias is very strong; Prentice (2007, p. 18) notes that "[m]ore than 60 percent of the subjects gave an obviously incorrect answer at least once."

6.4.6 False Consensus Effect

Ross, Greene, and House (1977, p. 279) define the False Consensus Effect when they write that "social observers tend to perceive a 'false consensus' with respect to the relative commonness of their own responses," where responses might be actions, choices or opinions. So, "raters estimated particular responses to be relatively common" (Ross, Greene, and House 1977, p. 279) —viz, the ones they had themselves made.

Ross, Greene, and House (1977, p. 279) conducted a number of experiments: one of them was called the 'supermarket story.' Subjects are asked to imagine that they are just leaving a supermarket, when they are asked whether they like shopping there. They reply that they do, since that is in fact the case. It is then revealed that the comments have been filmed, and the subject is requested to sign a release allowing the film to be used in a TV advertisement. The key question is then asked: the subject or 'rater' is asked to estimate the percentage of people who will sign the release.

The results were that raters overestimate the percentages of others who make the same choice they would. Ross, Greene, and House (1977, p. 294) conclude that "raters' perceptions of social consensus and their social inferences about actors reflect the raters' own behavioural choices."

6.4.7 Self-Presentation Bias

Igoe and Sullivan (1993, p. 18) give the definition when they write that "[i]ndividuals show Self-Presentation Bias by projecting personal behaviours that present themselves more positively than others." The Self-Presentation Bias is perhaps more of a natural psychological tendency than a cognitive bias, though that will not concern us since the effects are the same. Put simply, Self-Presentation Bias expresses the way that people generally wish to show themselves in a positive light. They may do this by selective story-telling or otherwise.

Igoe and Sullivan (1993) measure the rates at which individuals work at hard tasks and find that they systematically over-report their own likelihood of returning to a hard task. Thus, the individuals exhibit Self-Presentation Bias in that they make it appear as though they are more likely to work hard than they really are. Interestingly, the subjects also attributed a lower propensity to return to the task to a fictional character, thus enhancing their own position in relation to others.[5]

Kopcha and Sullivan (2006, p. 628) note that "self-report data often reflect a phenomenon known as self-presentation bias or social desirability bias —that is, a tendency of individuals to present themselves and their practices in a favourable way." They measure Self-Presentation Bias in a group of teachers, who all said that they engaged in an array of positively perceived teaching practices more than their colleagues. Similarly, Kopcha and Sullivan (2006, p. 629) cite Self-Presentation Bias as the cause in studies reporting that "medical professionals often overestimated their level of adherence to the guidelines for clinical practice." More generally, we may agree with Pronin, Gilovich, and Ross (2004, p. 788) who observe that there is "mounting evidence that people

[5]This assumes a continuity between how people assign properties to themselves and how fictional objects obtain their properties. For more on these vexed questions, see Short (2014).

are motivated to view themselves, their assessments, and their outcomes in a positive light."

6.4.8 Clustering Illusion

Gilovich (1993, p. 16) defines the Clustering Illusion as occurring when we believe falsely that "random events such as coin flips should alternate between heads and tails more than they do." For example, in a sequence of 20 tosses of a fair coin, there is a 25% chance of a sequence of six heads, which seems to us far too ordered to be random. Alternatively, consider the probability of the two sets of results of coin tosses: HHHTTT looks much more pattern-rich and therefore improbable than HTHHTT but they actually have the same probability. The Clustering Illusion is the tendency to see patterns in data that are not really there. Gilovich (1993, p. 15) provides further examples including a belief that the random pattern of bomb sites in London actually shows a pattern; this effect is due to selecting the quadrant frame almost in order to arrive at the view that some quadrants of London were more heavily bombed. In general, our abilities to handle random noise are poor; we see patterns everywhere and we even see faces in the side of cliffs.

6.4.9 Confirmation Bias

The remaining biases including Confirmation Bias have already been described above, so I will be brief for the rest of this section. As mentioned in Ch. 6, Evans (1990, p. 41) defines Confirmation Bias as the "fundamental tendency to seek information consistent with current [...] beliefs, theories or hypotheses and to avoid the collection of potentially falsifying evidence."

6.4.10 Belief Perseverance Bias

As Nestler (2010, p. 35) observes with copious references, "belief perseverance has been observed in social perception [...] and self-perception [...] and it is robustly shown that individuals cling to beliefs even when the evidential basis for these beliefs is completely refuted." The Belief Perseverance Bias was illustrated above in the discussion of the suicide note assessors in Ch. 6.

6.4.11 Endowment Effect

Kuehberger et al. (1995, p. 432) write that the "endowment effect [...] means that simply being endowed with a good gives it added value." It can be seen when students are asked to estimate the price of a visible item such as a mug with a university crest on it. They make an estimate and are then actually given the mug and asked what they would sell it for. It turns out that they demand a much higher price for the mug now that they own it than the figure they gave previously for its value.

The Endowment Effect was in fact behind the results discussed above in relation to the lottery ticket holders. They assigned a higher value to tickets they had chosen than to ones they were given, although the economic value of the tickets was identical irrespective of whether they had chosen them or not. It seems as though their sense of ownership was more awakened by choice.

6.4.12 Position Effect

Nisbett and Wilson (1977, p. 243) give the definition when they write that they measured "a pronounced left-to-right position effect, such that the right-most object in the array was heavily over-chosen" in the experiment with the shoppers. The shoppers had to say which of an array of identical pairs of stockings was of superior quality. This Position Effect was discussed above in Ch. 6.

Chapter 7

Bias Mismatch Defence: Motivation

7.1 Introduction

One apparent problem for my Bias Mismatch Defence would result from the possibility that biases might be thought to be in some way more systematic rather than temporary disruptions. That would make the explanation look a bit too much like 'we make systematic errors because we make systematic errors,' which has no explanatory value because it is circular. I agree that cognitive biases are a stable part of our cognition. So the motivational question becomes important. An account must be given as to why these bias mismatches occur. In the formulation of the structure of the Bias Mismatch Defence given earlier on p. 155, we had FACTOR and BIAS. What I will be suggesting in this chapter is that two candidates for FACTOR are affect and system (or both). FACTOR is important for motivational reasons: my account needs to give a reason why there is Bias Mismatch if it is to non-circularly explain ToM error. 'BIAS therefore BIAS' is not explanatory. My claim is not that bias is not simulated, it is that affective state and system

affect which biases to which an individual is subject.

What motivation for my defence can be given beyond the fact that it explains all the data? I will outline below three routes on which bias mismatch is a plausible outcome. It is important to note that all of these claimed routes represent auxiliary hypotheses. All of them could be false but the bias mismatch defence could remain intact.

The first auxiliary hypothesis postulates affect mismatch. This means that Bias Mismatch may occur because S and O are in different affective states. This would be likely in cases where it is known that affect makes biases more likely to be applied, as is observed when people are put under stress. Anecdotally, we would expect stressed or emotional people to apply cognitive biases more than calm, rational people. The same is observed experimentally. As Mineka and Sutton (1992, p. 65) observe, depression appears to be associated both with "mood-congruent judgmental biases" and "a memory bias for negative mood-congruent material." Notably, the authors go as far as to define cognitive biases as "any selective or non-veridical processing of emotion-relevant information" (Mineka and Sutton 1992, p. 65) showing the close link between affect and bias. I will discuss how affect mismatch can lead to bias mismatch further in §7.2.

The second auxiliary hypothesis postulates system mismatch. Dual Process Theory (Sloman 1996) suggests that there are two systems of reasoning, the quick but inaccurate System 1 and the slower but more rational System 2. If S and O apply different systems, simulation is again likely to fail. If S calmly and rationally simulates a panicked or depressed O, then S will likely be using system 2 and O will likely be using system 1. This is a formalisation of the intuition I have mentioned previously to the effect that thinking about someone hanging off a cliff by their fingertips is not as emotionally involving as actually hanging off a cliff by one's fingertips. I will discuss more fully how

system mismatch can lead to bias mismatch in §7.3.

It is apparent from the example above that there may be overlaps between the two auxiliary hypotheses. It appears as though in the cliff-top scenario, we have both affect mismatch and system mismatch. In other scenarios, we may have system mismatch without affect mismatch and also conversely, affect mismatch without system mismatch. I will discuss further how these other scenarios may produce bias mismatches in §7.4.

Taken together, the two auxiliary hypotheses and their interactions answer the motivational question. We expect bias mismatch to occur because we expect affect and system mismatch to occur. This also deals with the circularity problem. The account does not make a prediction of the form 'S's make systematic errors because S's make systematic errors.' It instead makes predictions of the form 'S's make systematic ToM errors when S's are in systematically different affective states to O's,' and 'S's make systematic ToM errors when S's are systematically using different reasoning systems to O's.' Now of course it becomes incumbent on the account to explain why that might be, and that task will be conducted systematically by looking at the experimental data, in Ch. 8 and Ch. 9. The idea will be that the experimental situations systematically induce affect mismatch, for example.

7.2 Affect Mismatch

My response to Saxe's challenge will be that Bias Mismatch between S and O can supply the missing element to ST (and so Weak S/T Hybridism) to allow it to explain the systematic ToM errors. Often, it will be the case that this Bias Mismatch is in turn a result of Affect Mismatch between S and O. It is acknowledged in the psychological literature that affect can lead to the application of cognitive biases. As Pronin, Puccio, and Ross (2002, p. 636) observe, not only do humans add information to the world, but "perceptions are

further biased by their hopes, fears, needs and immediate emotional state."[1]

We do not accurately allow for the biases of others because we are not as exposed as they are to the live situation. Even if we are present, it is much less involving to observe someone hanging from a cliff top by their fingertips than it is actually to be in that situation. S simply cannot feel or imagine the affective position of that O to any significant extent. S is more remote still if S merely hears a dry description of the situation given in a rather clinical fashion. There can be different degrees of such affective detachment, which will impede simulation, as Goldie (1999, p. 410) points out in a discussion of imagining being attacked by a jellyfish. He notes the different affective import of imagining the attack "whilst sitting at my desk in London, whilst swimming in a pool, and whilst swimming off the coast of South Africa." The fact that we can so easily do this and so easily agree with Goldie is to my mind in itself an argument for ST. As we will see, much of the empirical data on ToM errors falls into this category: perforce, if it is properly collected data, it has been collected in a scientific manner which excludes S feeling fully engaged in the situation of O.

The same distinctions can apply when S considers the position of S himself. As Goldie (2011, p. 129) notes in a discussion of S's views of S in the past, there are multiple ways in which the S now can differ from O as past S; the gap between S now and then "can be triply ironic: it can be ironic epistemically —I now know what I did not know then; it can be ironic evaluatively —I now evaluate what happened in a way that I did not at the time; and it can be ironic emotionally." The irony referred to is the 'dramatic irony' that exists in a theatre when the audience knows something that an observed character does not —this can form an interesting parallel to our examination of S and O. At least the last of these three forms of irony and probably the second as well

[1] Coplan (2011, p. 12) discusses how differences in affect may lead to simulation error.

have a strong affective component. We may also agree that even information asymmetry can have affective import, as suggested by the very term 'dramatic irony.'

Further evidence for this connection between affect and ToM may be derived from Boucher, when she cites Kanner as including affect in the original 1943 definition of autism. Boucher (1996, p. 228) writes that Kanner "originally suggested that autism [...] results from 'an innate inability to form the usual biologically provided affective contact with people' ." Given the well-known association of autism with ToM deficits, we can see that if Kanner's original definition is correct, lack of affective contact with others will impair ToM capacities.

Arguing positively for the connection between affect matching and successful mindreading, Biggs (2007) suggests that 'phenomenal simulation' —where S's phenomenal state resembles O's —may be an aid to mindreading and introduces claims that there is similar neurophysiology occurring in those who experience and merely observe disgust, pain, etc. Some might object here that 'phenomenal' simulation might proceed just as well without qualia, so long as the isomorphism of states and progress between them is the same in S and O, because S will still arrive at a good prediction of O's behaviour. I suspect that this is possible, but it seems less plausible and efficient than an account in which S simply matches O's states affectively in an attenuated way as well, unless one believes in zombies (Chalmers 1997).[2]

In the next two chapters, we will discuss a large number of cases of errors made by S's in assessing what O will do in certain, often stressful situations. Even though there will be processing differences between S and O, there will also be different inputs for S and O: namely, the affect actually felt by O in

[2]See also Mealey and Kinner (2002) for argument that psychopaths do not empathise as much as controls because they have flattened affect; and Short (2015, Ch. 10) for argument that the flattened affect of schizophrenic subjects causes their ToM impairments.

the situation. If S were able to model the stress of O accurately, it might lead to a reduction in ToM errors.

One objection will be to ask why this same bias does not apply when the model is run. S suffers from the same types of cognitive bias as O does. We need this to explain why there can be errors —if exactly the same system is run by S as by O, and there were no wrong inputs, then S would generally not be wrong about the mental state of O. The answer to this is that the specific bias occurs only for O's and not for S's. Why this is so may be because it is just not as engaging to be S as it is to be O —in any situation. Again, it simply is nothing like as fear-inducing to imagine hanging from a cliff by one's fingertips as it is actually to be in that situation. The biases are triggered more by the affect of the situation. While S will doubtless be experiencing some affect, and it may even be sufficiently engaging to trigger some of S's own biases, the affects will not be the same ones as those experienced by O.

One type of Affect Mismatch might be fear differentials. Gordon (1986, p. 161) picks up on the difficulty of adding really experienced fear to the simulation in his early paper, indicating it with his italicisation. He writes: "[i]f I pretend realistically that there is an intruder in the house I might find myself surprisingly brave —or cowardly." It might even be deleterious to simulate the fear well; S's might become unable to act when faced even with the prospect of danger. It is only possible for S to be surprised about S's bravery if S has a different level of affect, and thus different biases applying, in the simulated case and the real case. Gordon also notes here that self-deception may corrupt the simulation effort. This is highly consistent with the approach I propose here —perhaps S's often deceive themselves about the frequency with which S's use biased thinking. Dennett (1979, p. 37) notes that an affective involvement may lead to self-deception —which we may understand as a failure of ToM —when he writes that if S lacks "any

remarkable emotional stake in the proposition [p] [...] [then S] can quite safely assume that his judgement is not a piece of self-deception."

White (1988, p. 41) notes, "[S's] do not have the same practical concerns as [O's], because the judgements they are making do not relate to their own behaviours [...] there is less likelihood that accuracy will be low on their list of priorities." We can see that there will be more affective involvement for the O's who have after all been responsible for the behaviour in question than for the S's who are more dispassionately explaining it. Also, as Goldman (1989, p. 167) observes: the ST "approach can certainly insist that most simulation is semi-automatic, with relatively little salient phenomenology." Goldman is countering the objection that if ST is correct, then we should spend more time than we do experiencing vividly what it is like to be in others' shoes, but his point also supports the line I propose here. It might be that one of the conditions of making ST semi-automatic —which is needful given the requirement for efficiency —is that some of the elements, like bias-modelling or full affect simulation, not always be run. As D. M. Peterson and Riggs (1999, p. 82) point out, on "evolutionary grounds, it is plausible to consider strategies which involve minimum processing load." So the S's might need to exhibit Affect Mismatch on occasions, purely on efficiency grounds.

The proposal is not that the correct bias cannot be added to the simulation; merely that it often is not. As Gordon (1992, p. 20) writes, if you turn back on a country trail because you see a grizzly bear, you may be puzzled by your companion's standing her ground and taking out her pencil and notebook, unless you previously " 'prep' yourself with the appropriate intrepid naturalist attitudes and desires." The reasons you do not generally do this may derive simply from the additional cognitive load involved. As Gordon (1992, p. 25) goes on to observe, it may be that "readiness for simulation is a prepackaged 'module' called upon automatically;" that would be consistent with evidence

(Samson et al. 2010) suggesting that modelling just the perspectives of others is mandatory, fast and effortless but more complex ToM tasks involve significant cognitive load.

As Heal (2000, p. 16) notes, errors in simulation may be because S and O differ in "the degree of stress they are under in thinking of the problem." This view leads to a testable prediction of the Bias Mismatch Defence, which is that people with more active imaginations —who are perhaps more able to experience O's affect vicariously —would be less susceptible to Saxe's occasional systematic errors in ToM than others. The view I propose also allows for the relatively high success rate of our folk psychology: in the majority of everyday situations, there is not that much affect involved for either S or O; the lack of full bias modelling makes no difference to the outcome of the situation. This also explains part of why we find unpredictable people disconcerting.

The condition known as Williams Syndrome (WS) provides further evidence available for a link between affective nature and ToM ability. Segal (1996, p. 154) notes the following characteristics of WS: "average IQ of around 50;" "general impairments [in...] acquisition of [...] theoretical knowledge;" "high degree of social skills" combined with good ToM capacities. The social skills are most notable in the syndrome: Bellugi et al. (2007, p. 99) note that the "WS personality is characterised by hyper-sociability, including over-friendliness and heightened approachability toward others." This sociability will be driven by heightened enjoyment of social situations by WS subjects. They are therefore high affect individuals, when interacting socially. It is suggested by Bellugi et al. (2007, p. 100) that social ability and affective involvement go together when they note that WS children's stories "contained significantly more social and affective evaluative devices" than those of controls. We can see then that empathetic abilities can compensate in ToM for impaired intellectual capacities. The WS subjects are able to develop good

Table 7.1: Simulation Error Probability By System Type Of S And O

	S: System 1	S: System 2
O: System 1	Medium; S and O maybe different biases; but perhaps this covers much 'good enough' everyday ToM	High; frequent source of error in many situations where O is under more pressure than S e.g. Shoppers
O: System 2	Very High; quick simulation of slow reasoning; perhaps this is infrequently applied because ineffective	Low; any occasion where S rationally follows O's rational processes

ToM capacity despite impairments in their theoretical abilities; which makes it look like affect is more important than theory in ToM.

7.3 System Mismatch

A further illustration of occasions when the Bias Mismatch Defence may apply can be given by considering Dual Process Theory (Sloman 1996). Dual Process Theory postulates that there are two reasoning systems that persons use: System 1 and System 2. System 1 is quick and dirty; System 2 is more likely to produce the right answer but takes longer. It seems clear that if a particular episode of reasoning by O is simulated by S and S simulates in a different system to the one that O used, there will be ToM errors. Since System 1 basically is just a set of heuristics and biases, then this approach is another application of the Bias Mismatch Defence.

Table 7.1 shows rough estimates of how System Mismatch might allow for different simulation error probabilities depending on which of System 1 and System 2 are employed by S and O.

Three objections present themselves. The first objection urges that it is not plausible to claim that ToM can take place in different systems. The second objection claims that there is no difference in biases obtaining in System 1 reasoning and System 2 reasoning. The third objection claims that this approach over-generalises: it predicts too much ToM error. I will cover each objection in turn.

Defeating the first objection involves showing that ToM use takes place in both systems. Many commentators have described approaches to ToM which include two levels of processing. These map easily on to the System 1/System 2 division. I will provide three examples. Butterfill and Apperly (2013, p. 609), cite developmental and theoretical evidence to support the claim that "adults may enjoy efficient but inflexible forms of theory of mind cognition in addition to the full-blown form." In a second example, Goldman (2006, Ch. 7) argues for a division of simulation into low-level and high level forms. The quick and automatic simulation in System 1 might include the 'emotional contagion' that takes place when S observes O smiling. The more involved System 2 form of simulation might be more complex and explicitly conscious, though it need not be. As a third example, we may consider an episode of ToM implementation in System 1. Kahneman (2011, p. 91) notes that people judge competence by considering facial features such as "a strong chin with a slight confident appearing smile." Someone using their System 1 ToM will therefore predict competent behaviour by a person with such features. Naturally, it is not the case that facial features are a good predictor of competence. Someone using System 2 ToM might be aware of that. So every time O makes a judgement about competence using System 1 and S uses System 2 ToM to predict what judgement O makes, we should expect ToM error.

Someone holding the second objection can admit that ToM takes place in both systems, but deny that this will lead to bias mismatch and hence also

deny that system mismatch can lead to ToM error. Defeating this objection involves showing that different biases do indeed apply in the different systems. Biases as a whole are more prevalent in the System 1 mode; indeed the prevalence of biases is definitional of System 1: Kahneman (2011, p. 81) writes that "the confirmatory bias of System 1 favours uncritical acceptance of suggestions" meaning that Confirmation Bias is a central method of System 1. More broadly, we may regard Tversky and Kahneman's entire research effort as being a heuristics and biases programme.[3] Further, experiment shows that the application of various logical errors including the Conjunction Fallacy all resulted from System 1 based processing (Sloman 1996, p. 15).

The existence of the two systems explains how a person can reach different conclusions about the same question at different times, even if all the input data is identical. The selection between the two systems is driven by the difficulty of the question to be answered. Persons might use System 1 to decide on lunch arrangements, and System 2 to perform a more complex decision, or one that matters more. System 1 is adequate for the simple recall of when and where one is meeting someone for lunch; nothing very important depends on it. A more complicated task such as deciding how to get to the lunch engagement might engage the more complicated and rational System 2, especially if some selections have to be made between competing transport options involving some view of the weather and traffic conditions. If one is simply walking to the lunch, then again System 1 will likely be up to the task: many people have experienced walking somewhere 'on autopilot' and then noticing that they should have been walking to someone else's office rather than to the lunch venue, for example. This represents a phenomenological confirmation that different systems of reasoning exist and that using one to predict the output of the other will often fail.

[3] Cf. Nagel (2011, p. 8).

Defeating the third objection involves showing that the system mismatch account does not over-predict error. We can examine this by considering table 7.1. Making out the objection would involve showing that the cells in the table that predict a high probability of ToM error obtain much more often than those that do not. This cannot be done because calculating the relevant frequencies requires a method of counting occasions of ToM use, which in turn requires individuation criteria for separate episodes of ToM use. Such individuation criteria cannot be provided. It is also true that on many occasions the two systems will give the same answer. If that were not the case, then the error rate of System 1 would outweigh the value of its speed and light use of resources. So there will be some non-zero rate of system mismatch which does not result in ToM error, which will presumably also be cases where there is no bias mismatch.

A further prediction of this view is that the most accurate simulations will take place when both S and O are employing System 2. So we have another situation in which this account does not predict ToM error. This can easily be seen to be the case by recalling the example of Harris (1992). If two competent English speakers A and B are asked to decide which of a set of sentences are grammatical and which are non-grammatical, a clear result will become apparent. A will predict that B will make the same classification as A of sentences into the grammatical and non-grammatical categories, and this prediction of A's will be correct. Both A and B are using the same system; one can agree with this whether one believes that grammatical analysis is done explicitly using System 2 or has been automated into System 1. Absent reasons to think that S and O are employing different systems, the account predicts no ToM error.

Even if S and O are using different systems, the account might still predict an absence of error to the extent that both systems are good at producing

correct grammar, as seems plausible. It is also the case that persons differ significantly in various parameters that can affect ToM performance. For example, any of S's reasoning capacities, specialities, interests, and available executive function may differ from that of O. Also, the example given above may be a special case, because grammar is perhaps modular. All of these questions appear open to empirical investigation; the results of which would shape the resulting most plausible form of the Bias Mismatch Defence in its system mismatch incarnation and I suspect also strengthen it.

I conclude that the three objections to the idea of system mismatch can all be defeated and system mismatch leading to bias mismatch is a plausible explanation of ToM error for simulationist accounts.[4]

7.4 Mismatch Interactions

At this point, it will be useful to set out how Affect Mismatches can interact with System Mismatches to produce simulation errors. We will be interested in different routes to simulation error; and in predicting when simulation error is likely and when not. In Figure 7.1, the routes to systematic simulation error are shown. At this stage, these are mere template routes for occasions when Bias Mismatch could occur. How these templates work will become clearer in the next chapters when I illustrate examples of these routes in use.

Note that the dashed line is dashed merely to assist with the comprehension of the diagram rather than being a significant element of the argument. The dashed line is the 'yes' line from box 4 to box 3. It is dashed to distinguish it from the other two lines it crosses.

On the Bias Mismatch Defence I propose, whenever there is a Bias Mis-

[4]Other mechanisms producing bias mismatches that lead to systematic ToM errors can be imagined. One example would be the different cultural pressures on males and females; cf. Fredrickson and Roberts (1997).

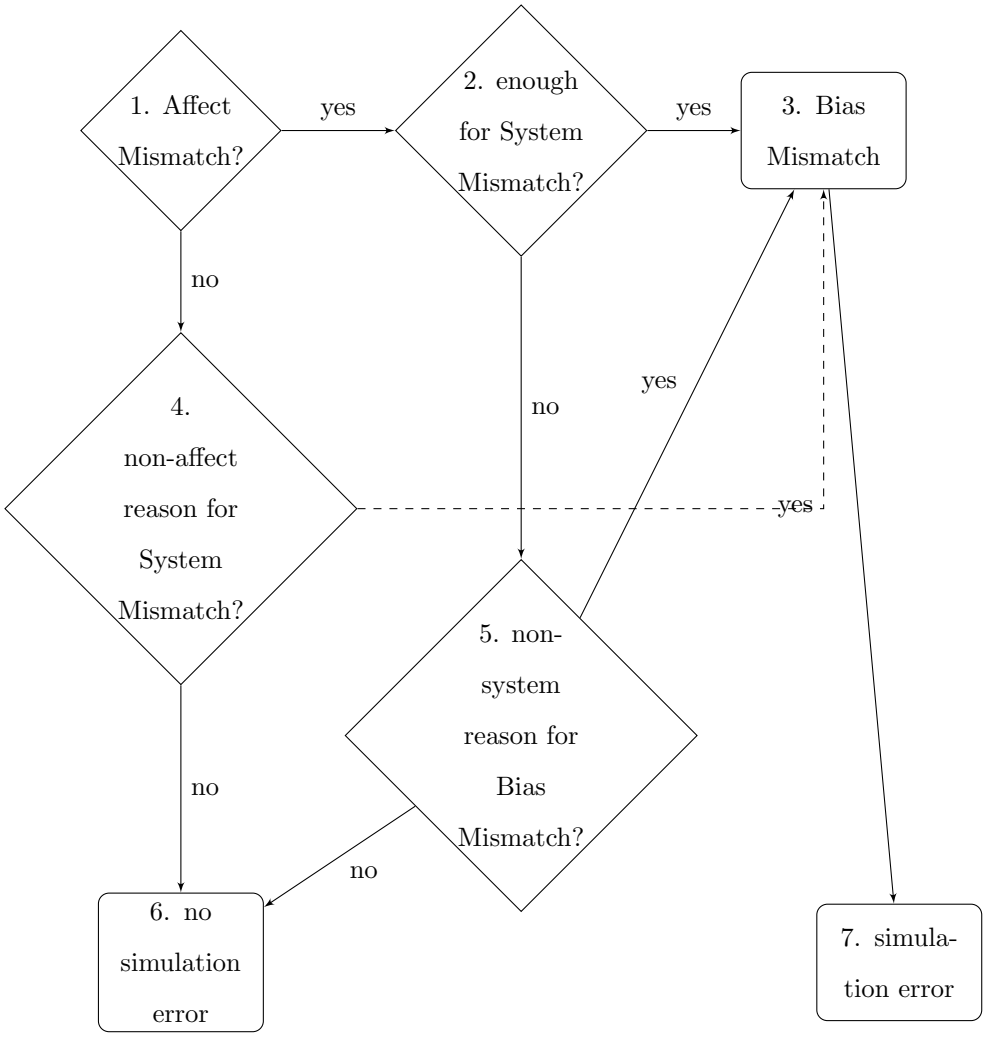

Figure 7.1: Systematic Simulation Error Routes

match between S and O there will be systematic simulation error. It can be seen that there are three routes to Bias Mismatch and so three routes to systematic simulation error. There are two routes which do not pass through Bias Mismatch and so do not result in systematic simulation error. I will outline these five routes through the diagram below. Each route is named by the sequence of boxes through which it passes. In each case of simulation error, it will be systematic because the Bias Mismatch will take place systematically. These five paths exhaust all possible complete routes through the diagram.

For each route, I state whether simulation error or no simulation error will result. In each case, I refer only to whether the route itself has resulted in error or not. It is also true however that different S's can differ quite a lot in their reasoning capacities, and specialisms, and interests, and available executive function. So some S's will be more likely to commit systematic errors than others and also the same S may perform differently at different times. For example, S may commit more errors when lacking available executive function. These caveats do not affect the bias mismatch idea I am outlining though.

- 1/2/3/7
  - There is an Affect Mismatch between S and O
  - This mismatch is significant enough to cause a System Mismatch between S and O
  - The System Mismatch causes a Bias Mismatch between S and O
  - This route results in systematic simulation error
- 1/2/5/3/7
  - There is an Affect Mismatch between S and O
  - This mismatch is not significant enough to cause a System Mismatch between S and O
  - There is nevertheless a Bias Mismatch between S and O, even though they employ the same system
  - This route results in systematic simulation error
- 1/4/3/7
  - There is no Affect Mismatch between S and O
  - There is nevertheless a System Mismatch between S and O, with non-affective causes

  - This route results in systematic simulation error

- 1/4/6
  - There is no Affect Mismatch between S and O
  - There is no other reason for System Mismatch
  - There is no System Mismatch
  - There is no Bias Mismatch
  - There is no simulation error

- 1/2/5/6
  - There is an Affect Mismatch between S and O
  - The Affect Mismatch does not suffice to cause System Mismatch
  - There is no other reason for System Mismatch
  - There is no System Mismatch
  - There is no Bias Mismatch
  - There is no simulation error

In the next two chapters, I will outline some of the situations where there is systematic error in ToM and explain how Bias Mismatch between S and O explains the errors in ToM. Saxe suggests a number of relevant circumstances. In some situations, we are 'too rosy' about the reasoning capacities of others and in other types of situation we are too cynical. I will cover both in turn.

Chapter 8

'Too Rosy' Evidence

8.1 Introduction

Saxe (2005a) cites Gilovich (1993) as one source of much of the data we will consider in this chapter. It all points to scenarios in which S's are systematically too rosy in their ToM. They predict that O's in the situations described will act more rationally, not to say ethically, than they do. These predictions will not be borne out, as we will see. Gilovich (1993, pp. 9-10) explains that the basic project of his book is to ask why "questionable and erroneous beliefs are learned, and how they are maintained." The fact that the beliefs are 'questionable' tells us that there are ToM errors involved. If the beliefs were not questionable, then they would presumably be predicted more often.

We will be interested in any biases that Gilovich cites as explanations of the questionable beliefs, because my proposal is that absence of specifically those biases in S at the time of simulation and as part of the simulation is what accounts for the surprise or the failure of ToM. Naturally I do not claim that S is free of the biases displayed by O; merely that the same biases are not triggered in S or used as part of the simulation because S is not actually in O's situation. That means that the full affective import of O's situation

is not felt by S, or S and O may employ different systems of reasoning. So there can be Affect Mismatch or System Mismatch between S and O, leading to Bias Mismatch and systematic simulation error.

Even if O's become motivated to remove their cognitive biases, this is very difficult. Tversky and Kahneman (1973, p. 218) found that undergraduates offered $1 —a significant amount in 1973 —to answer a mathematical problem correctly by avoiding the Availability Heuristic, did not do so. They conclude that: "[e]rroneous intuitions, apparently, are not easily rectified by the introduction of monetary payoffs." Failure of the S's to simulate cognitive biases in the O's will be a hard-to-remove source of systematic ToM errors in the S's, even when the O's might be expected to be trying hard to remove such biases. The difficulty of removing such biases may sometimes cause ToM errors in the other direction as well: S may apply his own biases. For these reasons, there are many ways of arriving at a mismatch in bias status in S as compared to O, and this will cause simulation errors.

A possible objection here derives from the fact that my account admits that there is wide-spread error in human cognition. It may be asked how this is possible, if our cognitive systems have evolved to help us survive. I will not address this issue at length, but merely outline the directions of two responses. Firstly, it is clear that we have the biases, and many of them. That is not necessarily irrational, because they save time and we simply do not have enough time or the inclination to consider every question that faces us on a daily basis with the maximum possible cognitive effort. Often, it is better to act on a 'wrong' decision and see what happens than agonise indefinitely. So it is clear that our thinking is not supposed to be even aimed at being error-free. Secondly, I might appeal to arguments of the sort raised in detail by McKay and Dennett (2009, p. 493), to the effect that some "misbeliefs" are "best conceived as design features. Such misbeliefs, unlike occasional lucky

falsehoods, would have been systematically adaptive in the evolutionary past." They give as examples unrealistically positive false beliefs about the self, which improve performance. Note that this could also suggest evolutionary grounds for systematic ToM error, since S employs ToM about S as well as O.

On the other hand, there will be many occasions when ToM succeeds because there is no Bias Mismatch, perhaps because there is no significant affect in either S or O. Or if tasks are selected such that S and O use the same system of reasoning, such as with the Harris (1992) grammar task, then simulation may proceed without error. My account also predicts that there should be occasions of successful simulation by bias matching, which should be empirically testable. It will be important though to ensure that S and O are not just employing the same bias. They would also need to be employing the same bias about the same data. S and O could well both be applying Confirmation Bias, for example, but unless they started with the same beliefs, that would not lead them to seek erroneously to confirm the same prior belief. Very careful experimental design will be needed here.

There can be two forms of evidence for ToM errors, which I will term 'hard' evidence and 'soft' evidence. Hard evidence will be constituted by statistical data on the ToM errors, of the form '75% of S's did not predict O's decision.' This will be the most important data. The softer form of evidence will be where no percentages are given, but we are surprised by the questionable beliefs. The surprise indicates that we have failed to predict the belief. This softer evidence, while still valuable, may suffer from the twin defects that surprise is both subjective and varies from mild to extreme. Moreover, many of the people likely to read this thesis will have extensive knowledge of the frailty of human reasoning and therefore be unsurprised by any revelations concerning it. It is possible that such S's are not using their ToM at all: they are merely consulting their relevant experience. The hard evidence will be

Group Studied	Response
Shock Appliers	(A): O exhibits Conformity Bias
Fake Prison Guards	(A): O exhibits Conformity Bias
Repenters	(B): S exhibits False Consensus Effect
Quiz Gamers	(B): S: Fundamental Attribution Error
Suicide Note Assessors	(A): O exhibits Belief Perseverance Bias
Lottery Ticket Holders	(A): O exhibits Endowment Effect
Gamblers	(A): O exhibits Confirmation Bias
Basketball Fans	(A): O exhibits Clustering Illusion
Cancer Cure Assessors	(B): S exhibits Confirmation Bias
Puzzle Solvers	(A): S exhibits Availability Heuristic
Shoppers Redux	(A): O exhibits Position Effect

Table 8.1: Response Type By Group Studied: Too Rosy

covered in the earlier sections with the soft evidence to follow.

Table 8.1 shows the Bias Mismatch response I will give in §8.2 to each of 11 cases discussed in the works cited by Saxe (2005a). I label the cases by the group of people studied. Some of the experiments have already been touched on previously. The explanations fall into two broad categories, as set out below.

- (A): O exhibits bias which is not simulated by S [eight entries]
- (B): S exhibits bias when simulating [three entries]

8.2 'Too Rosy' Data

8.2.1 Shock Appliers

The Milgram (1963) experiment was introduced in Ch. 5. The results include valuable hard evidence because there are some numerical data concerning S's confounded expectations of the likely behaviour of O's. Recall that $26/40 = 65\%$ of O's set the dial to 450 Volts while the psychology undergraduate S's predicted that that number would be 3% at most.

The Bias Mismatch Defence of ST that I propose must now explain this failure to predict. I will do this by noting the significant Affect Mismatch between S and O. The S's, whether ourselves or Yale seniors, consider the question as to how much they would be prepared to shock in a relatively calm, reflective state —precisely one arranged for the seniors so that they could "reflect carefully" (Milgram 1963, p. 375). The S's are not this instant under pressure from an authority figure in a lab coat, issuing stringent instructions. This is what makes the difference, as is confirmed by the summary of Plotkin (2011, p. 459), who writes that "Milgram repeatedly demonstrated how people without any history of cruelty or violence would, when ordered to do so by a figure of authority, inflict violent punishment upon others." We may also imagine that the effects of stress deriving from deference to authority would be much less in modern times than in 1963. All of these factors imply that we are unlikely to apply or to simulate the cognitive bias that tends to make us more obedient than we should be. We as S's fail to simulate the Conformity Bias of the O's, just as the carefully reflecting seniors of Milgram (1963) did.

Three questions may arise here. The Conformity Bias is often seen as being about conforming *judgements* to a reference group. Is the Milgram effect really about that? Is it driven by people judging that 450V is appropriate? And who is in the reference group that makes this judgement? In response to the first question, certainly the Milgram experiments are seen as part of the Conformity

Bias literature. Plotkin (2011, p. 459) confirms this when he writes that "the most powerful demonstration of what social psychologists call conformity, obedience or group cohesiveness was reported in a series of papers in the 1960s, summarized by Milgram." Also note the usage of the term Conformity Bias (Plotkin 2011, p. 459). Considering the facts of the experiment, one might also wonder what, if not Conformity Bias, can cause the O's to apply, as they think, high voltages. What other motivations do they have? Turning to the second question, the judgement that is conformed is not directly the one to the effect that '450V is appropriate.' It has that effect, but it is more appropriately phrased as approximately 'people in experiments will obey the instructions of the experimenter' perhaps combined with 'experimenters will not make unreasonable requests.' We can see this because we can confidently expect that similar results would be obtained in variants of the Milgram experiment which, for example, have the O's causing the hands of the dummy participants placed in cold water. Finally, we might say that the reference group is an imaginary one in the minds of the O's. It is as it were a creation of the O to represent 'how people generally behave in this situation.' This is distinct to the more usual situations where the reference group is actually present. As I will mention again below, this effect is outweighed when an actual reference group is present, lending strength to the assumption that we are in fact dealing with a reference group effect here.

In accordance with the framework set out previously, I will now explain why a Bias Mismatch arises. On this occasion, there is extreme Affect Mismatch between S and O. The S's are undergraduates sitting calmly, observing in a clinical fashion in the company of their distinguished professor. Nothing much hangs upon what the S's say or do in relation to the experiment; they are expected to make useful psychological comments. Nothing about the calmness and lack of involvement of the S's is true of the O's. As Milgram (1963, p.

375) writes, many of the O's exhibited extreme affect: "the degree of tension reached extremes that are rarely seen in socio-psychological laboratory studies. [...] Fourteen of the 40 [O's] showed definite signs of nervous laughter and smiling. [...] Full-blown, uncontrollable seizures were observed for 3 [O's]." There is a very clear affective mismatch between S and O; this explains the absence of the appropriate bias in the simulation which explains this failure of ToM. Milgram (1963, pp. 375-376) even makes remarks suggesting this in interpreting the results he obtained from the Yale seniors: 'it is possible that the remoteness of the respondents from the actual situation, and the difficulty of conveying to them the concrete details of the experiment, could account for the serious underestimation of obedience."

The hidden S's watching the experiment "often uttered expressions of disbelief upon seeing a subject administer more powerful shocks to the victim" even though the S's "had a full acquaintance with the details of the situation" (Milgram 1963, p. 377). Since these observer S's were relatively sophisticated associates of Milgram —'senior psychology majors' —we may presume that they were much less subject to the bias towards obedience. Or they may have been subject to a different strain of Conformity Bias in that they felt pressure to side with Milgram in his capacity as dispassionate observer. In any case, it is clear that the observers faced much less pressure than the O's, and in fact may have felt countervailing pressure to behave 'clinically.' So we are still entitled to conclude that Affect Mismatch driving Bias Mismatch causes the failure of ST here.

As the Editors (1992, p. 9)[1] write in their Introduction to the special double issue of Mind and Language, "I might conclude, after deliberating, that I would not behave sadistically in the notorious Milgram experiments. But, I might also be convinced, on the basis of scientific evidence, that the

[1]The Editors are not named but thank Tony Stone for his assistance.

balance of probability is that I would, in fact, so behave. Here, I will be baffled by the prospect of my action." The Bias Mismatch Defence I propose in this thesis eliminates this bafflement. The calm and unhurried simulation of the S's will not reflect at all the various extreme affects O's would face in the Milgram (1963) experiment. The S's would then not suffer from Conformity Bias at all or in the same way as the O's, and so S's simulations would fail.

There is good reason to think that it is Conformity Bias that causes the unexpected behaviour of the O's, because if the behaviour around them changes, so does theirs. In a variant of the standard experiment, known as 'experiment 17' or 'two peers rebel' (Milgram 1974, pp. 116–121) the role of teacher was split into three with two of the other teachers also being confederates of the experimenter. As the name suggests, two of the peers rebelled, which caused 92.5% of actual subjects to also rebel. It looks quite convincing then that what has happened is that a new more vivid reference group has been created which moreover gives the O's 'permission' to follow their consciences and reject the experimenter's instructions.[2]

This line must account for the observations made in a re-enactment described by Ross, Amabile, and Steinmetz (1977, p. 492). They write that "Bierbrauer (1973) [...] showed that even after personally participating in a verbatim reenactment of the classic Milgram (1963) demonstration, raters consistently and dramatically underestimated the extent to which Milgram's subjects would administer dangerous levels of electric shock in accord with the situational forces compelling 'obedience."' My response will be that a 'verbatim reenactment' is still not close enough to the real thing to make it count, affectively, for the S's. Stich and Nichols (1995a, p. 102) describe the

[2]Although note that Goldie (2002, p. 164) attributes the failure to predict behaviour in the Milgram (1963) experiment to the Fundamental Attribution Error. If he is right, a loose variant of the Bias Mismatch Defence still succeeds. (It is loose only because it would strictly speaking also require a specification of FACTOR.)

aim of the Bierbrauer reenactment as "to study the predictions that would be made by non-participant observers." Again, an observer S that is not a participant will not exhibit Conformity Bias to the same extent and about the same cognitions as participant O's.

A difficulty is raised for my account when Stich and Nichols (1995a, p. 102) note that there were still failures of ToM by S's when "they themselves played the role of a subject in a vivid reenactment." The force of this seems to be that there should not be simulation errors when S and O are identical, because S and O have the same mental machinery. However, two differences are apparent. This objection might suggest that S's should avoid simulation error in relation to themselves as O. This may not be true; it depends on how much 'playing the role' provides the full affective import of actually being in the Milgram (1963) experiment. I suggest that 'playing a role' is very different from 'being' the role. But in any case, the S's were here asked to predict the behaviour of other O's. The idea was that placing them in a closer situation to the one of the O's would give them better insight into the behaviour of the O's. But this will not work at all if there is still scope for Affect Mismatch, which I contend there very clearly is.

In general, many occasions of ToM error arise when S's are given the salient facts and asked to opine on them rationally. This differs dramatically from the affective position of the O's, who simply experience the world without the important, salient or significant facts being given to them as such.

It will be useful here to reiterate the structure of my approach, as initially set out on p. 155. I make two claims which may be approximately phrased as follows.

1. FACTOR X affects O but not S

2. FACTOR X modulates the probability of being subject to BIAS

One might think that early in this section it looks like FACTOR X is pressure of some kind but be concerned that here it seems like FACTOR X is merely being in the situation as opposed to observing it. I also therefore link both of these two items i.e. being in the situation inevitably involves more pressure than observing it. Also I aim here to avoid the less informative idea that biases are not simulated. The Bias Mismatch Defence is better described as being based on the claim that different factors can make O subject to a bias or not, and that it is this that causes the simulation errors. Note that it could also be that FACTOR affects S but not O; that could equally well produce a bias mismatch and a simulation error. As I explained previously, strictly speaking, the latter case would not be an exact instance of the Bias Mismatch Defence, since the formal structure above requires the FACTOR to affect O but not S. Likewise, any situations in which FACTOR is not specified are only loosely to be classified as instances of the Bias Mismatch Defence. I nevertheless maintain that all of the loose and strict instances can be of value in explaining ToM error.

This explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (affect: pressure from being in the experiment) and BIAS in the O's (Conformity Bias) leading to systematic ToM error.

8.2.2 Fake Prison Guards

Although Saxe (2005a) does not cite the infamous Stanford prison experiment, it is often considered together with Milgram (1963) as providing evidence of unexpected behaviour which may result from excessive deference to deemed authority. The prison experiment had a very simple design. A mock prison was constructed, and the O's were randomly assigned the role of guards and prisoners. The O's answered an "extensive questionnaire" designed to select

those who were "most stable (physically and mentally), most mature, and least involved in anti-social behaviour" (Haney, Banks, and Zimbardo 1973, p. 73). The O's were all male college students who were mostly middle class. The guards were given the instruction to "maintain the reasonable degree of order" needed for the "effective functioning" (Haney, Banks, and Zimbardo 1973, p. 74) of the prison, without being giving further specific instructions as to how this was to be achieved.

The results were that the guards were far more aggressive than expected. As Haney, Banks, and Zimbardo (1973, p. 69) write, "[a]t least a third of the guards were judged to have become far more aggressive and dehumanising toward the prisoners than would ordinarily be predicted in a simulation study." This is the hard evidence of ToM error; the authors as S's judged the behaviour of the guards as O's and were wrong about a third of them. Note that the experiment was a simulated prison which had no legal authority to hold persons; it was time-limited and yet dramatic and unexpected behaviours were observed. S's will generally fail to simulate O's accurately here in that they will expect that O's in the situation will not display marked aggression in the case of guards and marked passivity in the case of the prisoners.

Despite the apparent normality of the O's and the lack of instructions tending towards this outcome, "the characteristic nature of their encounters tended to be negative, hostile, affrontive and dehumanising" (Haney, Banks, and Zimbardo 1973, p. 80). A high proportion of the ten prisoners experienced extreme affect: "five prisoners [...] had to be released early because of extreme emotional depression, crying, rage and acute anxiety" (Haney, Banks, and Zimbardo 1973, p. 81). One prisoner even developed a "psychosomatic rash" (Haney, Banks, and Zimbardo 1973, p. 81). The guards on the other hand "enjoyed the extreme control and power they exercised" (Haney, Banks, and Zimbardo 1973, p. 81) and "on several occasions [they] remained

on duty voluntarily and uncomplaining for extra hours —without additional pay" (Haney, Banks, and Zimbardo 1973, p. 81).

The completeness of the failure here of ToM may be gauged from Haney, Banks, and Zimbardo (1973, p. 81) writing that these "differential reactions to the experience of imprisonment were not suggested by or predictable from the self-report measures of personality and attitude or the interviews taken before the experiment began." In fact, the study had to be terminated early after six days because of the "unexpectedly intense reactions" (Haney, Banks, and Zimbardo 1973, p. 88) generated. I will propose that the same explanation may be applied to Haney, Banks, and Zimbardo (1973) as was applied previously to Milgram (1963).

There will be significant affective mismatches between persons outside the experiment who are asked to use their ToM to predict the behaviour of O's in either role in the prison experiment. These mismatches permit the introduction of my proposed Bias Mismatch Defence, to the effect that such affective mismatches cause failure of ToM through inadequate simulation of biases exhibited by O's. There will be different Affect Mismatches between S's and the two subsets of the O's. The failure to simulate the guards accurately will be due to not fully reflecting their enjoyment of power, because once again actually enjoying that power is much more affectively engaging than merely simulating it. The failure to simulate the prisoners accurately with be due to not fully reflecting their depression and rage. The main bias active here in both the prisoners and the guards is again Conformity Bias. The prisoners conform with each other and defer to the guards. The guards also conform with each other, or perhaps more accurately we might say that they conform with the imagined group harshness. Although it is true that only a minority of guards exhibit extreme harshness, this minority will be setting the group standards in a vivid way, going against which would require significant courage, even if

it is the majority opinion that the harshness is excessive (Prentice 2007). So we have Affect Mismatch driving Bias Mismatch resulting in faulty simulation and systematic ToM error.

This explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (affect: fear in the prisoners and enjoyment of power in the guards) and BIAS in the O's (Conformity Bias) leading to systematic ToM error.

8.2.3 'Repenters'

We might expect that people will generally believe that other people agree with them only when there is some reason for that belief, for example testimony to that effect or perhaps polling data. However, Gilovich (1993, pp. 112-113) points out that there is in this respect a "systematic defect in our ability to estimate the beliefs and attitudes of others" whereby we "often exaggerate the extent to which other people hold the same beliefs that we do." This is held to be evidence for a failure in ToM because if people simulated accurately, they would not predict the presence of this agreement where it is absent. However, it can also be seen as positive for the ST or Weak S/T Hybrid accounts since they predict such default belief attribution while TT accounts do not, as discussed in Ch. 3.

In fact, Gilovich goes so far as to see the 'imagined agreement' problem as underpinning the wide variety of false beliefs he discusses, since maintaining the false beliefs without the imaginary agreement of others would be much more difficult. On that line, almost all false beliefs would represent a failure of ToM. However, as Gilovich (1993, p. 118) also points out, "we associate primarily with those who share our own beliefs, values, and habits;" insofar as that is true, we would not have a failure of ToM here at all.

Gilovich (1993, p. 114) cites Ross, Greene, and House (1977) to provide

data where we can see this failure. An experiment was conducted in which students were asked if they would be prepared to wear a large sign around campus bearing the legend 'REPENT.' A "substantial percentage" agreed. The critical question was then asked as to what percentage of their fellow students did they think would also agree to do so. It transpires that student S's thought, roughly, that their peer O's would decide as they had. Ross, Greene, and House (1977, p. 292) report that S's who agreed to wear the sign thought that 63.5% of peer O's would also agree while S's who declined to wear the sign thought that only 23.3% of peer O's would agree to wear it. This of course, is explained by the False Consensus Effect introduced in Ch. 6. The simulation of the S's is derailed by their exhibiting the False Consensus Effect.

Care must be taken here to avoid my account merely saying the following: – BECAUSE bias THEREFORE incorrect simulation – since the correct structure is instead approximately: FACTOR X distinguishes S from O – and – FACTOR X modulates the probability of being subject to BIAS. So what is FACTOR X here? My general claim is that the underlying reason for this and many cases of lack of simulation of cognitive bias is affective mismatch between S and O. Of interest then is the explanation offered by Gilovich (1993, p. 114) for the False Consensus Effect. There is a basic desire to "maintain a positive assessment of our own judgement" which is particularly likely to play a part when we "have an emotional investment in the belief." So, S's emotions can promote the False Consensus Effect. If S's believe particularly passionately that there will be a woman supreme court judge in the next ten years —one of the Ross, Greene, and House (1977) test questions —those S's are more likely to give a mistakenly high estimate of the number of O's who share that view. The effect is reduced or eliminated when the S's do not particularly care about the belief in question. When the test question was about the number of hours

of TV watched a week, S's were less likely to think O's were like them. So the False Consensus Effect is exacerbated in relation to beliefs that really matter to the S's. In other words, we have here once again Affect Mismatch driving Bias Mismatch. In this particular experiment, S's cared about the O's being like them, in that S's who agreed and thought that O's also would tended to view the exploit of wearing the sign as amusing. On the other hand, S's who did not agree and tended to think that the O's would also not agree tended to view the exploit as instead one only liable to be agreed to by compliant patsies.

Several experiments outlined by Pronin, Puccio and Ross may all be explained on the basis outlined above. Pronin, Puccio, and Ross (2002, p. 642) report on a similar experiment, in which the difference was that the sign read 'EAT AT JOE'S.' S's who agree to wear the sign believe that O's will agree more than they do, and those who do not agree think that O's will agree less than they do. The False Consensus Effect in S's results in systematic simulation error.

Similar results were obtained when S's were asked "whether they preferred music from the 1960s or the 1980s" (Pronin, Puccio, and Ross 2002, p. 642) and what percentage of their peers would agree. S's said that most people would agree with whichever choice they made, even though there was in fact more of an even split between the two choices. This further occurrence of the False Consensus Effect in S's results in systematic simulation error.

Pronin, Puccio, and Ross (2002, p. 643) discuss three further experiments conducted in their own lab which they say all show the False Consensus Effect. The first one was on "Encoding and Decoding Musical Tapping." S tapped out the rhythm of a well-known tune for a listening O who had to identify the tune. S was then asked to estimate the probability that O would recognise the tune. The results were that S's vastly overestimated how easy this would be,

because S can 'hear' the tune internally, and does not adjust for the fact that O cannot. S's thought that success rates would vary between 10% and 95%, with an average of 50%, while the actual success rate was a mere 3%. This experiment looks like a slightly unusual illustration of the False Consensus Effect in that it might be best termed as the generalisation 'everyone knows what I know' rather than 'everyone believes what I believe.' In some ways, it seems to be an adult version of failing the False Belief Task. We might explain the effects by appealing to the Availability Heuristic in its vividness incarnation. S's find the tune so vivid in their own mind that they are simply unable to simulate the total lack of vividness the tune has in the minds of the O's, who just have some dull tapping to interpret. Thus, there is a Bias Mismatch between S and O and so the Bias Mismatch Defence is available. The bias in question is the Availability Heuristic applied by S which is not applied by O for the obvious reason that the tune is not in fact highly available to O since it is unknown.[3]

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (emotional investment in the belief) leading to BIAS in the S's (False Consensus Effect).

8.2.4 Quiz Gamers

Ross, Amabile, and Steinmetz (1977) investigated the assessment of general knowledge by persons participating in one-on-one quiz game scenarios. The

[3]Some observers may maintain that there is nothing for ST to explain here, since there is just an information asymmetry between S and O, which on ST naturally causes S to make errors about O. If that is true, then ST is not harmed but TT accounts, especially modular ones like the versions of TT(Innate) so far described, need to explain how S's belief set in relation to the tune becomes available to ToM.

idea being investigated was 'social control,' meaning inequalities of power in social settings. For example, if one person works for another person, the latter person will have more social power. Another example discussed by Ross, Amabile, and Steinmetz (1977, p. 493) is that of the dissertation viva at a university. Here, the "candidate is required to field questions from the idiosyncratic and occasionally esoteric areas of each examiner's interest" while the "candidate has relatively little time for reflections." We might expect people involved in such situations of social control to take account of it when making ToM assessments. For example, if the candidate assesses the knowledge of the examiner, he should do so including consideration of the advantages of question selection, time and lack of pressure enjoyed by the examiner. Likewise, if the examiner assesses the knowledge of the candidate, he should take account of the corresponding disadvantages to which the candidate is exposed.

Ross, Amabile, and Steinmetz (1977) proposed to investigate the extent to which assessments of social control played a role in evaluation of others. They arranged for participants to pair off; the questioner would set questions and ask them of the contestant. The questioner set questions based on his own esoteric knowledge. Ross, Amabile, and Steinmetz (1977, p. 486) claim that this models many forms of social interaction, where they say that "[o]ne participant defines the domain and controls the style of the interaction and the other must respond within those limits." The questioner, then, is in a position of social control in relation to the contestant. The result, naturally enough, was that the contestants were not very successful in answering the questions, lacking the specialised knowledge of the questioner.

After the questions have been answered, the questioners and the contestants both made general knowledge evaluations of each other and themselves. It transpired that of 24 contestants, "20 contestants rated themselves inferior to their questioners" (Ross, Amabile, and Steinmetz 1977, p. 489). So, the

vast majority of contestants did not allow for the fact that it is much easier to set questions than to answer them. The contestants as S's made evaluations of the general knowledge of the questioner O's that completely failed to take account of the one-sided nature of the data available to them. Ross, Amabile, and Steinmetz (1977, p. 485) conclude that when "drawing inferences about [O's], [S's] consistently fail to make adequate allowance for the biasing effects of social roles upon performance."

This experiment was re-run with observers, who form new S's. The observers produced the same predictions as the participants viz. "[S's] impressions of the [O's] in the quiz game showed the same bias that was evident in the participants' own perceptions. Overall, the questioner is seen as tremendously knowledgeable" (Ross, Amabile, and Steinmetz 1977, p. 491).

What has happened is that the S's of both types have committed the Fundamental Attribution Error. Ross, Amabile, and Steinmetz (1977, p. 491) confirm this when they note that "the phenomenon we have described represents a special case of a more fundamental attribution error" meaning that S's attribute the underperformance of the O's more to the character of the O's than to the actually more important situational variables viz. the difficulty of the quiz questions. S "infers broad personal dispositions and anticipates more cross-situational consistency in behaviour than actually occurs" (Ross, Amabile, and Steinmetz 1977, p. 491).

The authors provide a further bias-related explanation for their data when they suggest that "the various raters' judgements were distorted precisely to the extent that they depended upon biased data samples" (Ross, Amabile, and Steinmetz 1977, p. 493) which would be an example of the Availability Heuristic. This would be because the results of the question and answer session would be highly available since salient yet inaccurate in assessing the general knowledge abilities of participants. S's do not take account of the obvious fact

that they have very little data and it is highly selective. So there are two ways of using biases in S to explain the systematic ToM errors made by S here.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It may loosely speaking be an instance though because we have FACTOR in the S's (affective effects of social power inequalities) leading to BIAS in the S's (Fundamental Attribution Error). It is also open to objectors to insist that FACTOR has not been adequately explained here.

8.2.5 Suicide Note Assessors Redux

This experiment was discussed previously in §6.2.1. The experiment, by Ross, Lepper, and Hubbard (1975) is described by Stich and Nichols (1995a). In this experiment, O's are given a test which "indicates that they are unusually good (or unusually bad) at a certain task [;..] an hour later it is explained to them that the test results were bogus" (Stich and Nichols 1995a, p. 100). (In fact, the delays were either 5 minutes or 25 minutes.) The task in the experiment was to assess whether a suicide note was fake or real. The odd result in the experiment is that O's continue to believe that they are unusually good or bad at the task even when the evidence therefor has been dismissed.

Stich and Nichols (1995a, p. 100) formed a body of S's from among their students and asked them to predict the results of the test. They found that "[t]he predictions the [S's] offered were more often wrong than right" Thus there is some quantification of failure of ToM here; more than half the S's exhibited such a failure. Stich and Nichols adduce this failure as evidence against ST by noting that the students would have exhibited the Belief Perseverance Bias had they taken the test as opposed to been asked to predict its outcome —which we may concede. If so, then they could not have been simulating,

according to Stich and Nichols, because they would not have made the error.

This is of course easily explained on the Bias Mismatch Defence I am proposing. The S's simulations failed because they failed to include the Belief Perseverance Bias of the O's in their simulation of the O's. In turn, they failed to include that effect because they were not in the situation faced by the O's, who had an affective involvement resulting from being told something about their competencies which may have been pleasing or displeasing. There was an Affect Mismatch between S and O and a resulting Bias Mismatch leading to systematic ToM error.

The experimental task in the Ross, Lepper, and Hubbard (1975) experiment is conducted while wired up to electrodes ostensibly intended to measure physiological responses. We may observe immediately that this is not a low affect scenario for the O's. In addition, the O's were randomly assigned to three groups —success, fail, average —and at least two of these will have had some influence on self-esteem which will in turn have had an affective component. This is confirmed by Ross, Lepper, and Hubbard (1975, p. 883) who note that "subjects in the success condition reported having felt more satisfaction than subjects in the average condition" who in turn felt more satisfaction than the subjects assigned to the fail condition.

I mentioned in §6.2.1 that the experimenters also had S's observe the O's who performed the suicide note assessment task, and these S's also exhibited the same Belief Perseverance Bias to some extent. Ross, Lepper, and Hubbard (1975, p. 885) recruited additional experimental subjects who were engaged in "observing and listening to an entire experiment through a one-way mirror." They also exhibit the Belief Perseverance Bias about the ability of the O's —they continue to believe that the 'success' O's are better at the task even after they also learn that the O's did not really succeed. This needs to be explained, because on my account so far, these new S's should not have an

affective involvement in the prowess of the O's. A further bias-related explanation is available here since the Belief Perseverance Bias can also be seen to be a result of Confirmation Bias. Ross, Lepper, and Hubbard (1975, p. 880) note that "once formed, impressions are remarkably perseverant and unresponsive to new input." That line is also suggested by Pronin, Gilovich, and Ross (2004, p. 796). who write "biased assimilation of new information, in turn, leads to unwarranted perseverance of beliefs." The point here then is that the observer S's exhibited Confirmation Bias which introduced simulation error. So we have a further Bias Mismatch explanation of the ToM performance of the new S's. Some observers may feel this may not be an example of the Bias Mismatch Defence, strictly speaking. The important factor is just that S's in the new conditions were *subject* to Confirmation Bias. If so, there is no appeal to *mismatch* of biases in explaining the result, although it is clearly still important to consider biases when examining ToM data. But my line would be that *any* bias in S or O which causes systematic ToM error can open up a form of the Bias Mismatch Defence.

What I have done here is appeal to this Belief Perseverance Bias to explain the discrepancy between prediction and performance in the suicide note assessors case (where people are first given incorrect feedback about their performance, then told the feedback is random, but still persist in believing that they are good/bad assessors of which the genuine suicide notes are). One might be concerned whether this is really an explanation rather than merely a restatement of the experiment. Again, for this reason, FACTOR is important as a motivator, as I explained on p. 186. We have avoided the circularity of the type 'BIAS therefore BIAS' which would indeed have little explanatory value. 'BIAS therefore ToM error' is of more use, but the key chain is 'FACTOR therefore BIAS therefore error.' Here, FACTOR is the affective nature of the competence information that O wishes to retain. This gives the account real

explanatory value. Put another way, the question is, why do people perform one way but predict a different performance if their prediction is based on (something underpinning) their performance? The answer is that FACTOR underpins the performance of O but not of S (in the strict formulation) and FACTOR underpins the performance of S but not of O (in one of the loose formulations).

This explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (affect: pleasure in competence or displeasure in incompetence) and BIAS in the O's (Belief Perseverance Bias) leading to systematic ToM error.

The remaining subsections cover the soft evidence of ToM error.

8.2.6 Lottery Ticket Holders Redux

This experiment was described previously in Ch. 6. The Endowment Effect is the bias of the O's that the S's do not simulate and this is why the S's simulation fails here.

The S's will be uninterested in the outcome of the Lottery, giving them an Affect Mismatch with the O's. Moreover, the S's have the question explained to them in a dispassionate way with the salient points for rational analysis prominent in that explanation. They could then have a System Mismatch with the O's as well. An Affect Mismatch between S and O is also suggested by Kuehberger et al. (1995, p. 429) when they write that "resale values would be lower when given in personal interaction with the experimenter rather than anonymously, since [O's] feel under more pressure of potentially having to justify their price." 'Pressure' means affect for the O's which is not there for the S's. Note also that all of the S's are in the presence of a different experimenter, and will feel different pressures. Together, these factors mean

that the Bias Mismatch Defence of ST which I propose predicts the actual outcome —the S's fail to simulate the O's in that the S's suggest lower, more reasonable, resale prices for the Lottery tickets. The S's feel pressure to be reasonable while the O's are in the grip of the Endowment Effect.

Nichols, Stich, and Leslie (1995, p. 443) respond to the Kuehberger et al. (1995) methodological criticisms by introducing new data. This new data is from an experiment showing that S's failed to predict their own later susceptibility to the Endowment Effect. This is of course exactly what my account predicts, since simulation is the basis of ToM when used for all O's, whether the O is another person or the S at another time. There is a Bias Mismatch between S and O at the later time, even though the O is the same person as the S. It is just that the S is not engaged in the actual situation at the time of simulation and so does not feel its affective import. So there can be an Affect Mismatch between S and O even when O is S at the later time. The objection that simulation cannot explain ToM in cases where S and O are identical ignores the fact that additional distinctions are available between S and O apart from their mental machinery, which we may concede is the same.

This idea is consistent with an observation of Kuehberger et al. (1995, p. 425) who note that "even five minutes of belonging might be difficult to simulate." The elements of belonging that might be difficult to simulate might be the affective elements and the biases that are triggered. This allows one to avoid having my claim here imply that simulation cannot ever enable S to identify the affective import of situations. That would be an unfortunate consequence since S's are often able to do so; earlier I gave examples such as when S's identify a situation as one in which O is likely to feel shame, or to be under pressure. There might be something specific about ownership which impairs affective forecasting, at which people are known to be poor (Sevdalis and Harvey 2007). For instance, persons believe that a new car will make

them very happy and a serious impairment such as becoming blind will make them extremely unhappy. Both of these claims are empirically false, and even repeated new car buying does not repair the ToM error that S makes about S. Another possible escape route here for the Bias Mismatch Defence, should one be required, will be to recall that simulated shame and simulated pressure are but shadows of the real thing. Perhaps sometimes the shadow suffices to match up the biases sufficiently to reduce ToM error and sometimes it does not.

In their response to the criticisms of Kühberger et al., Nichols, Stich, and Leslie (1995, p. 440) claim that ST is in trouble even though Kühberger et al. cannot identify the factors driving the Endowment Effect because "whatever subtle features of the situation triggered the difference in selling price, those features were presumably there for the observer subjects to see." Indeed, but to see is not to feel. The O's are affectively involved much more than the S's.

Similarly, Nichols, Stich, and Leslie (1995, p. 441) note that to "suggest that successful simulation requires more than the information that was available to our [S's] is to admit that simulation is a marginal ability that would fail in most real life situations." This is an interesting objection to which my account must respond. No plausible view of ToM can predict high error frequencies across the board when people use their ToM capacities. My account can respond by virtue of the analysis represented by Table 7.1 on p. 193 and Figure 7.1 on p. 198. These allow that there can be scenarios in which the following possibilities apply:

i there is no Affect Mismatch, no System Mismatch and no Bias Mismatch;

ii there is Affect Mismatch, but no System Mismatch and no Bias Mismatch;

iii there is no System Mismatch and there may be biases but not enough to cause a simulation error;

iv there is a System Mismatch, but the O's were not biassed and so there is no Bias Mismatch.

All four of these routes are ones on which my account predicts no simulation error, so my account can respond to the charge brought by Nichols, Stich, and Leslie (1995) to the effect that it makes ToM too error-prone.

This explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (affect: pleasure in ownership) and BIAS in the O's (Endowment Effect) leading to systematic ToM error.

8.2.7 Gamblers

Gilovich suggests we make inaccurate predictions of how unhelpful data are evaluated by others, thus indicating a failure of ToM. The others in question are gamblers, in one example. We know that betting shops are profitable which means on average that gamblers are not. We might think then that gamblers are in denial about their losses. They must somehow ignore or forget the data relating to their losses. They must be ignoring data which disconfirms the hypothesis they cling to: that they are successful gamblers.

The surprising element of this case is exactly how people dismiss disconfirmatory data. We may expect that they will simply forget it; they will pay it no attention. As Gilovich (1993, p. 62) says, "it is commonly believed that people are more inclined to remember information that supports their beliefs than information that contradicts them." Gilovich has shown however that disconfirmatory data are considered more, not less. He cites an experiment he conducted showing that gamblers remember their losses more than their wins —but they construct narratives in which the losses were actually 'near wins.'[4] As Gilovich (1993, p. 62) writes: "people often resist the challenge of

[4]See also Taleb (2008) for discussion of the 'narrative fallacy,' whereby even constructing

information that is inconsistent with their beliefs not by ignoring it, but by subjecting it to particularly intense scrutiny." We work hard to find flaws in the disconfirmatory data so as to accord it a lower weight in our considerations than the unquestioned confirmatory data. Note that this is subtly different type of Confirmation Bias to the one discussed above in Ch. 6. There, people seek the wrong sort of data, that which can only tend to confirm their hypothesis. Here, O's are presented with data tending both to confirm and disconfirm their hypothesis, and deal with that scenario in a way that does not optimise the potential value of the data.

The surprise is generated when the question is asked as to whether one should question the quality of new data even-handedly, irrespective of whether it is confirmatory or disconfirmatory. Everyone will answer that question in the affirmative. That could indicate a System Mismatch between S and O, because S is using System 2 to respond to questions about data handling while O is just behaving using System 1. In any case, once again, S's are affectively too remote from the O's situation, since it is known that gambling is highly affectively involving and indeed addictively so. There is then a significant Affect Mismatch between S and O here. As a result, the S's do not simulate the Confirmation Bias of the O's and thus exhibit ToM errors in their simulations.

This explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (system mismatch) and BIAS in the O's (Confirmation Bias) leading to systematic ToM error.

8.2.8 Basketball Fans

There is a belief among basketball fans in the 'hot hand' phenomenon. This holds that players shoot in streaks: if they have just made a shot, they are more likely to make the next shot, and if they have missed, they are more

stories based on the actual facts can be misleading.

likely to miss. Gilovich (1993, p. 12) studies data relating to an actual team, and finds that the hot hand phenomenon does not exist. On the contrary, "there was a slight tendency for players to shoot better after missing their last shot." Since we may expect basketball fans to have a close acquaintance with the relevant data because they spend a lot of time watching the reality generating it unfold, their belief in the hot hand phenomenon is unexpected and thus represents an error in our ToM.

It might be objected here that folk psychology is neutral as to what the fans believe, but I suggest this is only the case before being asked the question. Naturally, S cannot be asked the question as to what the O's who are fans will believe without being given the relevant facts. The idea then is that if we started with S's being told about Gilovich's data, S would predict that fans would not have this belief. It is, I concede, also possible that S's would simply not predict that the O's do believe in the hot hand phenomenon. Obviously this is an empirical question. In any case, even a failure to take a view represents a ToM error, since the O's hold the hot hand belief strongly. I might again mention here the unavoidable disconnect between the O's who are always already in the affect-laden situation and the disinterested situation of the S's who are given a dry description of the relevant facts.

Moreover, for at least those of us calmly considering the phenomenon, the fact that the data are random —or indeed, tending to show that there is an 'anti-hot hand phenomenon' if there is any effect at all —and do not support the phenomenon is made highly salient by its centrality in the discussion. Thus, we are surprised by the first order error; we make a second order error in our ToM because we are told dispassionately about the data. We would simulate better if we were able to place ourselves in the highly non-dispassionate position of a basketball fan, thus reducing the Affect Mismatch between S and O. One explanation of the situation here which is consistent with my account

is that the bias mismatch in this experiment derives from the fact that the O's fall prey to the Clustering Illusion described in Ch. 6. Naturally, this claim is speculative and could be supported (or refuted) independently of the view for which I argue. The structure of the Bias Mismatch Defence requires independent evidence for both FACTOR and BIAS, and for a link between them. Empirical work would be decisive here.

Gilovich makes the data more comprehensible for the fans by presenting it in numerical form: there is no statistical tendency for players' hits to follow hits more than misses. We might expect fans presented with this data to accept it, and admit that their previous belief in the hot hand phenomenon was mistaken. This does not happen: Gilovich (1993, p. 13) writes that most people question the data, "[t]he hot hand exists, the argument goes, it just did not show up in our sample of data." This again we will not expect as S's. The difference of course is that we are not at all committed to the hot hand, we are considering the matter dispassionately, and we will generally believe of ourselves that we will respond to convincing empirical evidence of the falsity of a belief by negating the relevant propositional attitude. That would give us a System Mismatch.

O's fall prey to one or other bias, but S's would not, unless they were shown the actual data. Thus, we could convert poor S's to good ones by changing the bias status of the S's. If we ask them whether basketball fans will believe in the hot hand in the face of contrary evidence, they will say no. If we ask them whether they will believe in streaks of shooting based on a sequence of six hits in 20 shots, and show them that sequence as a series of X's and O's, the S's will now be more closely tied in to the actual situation of the fans, and will now model the fans better by matching their Clustering Illusion bias. This would be another example of successful simulation by bias matching.

We may derive a further confirmation of this conclusion from Figure 2.2

presented by Gilovich (1993, p. 20). This shows the pattern of V-1 bombs falling in London during the war. There was a belief at the time that certain parts of London were safe from the bombs and others were not. As Gilovich (1993, p. 21) argues, this belief is created by dividing the map into quadrants, and observing that there are clusters in some quadrants. He notes that if the map were divided by diagonal lines, "there are no significant clusters." The point here is that shown the map, it appears to us that there are clusters: we would correctly simulate the Londoners who falsely believed the safe area/dangerous area hypothesis. But given the data dispassionately together with Gilovich's argument, we agree with him that randomness has fooled the Londoners and we do not simulate them correctly. Our simulation misses out the Clustering Illusion that leads O's to see patterns where none exist, unless we also see the same map. Then we would also exhibit the Clustering Illusion and we would have another instance of successful simulation via bias matching.

Strictly speaking, this does not appear to be an instance of the Bias Mismatch Defence because there is no mismatch of biases. That is, if S and O get the same inputs, S's prediction of O's belief is correct; while if S and O get different inputs, this is not guaranteed (even if the inputs carry basically the same information but are presented differently). So this is not exactly about *mismatch* in bias. It does though indirectly strengthen the Bias Mismatch Defence, because it gives it more explanatory power. Normally it is predicting ToM error associated with since caused by bias mismatch but here it is predicting the absence of ToM error because one of its causes is missing. Naturally this would only be a ceteris paribus prediction and would be subject to empirical confirmation.

8.2.9 Cancer Cure Assessors

Gilovich (1993, p. 30) cites data relating to whether people believe that cancer patients who engage in 'positive mental imagery' benefit their health status. He reports that people answering this question do not follow the scientifically correct rule based on the fact that instances "of cancer remission in patients who practice mental imagery do not constitute sufficient evidence that mental imagery helps ameliorate cancer" because there must be a control group i.e. the mental imagery practitioners might have improved anyway. What is needed is not a cure, but a correlation between a change in the independent variable —the practice of imagery —and a change in the dependent variable, the cure rate. Moreover, the cure rate improvement must be reproducible and not occur with changes in other independent variables.

This is an example of examining the wrong data. Gilovich's central charge is that people take evidence for a hypothesis as also confirmatory of that hypothesis, when they should use one dataset to form hypotheses for testing and further datasets to confirm them. He writes: "willingness to base conclusions on incomplete or unrepresentative information is a common cause of people's questionable and erroneous beliefs" (Gilovich 1993, p. 30). Often, these situations will be instances of Confirmation Bias, where people tend only "to focus on positive or confirming instances" (Gilovich 1993, p. 33) of a hypothesis they are testing. Whether one is surprised or not by this prevalence of Confirmation Bias —with the surprise being the indication of a failure in ToM —will depend on one's general level of cynicism in relation to the frailty of human reasoning. But one might be surprised at such a widespread lack of quality in data handling. This is not expecting untrained O's to be aware of correct scientific method so much as expecting them not to form scientific conclusions in unscientific ways. Even untrained O's are aware of the idea that A has not caused B if B would have happened anyway. Also note how uncongenial these

data are to TT(Scientific), which holds that pre-fives are experts on hypothesis selection and confirmation. What has happened to this expertise during maturation?

There is an affect disparity between the O's who are actually in situations where they must make some decision based on whatever evidence is available and the S's who model that decision making. The S's face little or no involvement or stress related to the question that the O's are considering. The S's are not exposed to the risk of failing to make a decision, where randomness in the O's may be beneficial in breaking a Buridan's ass-type deadlock. The S's have time to employ System 2 reasoning to come up with a more considered answer while the O's may be under pressure and thus employ System 1. Alternatively, the persons likely to assess non-standard cancer cure approaches which are accessible without training, money or hospital equipment are likely to be persons with cancer or persons who know someone with cancer. This of course is an extremely affectively involving situation; we as S's here have no affective involvement at all. The result is that we as S's do not simulate Confirmation Bias in the O's, thus leading to systematic simulation error.

On the latter line, this explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (affect: fear of death) and BIAS in the O's (Confirmation Bias) leading to systematic ToM error.

8.2.10 Puzzle Solvers

Saxe (2005a) cites Pronin, Puccio, and Ross (2002) in her 'too cynical' category, but it reports on one experiment which falls into the 'too rosy' category, so we may consider it here. The experiment in question involves the type of children's game where something seems blindingly obvious to the participants who are 'in on it' and yet extraordinarily opaque to those who are not. So the

systematic ToM error is too rosy in that S's overestimate how easy O's will find it to succeed at the game.

The experiment involved asking people to figure out what things existed in 'My World' (Pronin, Puccio, and Ross 2002, p. 644) from a series of clues, and then assessing how many clues would in general be needed by peers to solve the puzzle. For example, one clue was 'My world has trees and grass but not flowers.' The governing principle is that My World contains only things that have double letters in their name. Once one knows this, this factor seems to jump out of the page with extreme vividness, but before one sees the principle, it is possible to stare blankly at an enormous array of clues, forming and rejecting an immense number of baroque hypotheses. The results were, as expected, that successful solvers vastly overestimated the proportion of the class who would solve the puzzle —they thought that 78% of the class would succeed, while only 21% did. By contrast, those who failed to solve the puzzle gave quite an accurate assessment of how many would succeed —25%. This is explained by noting the extreme vividness to the solvers of the principle once seen and feeding that into the Availability Heuristic.[5]

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. There is BIAS in the S's (Availability Heuristic) but no obvious FACTOR, so critical observers may insist that there is a gap in the Bias Mismatch Defence here.

[5]Similarly, Pronin, Puccio, and Ross (2002, p. 643) give further examples of scenarios where the False Consensus Effect in S causes S to be too optimistic about O's current mental state. The examples are when S gives directions to O and where S asks O to decode musical taps.

8.2.11 Shoppers Redux

As discussed in Ch. 6, Shoppers were asked to consider which of a set of four identical stockings was the highest quality. They chose the rightmost item more often than chance, without reason to do so. The particular bias involved here is dubbed the 'Position Effect' (Nisbett and Wilson 1977, p. 243). When asked why they chose as they did, they confabulated reasons. The reasons they gave involved spurious claims about the superior quality of the selected pair. These claims could only be spurious since the pairs were identical. This I think I can say is surprising without fear of contradiction.

There is an Affect Mismatch between us as S's and the shoppers as O's. It is appropriate to make a quick decision in many low-impact, real-life circumstances. The Shopper O's, we may easily imagine, are already somewhat harried individuals who have moreover been unexpectedly approached to answer unusual questions at a busy time. As Goldman (1992, p. 116) observes in relation to the shopper case, "one is unlikely to replicate the uncertainties of the live situation" when one simulates. The quickest way for the Shoppers to be able to get on with shopping will be to make a choice. In addition, they might disappoint the authoritative figure of the questioner if they fail to respond, in a similar scenario to that seen in the Milgram (1963) experiments discussed in Ch. 8. S's are exposed to none of these affects and so they do not apply the same biases when they run their simulation. Here, the Bias Mismatch Defence makes empirical predictions that should be noted: (a) S would predict correctly if made to rush; and (b) S would predict correctly if O was given a lot of time to decide. My account also involves speculation that the subjects were harried, whereas it might equally be speculated that they were not because they had enough time to engage in a survey. Further empirical work would be decisive here.

It might similarly be objected that the 'harried O/relaxed S' explanation

does not obviously fit with the experiment, since as I will outline below, Nisbett and Wilson (1977) are puzzled by the effect they observed. This in itself is of course a good illustration of a systematic ToM error in need of a Bias Mismatch Defence. I make two observations in response. Firstly, Nisbett and Wilson (1977, p. 244) do note that on being asked to explain whether the position of the article chosen had any influence on their selection of it, they denied it "usually with a worried glance at the interviewer suggesting that they felt either that they had misunderstood the question or were dealing with a madman." This development is more consistent with the shoppers being harried than relaxed. Secondly, Nisbett and Wilson (1977) are unaware that they are describing what will later be deemed a systematic error causing ST difficulties as an account of ToM. So they are also unaware that a Bias Mismatch Defence. might be needed. They can thus be forgiven for not investigating the affective state of their O's. Further empirical work would again be decisive here.

We may also have a System Mismatch here. The shopping O's are under pressure to make a decision and aware that it is not of the first importance exactly which decision they make. It is often inaccurately reported that Nisbett and Wilson (1977) offered the O's the pair of stockings they selected, in which case they would care somewhat about which pair it was. However, note the exact words of Nisbett and Wilson (1977, p. 243): shoppers "were asked to say which article of clothing was the best quality;" no mention is made of giving them the stockings. The harried O's use a System 1 heuristic to make a decision. This could be designed to avoid Buridan's Ass-type paralysis in decision making, where one does not know which of two equally good options to choose.[6] The point is that one should just pick one; it does not matter which. The S's are at leisure to simulate the O's and therefore use the more rational System 2. For one or both of these reasons, there is a bias mismatch

[6]Taleb (2007, Ch. 10) discusses such useful sorts of randomness that help us to avoid Buridan paralysis.

between S's and O's here, since the O's are influenced by the Position Effect and the S's are not.

I have generally refrained from challenging experimental procedures on the grounds that accepting it gives the opposing view its best case. I will make an exception for the Shoppers case since it is heavily cited by TT proponents; it is the one experiment of the many reported by Nisbett and Wilson (1977) that is mostly selected for comment. The problem with the Shoppers experiment is its lack of what experimental psychologists term 'ecological validity.' The experimental task should be sufficiently similar to everyday tasks that one may reasonably expect to be measuring elements of everyday behaviour. The Shoppers experiment by contrast focusses on what Johansson et al. (2006, p. 689) rightly term "a rather strange and contrived task" with much less ecological validity than choosing between different stockings for a good reason.[7]

I suggest that this explanation represents even strictly speaking an instance of the Bias Mismatch Defence since we have FACTOR in the O's (affect: harried) and BIAS in the O's (Position Effect) leading to systematic ToM error. It is open to objectors though to hold that FACTOR is inadequately specified here.

[7]Part of a better approach is suggested by Heal (2003, p. 83) who notes that "the rightward bias is irrational and hence not something we need expect simulation to cope with." This can be seen as suggesting an approach congenial to mine in which 'irrational' biases in O are not something that S can 'rationally' simulate.

Chapter 9

'Too Cynical' Evidence

9.1 Introduction

I will consider three of the papers Saxe (2005a, p. 177) cites with the aim of showing systematically too cynical errors in ToM, as discussed in Ch. 5. These will be as follows: Pronin, Puccio, and Ross (2002), Kruger and Gilovich (1999) and Miller and Ratner (1998).

I summarise the responses I give in §9.2 to the elements of this different class of Saxe (2005a) data in table 9.1. The common link between all of the data considered in this chapter is that S's predict that O's will perform less rationally in their reasoning than they actually do. The Bias Mismatch responses to the remaining systematic ToM errors fall into three broad categories as shown below

- (A): S and O exhibit various biases [one entry]
- (B): S or S and O exhibit Availability Heuristic [four entries]
- (C): S exhibits Self-Presentation Bias [five entries]

I will conclude that the Bias Mismatch Defence does a reasonable job of accommodating the 'too rosy' data and the 'too cynical' data.

Group Studied	Response
Conflict Parties	(A): S and O exhibit various biases
Marriage Partners	(B): S and O exhibit Availability Heuristic
Video Gamers	(B): S exhibits Availability Heuristic
Debaters	(B): S and O exhibit Availability Heuristic
Darts Players	(B): S exhibits Availability Heuristic
Blood Donors	(C): S exhibits Self-Presentation Bias
Healthcare Consumers	(C): S exhibits Self-Presentation Bias
Campus Drinkers	(C): S exhibits Self-Presentation Bias
Smokers	(C): S exhibits Self-Presentation Bias
Statement Releasers	(C): S exhibits Self-Presentation Bias

Table 9.1: Response Type By Group Studied: Too Cynical

9.2 'Too Cynical' Data

9.2.1 Conflict Parties

Saxe (2005a, p. 177) writes that "[m]ost adults believe that reasoning can sometimes be distorted —both inevitably, by the limitations of the mind, and wilfully, as in wishful thinking and self-deception —and that this is more likely to be true of other people's thinking than of their own" and cites Pronin, Puccio, and Ross (2002) in support. This gives us three key claims. The first is that ToM predicts distortions in reasoning. The second is that these distortions may be voluntary or involuntary. The third is that S's predict more such distortion in O's than in S's. The first claim is not an example of ToM error, since it is true that there are many distortions in reasoning. The third claim seems by contrast to be a clear example of ToM error, since there is no justification across the board for S's reasoning to be less distorted than O's. The second claim is complex and interesting. We can agree that it is

true that there are some voluntary and some involuntary distortions, but the truth of that claim would not suffice to make it the case that there is no ToM error here. To avoid error, S's would need not only to recognise that there are voluntary and involuntary distortions of reasoning but also accurately to identify occasions when each are occurring.

The particular focus of Pronin, Puccio, and Ross (2002) is on how predictions of biased reasoning can bring parties into conflict so I use this as a title for the section. We might also use the term 'biased expectations of bias' to describe the focus of the authors. I will be arguing that the second claim is at the heart of their position. Conflict is caused not because S predicts bias in O, but because S wrongly sees O's bias as voluntary.

Debates about political questions are often highly affectively involving and so lead to conflict at some level. Pronin, Puccio, and Ross (2002, p. 637) give as examples debates about "capital punishment, abortion policy [and] the Middle East." S's considering the positions of O's who oppose their particular views on such topics will form negative views of the reasoning abilities of the O's. S's make "harsh evaluations of [O's] on the other side, whose perceptions and arguments [...] appear biased and self-serving" (Pronin, Puccio, and Ross 2002, p. 637). This is systematically too cynical ToM. Pronin, Puccio, and Ross (2002, p. 637) tell us that what fosters this is the way that partisans "accept at face value arguments and evidence congruent with their interests and beliefs" while subjecting opposing arguments to intense scrutiny. This of course is a definition of Confirmation Bias.

So the ToM errors here result from S failing to simulate Confirmation Bias in the O's in the right way. The S's are excessively cynical about the intentions or genuineness of the O's. The S's expect in one sense that the O's will exhibit biased reasoning, but do not ascribe it to Confirmation Bias —which may be an unavoidable aspect of human reasoning —but to a partisan and deliberate

failure properly to examine the facts and arguments. If simulation modelled bias in the right way, then we would expect the S's to predict the reasoning of the O's more accurately and also ascribe it less to deliberate partiality of the O's, which might reduce conflict. The key point for the Bias Mismatch Defence continues to be that the process is still well described as a simulation failure due to inaccurate bias modelling in S.

This conflict situation will be exacerbated if the S's also apply their own Confirmation Bias to the subject matter, because this will open the gap wider between S and O. This argument is intuitively compelling and empirical well-supported by citations supplied by Pronin, Puccio and Ross. In one citation, Edwards and Smith (1996) discuss a 'disconfirmation bias;' i.e. S's tend to apply a negative confirmation bias to the positions espoused by O's. The arguments of the O's are subjected to more intense scrutiny. The effects are worsened by Affect Mismatch: S's and O's may become passionately attached to their positions. As Edwards and Smith (1996, p. 20) note, "affective and motivational factors influence cognitive processes to produce biased conclusions." In sum, the overly cynical ToM errors here can be explained by S applying Confirmation Bias to S's own positions, a negative Confirmation Bias to O"s positions and then also failing to allow for O's own Confirmation Bias.[1]

Pronin, Puccio, and Ross (2002, p. 637) also note how the Availability Heuristic, the Representativeness Heuristic and Cognitive Dissonance reduction can all lead to conflict. I suggest the same mechanism applies here as described above. S's see "self-serving or ideologically determined biases in [O's] views." Again, if the S's simulated correctly, they would be aware that these biases are unavoidable. So we have the failed simulation of three more

[1]See also Kunda (1990) for discussion of selective memory access to bolster biased positions, and Short (2012) for arguments from Nietzsche to the effect that such memory selection is a feature of active and strong individuals.

types of bias generating systematically too cynical ToM error here.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It may loosely speaking be an instance though because we have FACTOR in the S's (affect: emotional impact of conflict situations) leading to BIAS in the S's (Availability Heuristic; Representativeness Heuristic; Cognitive Dissonance).

Pronin, Puccio, and Ross (2002, p. 644) describe how the False Consensus Effect may make teachers see students as "inattentive, unmotivated or even stupid" because the teacher fails to set aside her own mastery of the subject. Similarly, Pronin, Puccio, and Ross (2002, pp. 644–646) discuss an 'inadequate allowance' thesis in the context of a word game. S's who know the answer overestimate how easy it is to find the answer due to the False Consensus Effect. They then make "unwarranted negative inferences" about the O's. So a bias in S's leads them to make systematically too cynical ToM predictions.

These explanations do not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. There is BIAS in the S's (False Consensus Effect) but no obvious FACTOR, so critical observers may insist that there is a gap in the Bias Mismatch Defence here.

9.2.2 Marriage Partners

As discussed previously in Ch. 5, the predictions of marriage partners of one another's assessments of contributions to various activities were studied by Kruger and Gilovich (1999), being the second 'too cynical' citation of Saxe (2005a). They considered an array of positive and negative marriage activities, such as dog walking, or beginning arguments. They asked each S to rate S's own contribution to each activity, O's contribution to each activity, and

crucially, to state what S thought O would say that O's contribution to each activity was. The hypothesis was that S's would predict that O's would be self-serving in their responses i.e. O's would claim more responsibility than justified for positive activities and admit less responsibility that justified for negative activities.

The hypothesis was confirmed. Kruger and Gilovich (1999, p. 745) reported that "couples expected their spouses to claim more than their share of the credit for the desirable activities [...] —but less than their share of the blame for the undesirable activities." The spouses did indeed claim more than their share of the credit and accept less than their share of the blame. So the S's exhibited no systematic ToM errors in relation to the nature of credit and blame claims by the O's. The ToM errors were related to the amount of such differential claims. The O's engaged in making such differential claims to a lesser extent than predicted by the S's. The S's were thus 'too cynical' here in their ToM, in that they predicted more significantly differential claims to be made by the O's. These errors were symmetrical in that both spouses made them in relation to each other.

The explanation here is that both S and O exhibit the Availability Heuristic. The activities of each S are more available to that S than the activities of O are available to that S. This means that S's are likely to claim more responsibility for both positive and negative activities than is warranted. Then, these S's will make the opposite error in relation to O. As Kruger and Gilovich (1999, p. 744) point out, S's "may be surprised to find that others often claim too much responsibility for [negative] activities as well." This can be explained on the Bias Mismatch Defence for which I have been arguing. There is a Bias Mismatch between the S's and the O's. The S's are failing to allow for the application of the Availability Heuristic by the O's.

We have a Bias Mismatch though in a special sense of that term. Both

partners exhibit the Availability Heuristic, but they do so about different topics. This is how they can both apply the same bias but still make simulation errors. In each case, one partner as S employs the Availability Heuristic to overstate S's own contribution to the activities and understate the contribution of the other partner as O. Since the mirror image of this process occurs in the other partner when they are in the role of S, both of them come to overstate their own contribution and understate that of the other partner. This has the results seen: both partners predict that the other partner will be more self-serving than they actually are, which can lead to problems. Kruger and Gilovich (1999, p. 744) confirm this when they note the potential adverse effects of not allowing for the biases of others when they write that: "[i]nstead of attributing another person's inflated assessment to the availability bias, people are likely to see it as a motivated grab for excess credit." That gives us one affective distinction between S and O; it is also obvious that S will have an emotional investment in exaggerating his own positive contributions.

One might wonder whether the claim here is that S and O are subject to biases rather than an application of the Bias Mismatch Defence. Certainly, that is the explanation of the data, but I think we could still see this as an application of the Bias Mismatch Defence, judged solely on this parameter because the ToM failure continues to result from the fact that S has a bias of the same sort as O and about the same sort of activity, but about S's own activity rather than O's. This also applies to O – or put differently but equivalently, we obtain the same prediction if we switch S and O – so we obtain a prediction of the way that S's and O's errors mirror each other. So we can see that there are at least three aspects of bias that must match to avoid a bias mismatch: a). type of bias; b). subject matter of bias and c). subject of bias (meaning which person's activity in this example).

This Availability Heuristic explanation of these data is also given by the

original experimenters. Kruger and Gilovich (1999, p. 743) write of those original experimenters having: "offered an information-processing interpretation of this bias, one based on the differential availability of one's own and another person's contributions. Simply put, people have an easier time remembering their own input than someone else's." So, as Tversky and Kahneman (1973, p. 207) point out, "reliance on the Availability Heuristic leads to systematic biases" and we have explained exactly the systematic ToM errors which Saxe (2005a) cites.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in exaggerating own positive contributions) leading to BIAS in the S's (Availability Heuristic).

9.2.3 Video Gamers

The same explanation is available for a total of four studies reported by the same authors. In the second of four studies reported in the paper Saxe cites, Kruger and Gilovich (1999) examined assessments of bias in players of a two-person video game. This was a co-operative game where both players had to work together against a common enemy. There were two players and an observer. After the game, all three assessed the contributions of both players on eight parameters, evenly divided between negative contributions and positive contributions. Also, the players estimated how much each player would claim he contributed on each parameter, and what the observer would say.

As with the marriage partners, the video gamers expected more self-serving bias than was actually the case. "Participants expected their teammates to credit themselves with 23.0% more responsibility for the desirable game ele-

ments" and "less than their share of the blame for the undesirable game outcomes" but in "actuality, players took 8.3% more credit for the undesirable outcomes of the game" (Kruger and Gilovich 1999, p. 751). These results can be explained on the same grounds as the Marriage Partners case: S exhibited the Availability Heuristic.

The players thought that the observer would say the same as they did —i.e., unsurprisingly, the players thought that their opinions were objectively valid. This means they were unaware of the operation of the Availability Heuristic so did not correct for it. Pronin, Puccio, and Ross (2002, p. 662) cite evidence to the effect that "people are often unaware of their own unawareness" in the context of bias.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in exaggerating own positive contributions) leading to BIAS in the S's (Availability Heuristic).

9.2.4 Debaters

This study aimed to investigate a "more motivationally charged situation" (Kruger and Gilovich 1999, p. 748) with the aim of examining whether motivation affected ToM errors. Kruger and Gilovich (1999) did this by studying undergraduates taking a debating course, who wanted to do well, since they sought careers in law and politics and the like. Participants debated a political topic in teams of two, and were subsequently asked anonymously to apportion responsibility for positive and negative aspects of the debate between themselves, their team-mates and their opponents. They were also asked to predict what apportionments the team-mates and opponents would make. There are

two factors in play here. There is to some extent an objective fact of the matter about who did what in the debate —or at least a less subjective reality such as one might expect from an impartial observer. The second factor is a subjective 'overlay' on the objective facts, reflecting the hypothesis that S's would give themselves more credit for positive aspects of the debate but also expect O's to do the same.

The results were consistent with the hypothesis that S's would predict more self-serving bias in the O's than the O's actually exhibited. Kruger and Gilovich (1999, p. 749) found that "debaters expected their opponents to claim 69.8% more of the credit for the desirable outcomes than for the undesirable outcomes" but "this assumption was wildly exaggerated" since "[d]ebaters in fact credited their own team with 21.0% more of the credit for the desirable outcomes than for the undesirable outcomes." This prediction of biased estimation still appeared when S's considered their team-mates, but much less so, with S's predicting team-mate O's would claim "26.0% more of the credit for the desirable outcomes than for the undesirable outcomes" (Kruger and Gilovich 1999, p. 749).

These results can again be accounted for by assuming that the S's applied the Availability Heuristic, with some extension from themselves to their team-mates. In short, there is a hierarchy of availability which follows the order S; team-mate O; opponent O. Thus we can explain why S's predicted that their team-mate O's would take some more credit for desirable outcomes than justified, and opponent O's a lot more. It is because S's own activities are somewhat more available than those of team-mate O's and much more available than those of opponent O's. This line is suggested when Pronin, Puccio, and Ross (2002, p. 637) summarise the literature in writing "[i]intergroup enmity can arise from simple availability and representativeness biases."

S's "thought their opponents would claim [...] 32.7% less than their fair

share of the undesirable debate elements" (Kruger and Gilovich 1999, p. 750). In partial contrast, they thought their team-mate O's would also admit less than their full share of responsibility for undesirable debate elements, but would not do so to the same extent as the opponent O's. The reality was that O's of both types admitted to more responsibility than expected. This can be explained by an extension of the 'marriage partners' account to allow for the additional participants in this experiment. S exhibits the Availability Heuristic such that S's own actions loomed larger in the debate on both positive and negative sides than those of the team-mate O's. S also exhibits the same heuristic in relation to the even less available actions of opponent O's. These biases explain the systematically too cynical ToM of S in this scenario together with the different levels of cynicism in relation to team-mate O's and opponent O's.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It also may not even loosely speaking be an instance because the FACTOR in the S's is difficult to disentangle; we do at least have BIAS in the S's (Availability Heuristic). Critical observers may insist that there is a gap in the Bias Mismatch Defence here.

9.2.5 Darts Players

Kruger and Gilovich (1999, p. 751) found in the final study reported in the paper Saxe cites that "darts players thought their opponents would be more self-serving than their teammates and more self-serving than they actually were." These results are similar to the ones about the debaters and can be explained in the same way using a hierarchy of availability.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in

the O's leading to systematic ToM error. It also may not even loosely speaking be an instance because the FACTOR in the S's is difficult to disentangle; we do at least have BIAS in the S's (Availability Heuristic). Critical observers may insist that there is a gap in the Bias Mismatch Defence here.

9.2.6 Blood Donors

Five studies in Saxe's final 'too cynical' citation, Miller and Ratner (1998), are claimed to show evidence of systematic cynicism in ToM. I will be explaining them all by appealing to Self-Presentation Bias in the S's.[2] In each case, responses by participants are likely to be dominated by what they think they should say —or what shows them in a positive light —rather than what they actually think. Note that participants need not be aware of the operation of Self-Presentation Bias.

The first Miller and Ratner study examined the number of S's who would donate blood with and without payment and compared this to the estimates of the S's as to how many O's would donate blood with and without payment.

The results of the study are shown in Table 9.2. Miller and Ratner (1998, p. 54) found that "(63%) indicated they would agree to give blood if not paid, and [...] (73%) said they would agree to give blood if paid $15." So the cash incentive had little effect because there were only an additional 10% of participants whose minds were changed by the payment.

However, S's "estimated that roughly twice as many [O's] would agree to donate blood for $15 as would agree to donate blood for free;" the S's estimated that 63% of O's would donate for payment whereas only 32% would donate without payment. This study then is an example of too cynical ToM, because the S's expected that more of the O's would agree to donate only if paid than was actually the case. So the challenge here for ST is that the S's

[2] Pronin, Puccio, and Ross (2002, p. 665) also discuss the Self-Presentation Bias in terms of a "holier than thou" effect.

	Volunteer Rate	
Incentive	Actual	Estimated
Payment	73%	63%
No Payment	63%	32%

Table 9.2: Actual Versus Estimated Number Of Individuals Volunteering To Give Blood For Payment Or No Payment

predicted that the O's would be more motivated by payment than by altruism than was in fact the case; and the S's made this too cynical prediction even though the S's themselves were not in general more motivated by payment than by altruism.

One way for S's to promote a positive self-image in this experimental scenario is to make it look as though they are uncommonly altruistic. This they can do by saying that they would themselves volunteer to give blood for no payment but also by saying that few others would do so. It is not sufficient merely to volunteer if everyone else does as well. So the data are explained by Self-Presentation Bias in the S's.

The same objection can arise here as came up in §9.2.2 viz.: objecting that the data are explained merely by S and O exhibiting biases rather than this being a case where the Bias Mismatch Defence applies. However, the same response that I set out on p. 243 is available; S and O do indeed both exhibit Self-Presentation Bias and it is indeed about the same type of subject matter, but, crucially, S's Self-Presentation Bias relates to S and not to O.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in own self esteem) leading to BIAS in the S's (Self-Presentation Bias).

9.2.7 Healthcare Consumers

The second Miller and Ratner study examined the effect of sex on views of a putative US government programme to make abortion available at public expense.

A large majority of S's agreed that such a programme would benefit women more than men. They also thought that this would mean that women would be more in favour of the programme than men. Miller and Ratner (1998, p. 56) write that the "majority of [S's] in this study perceived women to have a greater stake in, and to be more supportive of, a proposed health care plan than men." This was a systematically too cynical ToM error though, since in fact "there was no difference in the degree of support expressed by men and women" (Miller and Ratner 1998, p. 56).

One way for S's to promote a positive self-image here as in the other experiments in this class is to maintain that S is less subject to biased reasoning than O's. So both male and female S's here thought that S's own opinion was free of bias but that O's opinion would be heavily biased by self-interest. Self-Presentation Bias in S explains this desire to predict that O is more prone to bias than S, and in this case to predict wrongly that female O's would favour the programme more than male O's.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in own self esteem) leading to BIAS in the S's (Self-Presentation Bias).

9.2.8 Campus Drinkers

The third Miller and Ratner study related to attitudes to alcohol pricing on campus. We learn that there was a ban on the sale of kegs of beer at Princeton

which affected younger ('sophomore') undergraduates more than older ('senior') ones, because the latter were members of dining fraternities untouched by the ban. The experiment involved examining the interaction between three items: age, condition and performance. The age parameter was binary between junior and senior; the condition was binary between favour or oppose; the performance was binary between whether or not there was ToM error.

The results were similar to those seen in the two studies reported above: the "majority of participants in this study perceived Princeton sophomores to be more adversely affected by, and to be more opposed to, the keg ban than Princeton seniors" but in fact "there was no difference in the opposition expressed by sophomores and seniors" (Miller and Ratner 1998, p. 57). Again, this is explained by the S's ascribing more biased reasoning to O's than to themselves. It could also be that the sophomores wanted to think something like 'my peers are more addicted to alcohol than I am, so they will oppose the ban, while I am health-conscious enough to favour it.' Either way, the results are driven by Self-Presentation Bias in the S's.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in own self esteem) leading to BIAS in the S's (Self-Presentation Bias).

9.2.9 Smokers

The fourth Miller and Ratner study looked at whether smokers and non-smokers favoured smoking bans, and whether smokers and non-smokers were expected to reason in their own interests. There was a change here to the prior studies. The prior studies had shown predictions that there would be a relationship between self-interest and reasoning in scenarios where no such

relationship existed. The hypothesis in the smoking study was that there would be such a relationship, but that its strength would be overestimated. As Miller and Ratner (1998, p. 57) point out, they "do not claim that vested interest never affects attitudes, only that it does not affect attitudes as much as lay theories assume." The hypothesis was borne out by the results. Over-prediction by S's of self-serving bias in O's combined with S maintaining the belief that S is himself free from such self-serving bias explains the data and represents Self-Presentation Bias in S. This explains the systematic error in S's ToM in this scenario.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in own self esteem) leading to BIAS in the S's (Self-Presentation Bias).

9.2.10 Statement Releasers

The fifth Miller and Ratner study looked at behaviour rather than attitudes. Participants were told of a purported health threat that affected either only men or only women, and a proposed cut in government funding of research into it. The questions were whether participants would agree to release a statement about the cut to a local political organisation; whether members of the sex with a vested interest would agree more than members of the opposite sex; and whether S's would predict that O's with a vested interest would agree more. As before, S's "predictions significantly overestimated the actual impact of self-interest on behaviour" (Miller and Ratner 1998, p. 59) in that S's predicted that the affected group would release more than the unaffected group, even though in reality both groups exhibited similar high release rates. Over-prediction by S's of self-serving bias in O's combined with S maintaining

the belief that S is himself free from such self-serving bias explains the data and represents Self-Presentation Bias in S. This explains the systematic error in S's ToM in this scenario.

This explanation does not represent strictly speaking an instance of the Bias Mismatch Defence since we do not have FACTOR in the O's or BIAS in the O's leading to systematic ToM error. It is loosely speaking an instance though because we have FACTOR in the S's (affect: emotionally invested in own self esteem) leading to BIAS in the S's (Self-Presentation Bias).

Based on the discussion of data in this chapter and the previous one, I conclude the following. While the Bias Mismatch Defence has not completely covered all of the data, often because FACTOR is inadequately specified, overall it has done a very reasonable job of covering it. And even absent FACTOR, it is still of value to note that BIAS has caused a systematic ToM error. In conclusion, much of the 'too cynical' data from this chapter and the 'too rosy' data from the previous chapter can be explained by appealing to the Bias Mismatch Defence, especially when one considers, as I contend is reasonable, its looser formulation along with the very strict formulation. It is also worth noting that the defence will be available for much other existing data showing systematic ToM error and also future data of the same type.

Chapter 10

TT: Inaccurate Generalisation Defence

10.1 Introduction

As Saxe (2005a, p. 175) explains, if an experiment shows systematic error in ToM, for example where a child systematically and wrongly attributes error instead of ignorance, then "the actual result is best explained by an inaccurate generalisation in the child's developing theory of mind." This Inaccurate Generalisation Defence is the TT answer to all cases of such systematic ToM error. Recall that I mentioned early on (cf. p. 28) that I would not be considering any putative forms of TT which are not based on generalisations. Any such form of TT which was enable to avoid generalisations while remaining a theory-theory view would have the same problem that Saxe (2005a) presses against ST; viz. how can TT explain systematic ToM error? No generalisations means no Inaccurate Generalisation Defence. Recall also that there is a 'body of folk psychological knowledge' under TT accounts (cf. p. 36); upon what is that based if not on generalisations? It remains the case though that TT is not committed to any particular theory or any particular view of what

a theory is; though I have suggested (cf. fn. on p. 85) that attempting to base a theory on a structure like DNA without generalisations omits what is useful about theories. It would be valuable to have these underlying views on what constitutes a theory spelt out by TT proponents.

Saxe's argument is that TT or Strong S/T Hybridism is to be preferred to ST or Weak S/T Hybridism because TT can appeal to the Inaccurate Generalisation Defence and ST has no response. I have already provided a 'weak' defence of ST against this charge by showing that it can in fact appeal to a Bias Mismatch Defence. In this chapter, I will not go further by aiming to provide a strong defence of ST by showing that the Bias Mismatch Defence forms with ST a more plausible account of the data than TT with the Inaccurate Generalisation Defence. I will show however that the game has changed and now TT must compete with ST on the aspect of explaining systematic ToM error; I will also show that this does not look straightforward.

It might be thought that Saxe may have in mind something much less ambitious than a fully general explanation of error: the kinds of errors we observe could perhaps be explained by persons having, or lacking, certain beliefs about minds. We should note though that Saxe (2005a, p. 175) writes that "the argument from error suggests that aspects of the observer's naïve theory of psychology (like over-attributing rationality, and naïve cynicism) play a pervasive role in reasoning about the mind." However, Saxe (2005a, p. 175) explicitly only appeals to the Inaccurate Generalisation Defence to explain the systematic ToM errors seen in the Ruffman (1996) data on variants of the False Belief Task. But I think in fact that since I have now provided ST with a fully general account of systematic ToM error by adding the Bias Mismatch Defence, the same is needed for TT if it is to compete with ST on this figure of merit. It is unclear what defence TT will have beyond such an appeal to inaccurate generalisations. Perhaps alternative defences could

include, for example, a theory that omits certain facts, or which makes certain predictions very expensive to compute. This seems possible, but is otherwise unmotivated and will be less simple an explanation that ST combined with the Bias Mismatch Defence. So this means that it is likely many inaccurate generalisation scenarios will need to be constructed by TT proponents.[1]

I will examine the feasibility of this task in §10.2 by attempting to set out what the inaccurate generalisation will in fact look like in the various experimental situations. I have previously suggested that the reason TT proponents have been extremely reluctant to attempt this is because it cannot be done, or at least will prove enormously cumbersome. In what follows, I will nevertheless make the attempt, raising questions along the way as well as attempting to construct the inaccurate generalisations. One such question will be why we should accept that adults as well as young children continue to employ such inaccurate generalisations when they have a wealth of disconfirmatory data.

I will address the majority of the experiments previously discussed, covering both the too rosy and the too cynical classes of data here. I will concede that some do in fact comport well with the Inaccurate Generalisation Defence. Overall though, we must judge the success of the Inaccurate Generalisation Defence across the board. How parsimonious is it across the range of experiments and how explanatorily powerful is it? I will conclude that in many cases, plausible inaccurate generalisations cannot be constructed. For this reason and because of the mysterious way in which these generalisations must be maintained, I will conclude directly that the Inaccurate Generalisation Defence of TT does not succeed and indirectly that ST or Weak S/T Hybridism augmented with the Bias Mismatch Defence is a far superior account of ToM.

[1]The best strategy for TT may depend on the earlier issue about whether TT needs to hold that a theory is a body of generalisations.

10.2 Constructing The Generalisations

The systematic ToM error in the Milgram (1963) data (§8.2.1) is seen in the conflict between the prediction by S of the behaviour of O and the actual behaviour of O. Many of the O's are prepared to set the dial to 450V while we as S's and the S's in the experiment generally do not predict this. The inaccurate generalisation seems to be 'O will not harm others without adequate justification.' This appears reasonably promising as an explanation of the systematic ToM error seen in this experiment, though we might pose questions as to the amount of work that 'justification' is doing in the inaccurate generalisation. It seems as though this is an inaccurate generalisation that adults could be expected to use, even though a cynical observer might ask how it is maintained in the face of much of the behaviour observed around one and in the media. Similarly, the same inaccurate generalisation will likely be successful in explaining the ToM errors illustrated in the prison experiment (§8.2.2). I will therefore concede that the Inaccurate Generalisation Defence succeeds in giving TT an account of these two experiments, but note that in both cases, the inaccurate generalisations include notions of justification. Introduction of such a fraught ethical issue, part of a heavily discussed and controversial domain of philosophy, means that while the generalisation may be simply stated, it cannot be called parsimonious[2] in application.

The Inaccurate Generalisation Defence seems to have greater difficulties in the case of the repenters (§8.2.3). There are two levels of ToM error here. In the experiment, we have the repenters making ToM errors about other participants in the experiment. The errors flow from the repenters inaccurately predicting the behaviour of those others by over-attributing their own attitudes to the others. This is a systematic ToM error which is readily explainable on

[2]Here, somewhat in contrast to the original definition of parsimony on p. 32, lack of parsimony means a single generalisation will have to capture a very complex concept.

the ST approach. However, Saxe (2005a) has cited this experiment and others like it as an example of a 'surprising' result in social psychology. The idea is that we as S's would not be surprised by what the repenters as O's do in this experiment if ST were correct. If it were correct, we would not be surprised because we would simply predict the result: we would predict the errors that the repenter O's make.

Since we have two levels of ToM error going on, we will need some new terminology. I will introduce 'T' for 'third person.' The roles are as follows: we are the S's who make systematic ToM errors about the repenter O's who themselves make systematic ToM errors about other repenters, the T's. The error that we as S's make is to fail to predict that the O's will predict that the T's will agree with the O's about whether to wear the sign or not much more than the T's actually do. So the O's over-attribute the attitudes of the O's to the T's. This new terminology is outlined in fig. 10.1 below, with the arrows representing ToM use. S can use ToM in relation to O, S can use ToM in relation to T, and O can use ToM in relation to T. All three of these occasions of ToM use can result in ToM errors.

Saxe has cited our surprise at the results of experiments in social psychology as an example of systematic ToM error. She presumably was mostly thinking of the ToM errors made by S about O outside the experiment. The Inaccurate Generalisation Defence of TT was then constructed to give TT the resources to account for these particular errors. However, in my view the Inaccurate Generalisation Defence must also account for the ToM errors made *within the experiment* by O about T. After all, this is also a systematic ToM error. If TT proponents employ a different defence here, then they are admitting that some cases of systematic ToM error are accounted for by the Inaccurate Generalisation Defence and some are not. That would be highly ad hoc and unmotivated. Below I will suggest that there are difficulties in many

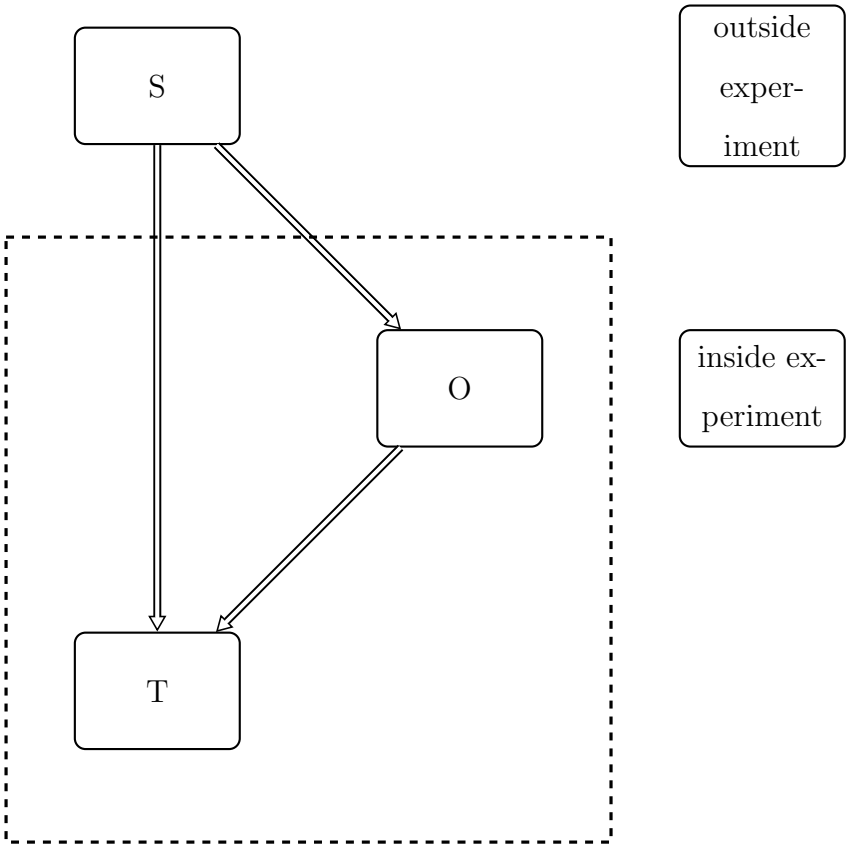

Figure 10.1: S's Inside And Outside Of Experiment

experiments in coming up with generalisations to describe the ToM errors made by O about T.

It seems as though the inaccurate generalisation in this experiment will be expressed as 'T believes what O believes,' together with the related 'T knows what O knows.' While this will handily encapsulate the observed data, it lacks parsimony in the same sense as I specified earlier in this chapter and plausibility. It will be particularly difficult for a modular account of ToM such as TT(Innate) to allow the entire belief set of O to be available within the ToM for ascription to T, because informational encapsulation is definitional of a (Fodorian) module. There are different problems with TT(Scientific), which must explain why a child learns the inaccurate generalisation that O

knows everything that S knows. So there seem to be difficulties for both TT accounts in explaining why such an inaccurate generalisation could arise. Moreover, TT proponents must explain why this inaccurate generalisation continues to be used by adults. Perhaps adults do not need to accept exactly this generalisation, but a replacement would need careful caveating to allow for motivated exclusions which fit the data. The at least occasional use by adults of the inaccurate generalisation 'T believes what O believes' seems to occur despite a large amount of data that must be constantly available to all S's that confirms that T does not in fact know everything that O knows or believe everything that O believes; or S for that matter. Indeed, the existence of the latter inaccurate generalisation would prevent O from passing the False Belief Task, which is not what is observed. Naturally, ST suffers from none of these difficulties, since on the ST account, O starts from O's belief set in simulating T and then modifies that belief set as required.

The Inaccurate Generalisation Defence of TT in the case of the quiz gamers (§8.2.4), will include a generalisation of the sort 'the level of general knowledge of O is to be assessed solely on the basis of O's ability to set specific questions.' Let us call this generalisation one. In order to explain the data and their perversity, there must be no modifying generalisations of the following sorts: 'generalisation one is to be adjusted to allow for O's advantage in setting the questions' or 'generalisation one does not apply across the whole range of general knowledge but only to the specific field about which O set the questions,' and especially not 'O's level of general knowledge is more accurately assessed by O's ability to answer questions rather than answer them.' It is the omission of modifying generalisations such as these that makes generalisation one an inaccurate generalisation. This will work for TT as an explanation of the data, but the same questions arise in relation to this experiment as in the case of the repenters discussed above. We may ask why it would be the case

that adults will employ such an inaccurate generalisation when they will have plenty of opportunities to observe disconfirmatory data. In both this experiment and the one relating to repenters, it also seems as though no-one would affirm the inaccurate generalisation if asked. The modifying generalisations do not seem to be greatly more complicated than the unmodified generalisation one, and so one might wonder why ToM would include only the unmodified version of generalisation one, since the modified one would be so much more accurate. On ST of course, there are no similar issues around why the generalisation set is so poor. S simply places himself in the position of O who has just set a number of rather impressive specialist questions, concludes that S has impressive abilities, and then ascribes those impressive abilities across the board to O. S does not at any point need to affirm the generalisation explicitly under ST, because ST does not include any generalisations.

The suicide note assessor data (§8.2.5 and §6.2.1) meet with mixed results on the Inaccurate Generalisation Defence of TT. The ToM error here is that S does not predict that O will continue to believe he is good at assessing suicide notes after the evidence therefor has been eliminated. The inaccurate generalisation here is approximately 'O's beliefs will conform to O's relevant evidence.' This represents a good explanation of this and many items of 'too rosy' data and so to that considerable extent, the Inaccurate Generalisation Defence of TT is here successful. Where there will continue to be questions will be around how such a generalisation is acquired. It is certainly a good starting point, which might be also a good argument for TT proponents. But why has it never been improved? There are plenty of observational data around suggesting that O will often in fact conform his beliefs to his evidence when that suits him. Indeed, that must surely also be an inaccurate generalisation of ToM, since it underlies much of the 'too cynical' data. So does not then TT face a similar problem in explaining systematic ToM error as ST does, with

the difference that the problem may be couched in 'conflict of generalisation' terms?

Prima facie, it seems as though the inaccurate generalisation in the case of the lottery ticket holders (§8.2.6) is the negation of the Endowment Effect. While stating a generalisation to achieve this is difficult, that may not be a serious problem for TT proponents. They merely need to say that the Endowment Effect is unknown so not part of ToM, whatever the form taken by the underlying theory. This line may make a testable prediction: if one tells S's about the Endowment Effect, they might make different predictions in future. One would certainly expect that if one told them about it and then immediately asked them to describe the behaviour of O in exactly the same circumstances. If this transpired to be the case, then this would appear to be evidence for an element of TT beyond what Weak S/T Hybridism already allows. Further empirical questions would then become pressing. To what extent does the new information become part of ToM? If the S forgets the data immediately, or even only after some time, it would be difficult to argue that the improved generalisation had become part of ToM. On the other hand, if TT avoids that difficulty, then it would be committing to a highly plastic picture of ToM, whereby ToM development is a lifelong process. That too might be fine, but it appears as though a further prediction of this line is that academic psychologists, especially those focussing on biases, will make many fewer systematic ToM errors than the general population, and their ToM will continue to improve as their professional knowledge does. It would be fascinating to see this prediction tested; I suspect that it might be found that it is partially true: some ToM errors are eliminated by knowledge and some cannot be shifted. If so, TT would owe an explanation.

There seems in fact to be a wider problem with economic generalisations of the sort seen here. The generalisation above seems to be a special case of

some broader scope generalisation like 'O will consider only relevant economic factors when making economic decisions.' The problem here for the Inaccurate Generalisation Defence seems to be that this is not merely a inaccurate generalisation, it is an exceptionally poor one: a very large proportion of people will fail to behave in accordance with such a generalisation on a very large number of occasions. Indeed, the entire literature on behavioural economics rests on exactly how often people fail to behave in such a way. One might even think that the number of occasions on which an individual makes a pure perfectly optimised economic choice is very limited. So ToM is here, on the Inaccurate Generalisation Defence of TT, going to make errors not just on a large number of occasions, but practically every time it is applied. There is no reason to accept this. Why do sophisticated adults with a lifetime of experience apply a ToM generalisation which fails on almost every occasion of use? Here, we should again prefer the ST explanation which just has S place himself in the lottery situation but fail to apply the Endowment Effect because a more intense affective involvement with the situation, one that might come from touching the lottery tickets and owning them physically, is needed to trigger such an application.

The initial proposal for the inaccurate generalisation in the gamblers experiment (§8.2.7) will be something like 'O will evaluate relevant data objectively.' This will probably appear too strong given the points made above about how often O's rationality is suspect. It might be weakened and made more economically relevant by focussing more narrowly on what exactly it is about this experiment which is surprising. That is that O seems to handle relevant data poorly even when the data matters crucially and economically to O. After all, gamblers lose money when they gamble badly, so surely they should look very closely at relevant data.

What seems to be going on in this experiment is a conflict of two general-

isations. The first generalisation will be a version of the above generalisation modified in the way suggested, viz. something like 'O will evaluate relevant data objectively when the question is of sufficient importance for O.' The conflict seems to result in a tension between the behaviour called for from O were this generalisation to be accurate, and O's desire to continue to gamble. Such a desire might result from O's desire to maintain his self-image as a successful gambler – which line of course fits neatly with the Bias Mismatch Defence – or possibly O is a compulsive gambler, bypassing his decision making system – which would open the possibility of employing the Harris (1992) Translation Defence of ST. In the former case, the second generalisation would be something like 'O wishes to maintain his self-image.' Both of these generalisations seem to be sound in terms of applying ToM on relevant occasions, so the problem seems to lie in combining the two. It seems that the Inaccurate Generalisation Defence of TT must claim that the first generalisation is given more weight than the second generalisation, when in this particular case, a more accurate prediction of O's behaviour would be given by applying more weight to the second generalisation. As usual, this approach seems complex and unmotivated.

The basketball fans experiment (§8.2.8) seems to require the Inaccurate Generalisation Defence of TT once again to appeal to a bad combination generalisation with the first one being a good data handling generalisation. Generalisation one would be something like 'O will make statistical judgements using the available statistical data appropriately.' TT defenders will not want that generalisation to apply across the board, because people are notoriously poor at making statistical judgements. Yet they might reasonably expect something like generalisation one to apply in relation to the hot hand phenomenon. After all, that is a pure statistical judgement which can be made solely on the basis of hits following misses and that data is right in front

of the basketball fans. It is not clear what generalisation is in conflict with generalisation one in this case. As I have suggested above, it looks as though the fans wish to maintain their self-image as basketball experts. That might explain why, once they have espoused the hot hand illusion, they continue to maintain their belief in it. On this view, the conflicting generalisation two is something like 'O will maintain his belief in his own expertise.' This looks somewhat like a 'bias mismatch defence' of TT. That might acquire some initial appeal since it fits the data, but there would be a parsimony cost to pay for the TT defender here. That cost arises because ST gets the biases for free while TT must engineer them in through careful generalisation selection. Moreover, this version of the Inaccurate Generalisation Defence still does not explain why basketball fans acquire their belief in the hot hand phenomenon to begin with. Perhaps TT defenders could suggest that the belief is culturally acquired, but they would have then move beyond the purview of a Inaccurate Generalisation Defence and into ad hoc territory.

It would be uncharitable in relation to the cancer cure assessor experiment (§8.2.9) to saddle the Inaccurate Generalisation Defence of TT with a generalisation that predicts that O will employ sound scientific reasoning in arriving at medical conclusions. A generalisation insisting that O constructs hypotheses based on one data set and tests them by examining their predictions on another data set would generate almost total ToM error. Weakening that, TT defenders might retreat to an expectation that O will often follow a dictum to the effect that extraordinary claims require extraordinary evidence (Hume 1993, p. 73). This still seems much too strong to be a plausible candidate generalisation of ToM, because if people generally behaved that way, we would see it and there would have been no point in Hume arguing for his dictum. The claim that mental imagery can have physiological effects including curing a very severe disease seems to fall into the extraordinary claim category,

though, so proponents of the Inaccurate Generalisation Defence will need to construct a generalisation that refers to it. An escape route suggests itself: TT defenders can appeal to what seems to be going on here, which is that O's with a vested interest in believing the proposition will find a way to believe it, despite the lack of evidence. So the generalisation could be something like 'O will require extraordinary evidence before believing extraordinary claims unless O has a vested interest in the claim in question.' This now looks like a good ToM generalisation in that it begins to describe what O actually does. And yet here is a problem for proponents of the Inaccurate Generalisation Defence of TT because this generalisation does not predict ToM error. The ToM error may in fact not exist. O believes, roughly, 'positive imagery cures cancer' absent a control experiment showing that. If S predicts this, then there is no ToM error and TT has nothing to explain – and neither does ST, so this experiment drops out of the frame as far as the ST vs TT debate is concerned. If S does not predict this, then it will be because S has applied the alternate generalisation discussed above: 'O will conform his beliefs to his evidence.' We may then ask again why this generalisation has not been improved. Again, all of these questions appear amenable to empirical work.

The explanation of the ToM error in the puzzle solvers experiment (§8.2.10) seems again to require wholesale belief attribution as in the case of the repenters discussed above. We take the role of S here, and the O is the person in the experiment who is setting the 'My World' task. There is a third person involved, who is the T that the O asks to solve the 'My World' puzzle. It is again important to keep our levels of ToM error separate here. There seems to be a systematic ToM error made by O about the solver. This is interesting, because it is explicable on the ST account, but Saxe (2005a) again cites these data as examples of 'surprising' results in social psychology. The idea is again that if our ToM were not subject to systematic error, we would not be

surprised by the results of such experiments because we would simply use our ToM to predict them. So that is the level of ToM error on which we should focus.

The problem here is that O wrongly attributes to T the entirety of the knowledge base of O. As discussed above in the case of the repenters, this is hard to achieve for the TT defenders while it is an immediate consequence of ST. The Inaccurate Generalisation Defence of TT could initially postulate a generalisation like 'T knows what O knows.' This is a start, but it will need modification to allow for obvious errors that will result from scenarios where O knows that T does not know a fact of key relevance. It is being able to account for such lacunae in the knowledge base of T that allows O to pass the False Belief Task when O is a typically-developing individual older than five. So the generalisation needs to be modified to 'T knows all facts that O knows less those facts that T clearly does not know.' This will basically allow the Inaccurate Generalisation Defence of TT to explain the False Belief Task data. Now there is a problem for the defence though, because it is clear to O in the 'My World' experiment that T does not know the rule. Precisely that is the point of the game, together with O's remarkable difficulty in working out just how difficult the task is for the T's, and our difficulty as S's in understanding why O makes such a remarkable ToM error. So the Inaccurate Generalisation Defence of TT needs to explain why the modified generalisation which explains the data in the False Belief Task does not apply here to allow O to make more accurate predictions of the likelihood of solving the puzzle when O himself solved it.

The Inaccurate Generalisation Defence of TT will claim in the shoppers experiment (§8.2.11) that the problem derives from S employing an inaccurate generalisation like 'O will make a quality selection based on relevant quality factors' unmodified by a strange generalisation to the effect that 'O will often

make a quality selection based on irrelevant quality factors like position.' Here, the Inaccurate Generalisation Defence succeeds, though there is some risk for TT defenders that the quality selection generalisation risks being far too optimistic about human behaviour.

I will now move on to the 'too cynical' data from Ch. 9.

The Inaccurate Generalisation Defence of TT will explain the data on conflict parties (§9.2.1) by including a inaccurate generalisation to the effect that 'T will make self-serving judgements when T is vitally interested in the outcome of the judgement.' (As previously, I will use the tripartite labels such that we are the S's who make the ToM error of interest about what the O's in the experiment say about the third party T's also in the experiment.) This generalisation seems useful for TT defenders, since it explains the data. The major problem for TT though is that ST defenders can bring the mirror-image of the charge of Saxe (2005a) here, demanding to know exactly why this 'cynical' generalisation will come to the fore in this experiment when previously generalisations like 'T will evaluate the relevant data dispassionately in making an important judgement' were more to the fore. These two generalisations are in direct conflict with one another and both must form part of ToM, because both are appropriate in some circumstances. It seems otherwise unmotivated to claim that this particular inaccurate generalisation will systematically be employed in all similar circumstances. Further, in the case of the teachers, it seems that they have misapplied the generalisation 'T knows what O knows' in the very face of the obvious fact that their job is to teach facts to their student T's that those students do not know. This opens up another difficult question for the Inaccurate Generalisation Defence of TT, which asks to what extent will inaccurate generalisations be modified, improved or deleted as a result of germane experience. Neither of the possible answers – no improvement versus

some improvement – appear promising.

The Inaccurate Generalisation Defence will in the marriage partners experiment (§9.2.2) again be postulating a 'cynical' generalisation, to the effect that 'T will take credit or admit responsibility for positive and negative actions in self-serving ways.' Then, O's ToM predicts that T will admit less responsibility for negative actions than is actually the case. Here, as with much of the 'too cynical' data we will be examining, the ToM errors in the experiment seem to all derive from O expecting more self-serving bias in various forms than is actually exhibited by T. The ToM errors made by us as S's outside the experiment relate to the fact that we do not predict what O will say about the T's. T does indeed exhibit some self-serving bias, but less than O predicts. Now we are entitled to demand why in the experiment there is such a systematic mismatch between the quantum of self-serving bias predicted by O and that exhibited by T. If TT(Scientific) is correct, then it might seem that O has learned the miscalibrated generalisation from experience which does not support it. Learning from experience should precisely have the function of improving the calibration of such a generalisation. TT proponents here may object that this is too crude, generalisations are indeed learned, but maybe the inaccurate generalisations are a side-effect of learning otherwise reliable heuristics. That line might work, but would need some specification. Otherwise, TT defenders might want to appeal to TT(Innate) and argue that the miscalibrated generalisation is part of an inherited module. On this picture, everyone inherits an inaccurate generalisation because the all inherit the same otherwise reliable heuristics. This line predicts that everyone will operate with the same inaccurate generalisation that is in all S's similarly miscalibrated. Why does no-one ever calibrate it correctly? This is particularly a problem for TT(Scientific), when according to major proponents Slaughter and Gopnik (1996, p. 2969) "another important feature of intuitive theories is that they

may be revised as a result of new evidence [which] differentiates theories from other types of cognitive structures, such as modules" found in TT(Innate).

Since much of the data in this section falls into the category described above, this 'Miscalibration Objection' will apply as well in many cases below. The Inaccurate Generalisation Defence of TT must explain not just the ToM errors made by us as S's about the experiment but the ToM errors made within the experiment by the O's. In my view, TT defenders have no response to this objection.

TT defenders will again claim in the video gamers experiment (§9.2.3) that the inaccurate generalisation within the experiment is 'T will take credit or admit responsibility for positive and negative actions in self-serving ways.' However, the Miscalibration Objection will apply again here. Further, we may ask why the miscalibration is different in this scenario than in the Marriage Partners experiment. Also, Kruger and Gilovich (1999, p. 747) note evidence that cooperative video games such as the one they employed often engender "other-serving judgements of responsibility." So the Inaccurate Generalisation Defence needs to modify the generalisation here to include a term like '[. . .] with the quantum of self-serving depending on whether the activity in question is cooperative or not.' That appears to be an unlikely generalisation modification. Again, we may step outside the experiment and ask why we as S's do not predict the errors that the O's make about the T's. ST can suggest instead here that the reduction in ToM error within the experiment comes about because of an improvement in simulation accuracy engendered by the closer relation involved in a cooperative scenario. As Kruger and Gilovich (1999, p. 747) observe, "intergroup rivalries tend to increase in-group cohesion." ST can continue to predict ToM error by S's outside the experiment though by noting that the external S's are non-participants and so immune to group cohesion effects.

Similarly, TT defenders will claim in the case of the debaters experiment (§9.2.4) that the inaccurate generalisation within the experiment is 'T will take credit or admit responsibility for positive and negative actions in self-serving ways.' The Miscalibration Objection will apply again here. There are now multiple levels of miscalibration in O's ToM relation to team-mate T's and opponent T's. The inaccurate generalisation needs to have O's predicting that T's who are opponents will make even more self-serving allocations of responsibility than team-mate T's who will be more self-serving than the O's take themselves to be. So the inaccurate generalisation within the experiment will need to make some reference to the status of the T as opponent or team-mate which seems implausible. And as usual, we as S's outside the experiment predict none of this.

TT defenders will again claim in the case of the darts players (§9.2.5) that the inaccurate generalisation is 'O will take credit or admit responsibility for positive and negative actions in self-serving ways.' The Miscalibration Objection will apply again here. The inaccurate generalisation needs to explain why in this experiment O's expected opponent T's to over-claim 14.6% of the credit for desirable outcomes than their team-mate T's and why O's expected opponent T's to admit 13.9% less of the responsibility for undesirable outcomes than team-mate T's. And outside the experiment, we as S's will it appears need to have an inaccurate generalisation that results in differences expressible in percentages.

The ToM error in the blood donors experiment (§9.2.6) is that O's predicted that T's would be motivated by money much more than they actually were. The inaccurate generalisation must be something like 'T's will agree to perform an unpleasant duty that benefits others more often if paid than not paid.' This is shown by the experiment to be true, but not to anything like the extent predicted by O's. Here, the Miscalibration Objection will ask why

the generalisation within the experiment is off by the amount it is; the 'more often' here is 32% when it should be 10%. It is not a good response here for TT defenders to suggest that ST also cannot explain these percentages, because ST, employing no generalisations, does not need to plug in any percentages into any generalisations. Since these are the percentages measured, these must be the amount by which Self-Presentation Bias affects reported judgements.

The Inaccurate Generalisation Defence of TT in the case of the healthcare consumers experiment (§9.2.7) will not presumably be specific to the level of healthcare, but will have O's predicting that T's act to their own advantage. The generalisation within the experiment will be something like 'T will be more likely to favour public expenditure which benefits T.' There seem to be a number of difficulties with this generalisation. Firstly, the generalisation will need to key off O's view of what benefits T, because O does not have access to either T's view of what will benefit T or what may be very different, what will actually benefit T. This then raises the problem that the generalisation will indeed be systematically bad if O's views of what benefits T are adrift from T's views, but that that systematic ToM error does not seem apt to match the systematic error required to explain the data in this experiment. The data show that the generalisation canvassed above is totally false in that T is apparently not more likely to favour public expenditure which benefits T. Secondly, it also appears overly specified and implausible in what is after all an evolved ToM that it includes generalisations about public expenditure. On the other hand, if this is avoided by changing the generalisation to 'T will favour decisions that benefit T' it seems to be too weak to have sufficient predictive power. The situation will not become simpler when extended to us as S's outside the experiment.

The campus drinkers experiment (§9.2.8) appears similar to the previous one, but in the reverse direction. The inaccurate generalisation would appear

to be something like 'T will disfavour policy decisions that adversely affect T.' This generalisation is open to both of the objections set out previously; viz. systematic mismatch between what O believes will adversely affect T and what T believes will adversely affect T will not predict the data seen. This again shows that in this case, T does not disfavour policy decisions that adversely affect T. The generalisation similarly makes reference to policy decisions which seems implausible. Perhaps TT defenders can say that this inaccurate generalisation is not triggered in this case because T believes that in fact the policy decision will not adversely affect T: T may consider the longer-term health benefits of drinking less. Taking that way out though means that TT defenders will need to specify axiomatically when the inaccurate generalisation applies and when it does not. They will need a large number of further generalisations to do that. Again, there seem to be sufficient difficulties for the Inaccurate Generalisation Defence of TT within the experiment before we step outside to us as S's making ToM errors about the O's in the experiment.

The smokers experiment (§9.2.9) introduces yet another complication for TT defenders to add to the previous two. The previous two experiments showed the absence of self-serving bias in decisions made by T in different directions: firstly there was contrary to O's predictions no bias in favour of decisions seen as favouring T and secondly there was no bias against decisions seen as disfavouring T. Here we have a weaker variant of the first effect in that there was a bias towards decisions favouring T, but not as much as the O's predicted. This might allow TT defenders to avoid the problems described in the previous two sections, but here they will again have to face the Miscalibration Objection. Why is the generalisation not calibrated better to the quantum of bias actually observed? We might also note that when an experiment is conducted and an average result over a population given, that is generally not the answer all of the experimental subjects gave. They gave

a range of answers which were averaged to arrive at the answer given. So the Inaccurate Generalisation Defence now has to explain why there is such a range of miscalibrations. And as ever, we may ask what the account says that S's outside the experiment will predict about the O's.

The statement releasers experiment (§9.2.10) showed a similar miscalibration to the one described in the experiment discussed in the previous section. There was a self-serving bias in the responses given, but it was not as strong an effect as the S's predicted. All of the objections mentioned in the previous section apply. We may add to the problem about the range of answers given one as to what distinguishes this experiment from the previous one in terms of which inaccurate generalisations apply. If TT defenders wish to apply the same generalisation to this experiment and the previous one, they will have to explain why different percentage errors are found. If they wish to apply different generalisations to the experiment and the previous one, they will have to specify the new generalisations and provide a third generalisation which decides which of the first two generalisations will apply in which case. They will have to repeat that process for all possible current and future experiments.

10.3 Conclusions

Let us review progress made across the board. In table 10.1, I give in summary my assessment of the extent to which the Inaccurate Generalisation Defence of TT has been seen to be successful.

The results show a mixed picture. Bear in mind that the game has now changed, since ST combined with the Bias Mismatch Defence now has a parsimonious and comprehensive explanation of all these data, together with a great deal more where there are surprising data in social psychology resulting from biases. This data set will continue to expand over time, compounding the problem for TT.

Experiment	Success?	Issues
Shock Appliers	✓	'Justification' in generalisations
Fake Prison Guards	✓	'Justification' in generalisations
Repenters	✓	Requires full belief set of S
Quiz Gamers	?	Why does S have this generalisation?
Suicide Note Assessors	?	Conflict of generalisations?
Lottery Ticket Holders	✓	ToM static or dynamic?
Gamblers	✓	Conflict of generalisations?
Basketball Fans	✓	TT needs to add in biases
Cancer Cure Assessors	✓	Conflict of generalisations?
Puzzle Solvers	✗	test
Shoppers Redux	✓	Conflict of generalisations?
Conflict Parties	✗	ToM static or dynamic?
Marriage Partners	?	Miscalibration
Video Gamers	?	Miscalibration
Debaters	?	Miscalibration
Darts Players	?	Miscalibration
Blood Donors	?	Miscalibration
Healthcare Consumers	✗	Who benefits?
Campus Drinkers	✗	Opposite of above
Smokers	✗	Miscalibration
Statement Releasers	✗	Miscalibration

Table 10.1: Inaccurate Generalisation Defence: Data Issues

Observers might disagree with my assessment in a number of cases, and argue that in fact, the Inaccurate Generalisation Defence has been successful in more cases than I allow. So be it: it remains the case that the Inaccurate Generalisation Defence needs to explain all of the data in order to reach parity with augmented ST. And we may enquire as to the parsimony cost. I have several times asked about conflicts of generalisations. Remember that all TT has in such cases is more generalisations, which risks exacerbating the very problem the extra generalisations were introduced to address. There will be a need for a great number of generalisations, all of which will be wrong in all cases of systematic ToM error. The overall account, if fully specified as I insist it ought to be, will be severely lacking in parsimony if it can be produced at all.

Three issues came up several times. There is a major problem of how to resolve conflicts of generalisations; there is another one as to whether ToM is static or dynamic in adults and there is the problem of a wide array of miscalibrated generalisations. I conclude that while this set of problems does not do enough to terminate the viability of TT, it certainly poses a very difficult set of questions which must be dealt with if TT is to reach parity of plausibility with ST or Weak S/T Hybrids.

Chapter 11

Conclusions

I began in Ch. 1 by setting out the central task. This was to defend ST and Weak S/T Hybridism against a serious challenge from TT and Strong S/T Hybrid theorists. The challenge was that ST alone could not account for the observed systematic errors in ToM performance of different kinds under different types of scenario. No response had been provided by the ST side to this challenge, and therefore it could be seen that there was an urgent need for ST proponents to provide a response. Such a response is now in place, and it may now be seen that the response is comprehensive, parsimonious and convincing. I therefore conclude that Weak S/T Hybridism is a better account of ToM than had been thought. It is in a much stronger position than its close relation pure ST had been thought to be in.

The story began in earnest in Ch. 2 with a consideration of the possible logical variety of accounts of ToM. I examined both the scientific and Modular versions of TT, and the transformation and replication variants of ST. I spent some time considering further possible variants of ST, all of which appeared worthy of further consideration. I did not select a champion; the purpose of this thesis was to defend all versions of ST. The debate as to which version if any is superior may be deferred. The more serious problem of whether ST

was in fact separate to TT was considered; it was seen that the 'collapse risk' of ST entailing TT was in fact avoidable. Avoiding the error of 'setting the bar too low' was crucial here. With that in hand, one could be confident that there was a version of ST which would be separate from TT and could succeed without needing to decide which exact one it was.

In the next two chapters I examined difficulties with accounts that I aimed to challenge in this thesis. In Ch. 3, I considered some objections to pure TT accounts already proposed. I concluded that there were difficult objections to both of the types of TT that have received widespread support in the literature. I then moved on in Ch. 4 to consider whether the mainstream Strong S/T Hybrid approach could deal with these problems. Although mixing in some of TT(Innate) to a TT(Scientific) account does mean that some of the objections can be addressed, some of them still remain. The next idea was to add in some simulation also to the mix. Strong S/T Hybrid approaches must be the right answer, on this sort of view, because the challenge set out above to TT must mean that some ST is needed in response. It was seen though that there were two sorts of difficulty for Strong S/T Hybrid positions. One sort was that such Strong S/T Hybrid positions involve TT and therefore inherit all of the objections to TT (though we also saw that if Strong S/T Hybrids accept the lack of parsimony involved in including TT(Scientific), TT(Innate) as well as ST, they could solve three of the six objections). But a further unique sort of difficulty arises from trying to make TT and ST work together. These problems around resolving questions such as whether and how TT and ST interact and how that works led to difficulties for the mainstream Strong S/T Hybrid position.

I then, in Ch. 5, gave the TT opposition, ably represented by Saxe, their best case. I agreed that there was a serious problem for ST in explaining the systematic errors in ToM. I did not attempt to deny that these errors

occurred or that they were systematic. I agreed that TT could explain them parsimoniously by assuming the employment of a false ToM generalisation. I agreed that ST needed a response. I conceded that the difficulty for ST was sharpened by Saxe's astute observation that the errors were systematically different in different scenarios. Why would ToM errors be like that if ST were correct? After all, one of the ST claims is that we use our own minds to simulate our own minds in different circumstances, so how could be be wrong? I considered two different types of data introduced by Saxe: the 'too rosy' ToM error cases and the 'too cynical' cases. The conclusion at this stage was that the TT opposition to ST had a very strong case which needed answering.

The stage was then set for a new approach. I introduced the Bias Mismatch Defence in Ch. 6, suggesting that there were new resources available to ST to allow it to respond to the various forms of the systematic error challenge by allowing that S and O may apply different cognitive biases. As I suggested in Ch. 7, they may do so systematically because either they are differently involved affectively speaking in the particular scenario, or because they use different systems of reasoning. I outlined the various biases that would later be employed to explain a large array of data that TT proponents use to show systematic ToM error.

The next two chapters, Ch. 8 and Ch. 9, were the data-driven heart of the argument. I showed how different biases being applied by S and O could explain dozens of experiments in both the 'too rosy' and 'too cynical' directions of ToM error. Naturally, if some commentators suggest different combinations of biases to explain the data, that would constitute a 'friendly amendment,' remaining entirely consistent with the Bias Mismatch Defence.

In Ch. 10, I showed that constructing the inaccurate generalisations needed for various experiments was sometimes difficult. Given that the defence, to be deemed successful, needs to account for all or at least a wide variety of data,

this made it look as though the Inaccurate Generalisation Defence would not be a panacea for TT proponents. There were three major problems involving conflicts of generalisations, whether ToM is static or dynamic in adults and miscalibration of generalisations. Significantly more working out of the defence would be needed for TT proponents to be able to answer the mirror image of the question Saxe posed to simulationists: 'how can TT account for systematic ToM errors?' I concluded that TT must now handle a difficult set of questions which must be dealt with if TT is to reach parity of plausibility with ST or Weak S/T Hybrids.

In sum: we have seen that ST can not only respond to the systematic error challenge but it appears to do so more parsimoniously than the alternatives. We may therefore conclude that the current mainstream Strong S/T Hybrid/TT consensus is now in need of further support on the systematic error question while Weak S/T Hybridism, which is very close to pure ST, can now be seen as, pending that work, a better account of ToM.

Index

Bibliography

Ames, D. R. (2005). "Everyday Solutions To The Problem Of Other Minds: Which Tools Are Used When". In: *Other Minds*. Ed. by B. Malle and S. Hodges. Guilford Press, pp. 158–173. ISBN: 9781593854683. URL: http://www.worldcat.org/title/other-minds-how-humans-bridge-the-divide-between-self-and-others/oclc/59280150.

Andrews, K. (2008). "It's In Your Nature: A Pluralistic Folk Psychology". In: *Synthese* 165.1, pp. 13–29. DOI: 10.1007/S11229-007-9230-5.

Apperly, I. A. (2008). "Beyond Simulation-Theory And Theory-Theory: Why Social Cognitive Neuroscience Should Use Its Own Concepts To Study "Theory Of Mind" ". In: *Cognition* 107.1, pp. 266–283. DOI: 10.1016/J.Cognition.2007.07.019.

— (2009). "Alternative Routes To Perspective-Taking: Imagination And Rule-Use May Be Better Than Simulation And Theorising". In: *British Journal Of Developmental Psychology* 27.3, pp. 545–553. DOI: 10.1348/026151008X400841.

Arkway, A. (2000). "The Simulation Theory, The Theory Theory And Folk Psychological Explanation". In: *Philosophical Studies* 98.2, pp. 115–137. DOI: 10.1023/A%3A1018331121169.

Asch, S. E. (1946). "Forming Impressions Of Personality". In: *Journal Of Abnormal Psychology* 41.3, pp. 258–90. DOI: 10.1037/h0060423.

Asch, S. E. (1952). *Social Psychology*. Prentice-Hall. URL: http://www.worldcat.org/title/social-psychology/oclc/254969.

Avis, J. and P. L. Harris (1991). "Belief-Desire Reasoning Among Baka Children: Evidence For A Universal Conception Of Mind". In: *Child Development* 62.3, p. 460. DOI: 10.2307/1131123.

Bach, T. (2011). "Structure-Mapping: Directions From Simulation To Theory". In: *Philosophical Psychology* 24.1, pp. 23–51. DOI: 10.1080/09515089.2010.533261.

Baron-Cohen, S. (1993). "The Concept Of Intentionality: Invented Or Innate?" In: *Behavioral And Brain Sciences* 16 (01), pp. 29–30. DOI: 10.1017/S0140525X00028661.

— (2001). "Theory Of Mind In Normal Development And Autism". In: *Prisme* 34, pp. 174–183. URL: http://www.autism-community.com/wp-content/uploads/2010/11/TOM-in-TD-and-ASD.pdf.

Baron-Cohen, S., A. M. Leslie, and U. Frith (1985). "Does The Autistic Child Have A "Theory Of Mind" ?" In: *Cognition* 21.1, pp. 37–46. DOI: 10.1016/0010-0277(85)90022-8.

Bello, P. and N. Cassimatis (2006). "Developmental Accounts Of Theory-Of-Mind Acquisition: Achieving Clarity Via Computational Cognitive Modeling". In: *Proceedings Of The Twenty-Eighth Annual Meeting Of The Cognitive Science Society*. Ed. by R. Sun and N. Miyake, pp. 1014–1019. ISBN: 978-0-9768318-2-2. URL: http://www.amazon.co.uk/Proceedings-Conference-Cognitive-Science-Proceedin/dp/080586296X.

Bellugi, U. et al. (2007). "Affect, Social Behavior, And The Brain In Williams Syndrome". In: *Current Directions In Psychological Science* 16.2, pp. 99–104. DOI: 10.1111/j.1467-8721.2007.00484.x.

Biggs, Stephen (2007). "The Phenomenal Mindreader: A Case For Phenomenal Simulation". In: *Philosophical Psychology* 20.1, pp. 29–42. DOI: 10.1080/09515080601108013.

Bishop, M. A. and S. M. Downes (2002). "The Theory Theory Thrice Over: The Child As Scientist, Superscientist, Or Social Institution?" In: *Studies In History And Philosophy Of Science* 33.1, pp. 117–132. DOI: 10.1016/S0039-3681(01)00029-2.

Blackburn, S. W. (1992). "Theory, Observation, And Drama". In: *Mind And Language* 7.1-2, pp. 187–203. DOI: 10.1111/J.1468-0017.1992.Tb00204.X.

Bloom, P. and T. P. German (2000). "Two Reasons To Abandon The False Belief Task As A Test Of Theory Of Mind". In: *Cognition* 77.1, pp. 25–31. DOI: 10.1016/S0010-0277(00)00096-2.

Bora, E. and C. Pantelis (2012). "Poster #141 Theory Of Mind Impairment At Risk Conditions To Psychosis And In First-Degree Relatives Of Schizophrenia: Systematic Review And Meta-Analysis". In: *Schizophrenia Research* 136, Supplement 1. Abstracts Of The 3rd Biennial Schizophrenia International Research Conference, S142. DOI: 10.1016/S0920-9964(12)70455-3.

Boucher, J. (1996). "What Could Possibly Explain Autism?" In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 223–241. ISBN: 9780521559164. URL: http://www.worldcat.org/title/theories-of-theories-of-mind/oclc/32311136.

Butterfill, S. A. and I. A. Apperly (2013). "How To Construct A Minimal Theory Of Mind". In: *Mind And Language* 28.5, pp. 606–637. DOI: 10.1111/Mila.12036.

Carey, S. (2009). *The Origin of Concepts*. Developmental Cognitive Neuroscience. OUP USA. ISBN: 9780195367638. URL: http://www.worldcat.org/title/origin-of-concepts/oclc/233697385.

Carruthers, P. (1996). "Simulation And Self-Knowledge: A Defence Of Theory-Theory". In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 22–38. ISBN: 9780521559164. URL: http://www.worldcat.org/title/theories-of-theories-of-mind/oclc/32311136.

— (2009). "Simulation And The First-Person". In: *Philosophical Studies* 144.3, pp. 467–475. DOI: 10.1007/S11098-009-9357-Y.

Cassidy, K. W. et al. (2005). "Theory Of Mind May Be Contagious, But You Don't Catch It From Your Twin". In: *Child Development* 76.1, pp. 97–106. DOI: 10.1111/j.1467-8624.2005.00832.x.

Chalmers, D. J. (1997). *The Conscious Mind: In Search Of A Fundamental Theory*. Oxford University Press. ISBN: 9780195117899. URL: http://www.worldcat.org/title/conscious-mind/oclc/439674143.

Coplan, A. (2011). "Understanding Empathy: Its Features And Effects". In: *Empathy: Philosophical And Psychological Perspectives*. Ed. by A. Coplan and P. Goldie. Oxford University Press. ISBN: 9780198706427. URL: http://www.worldcat.org/title/empathy-philosophical-and-psychological-perspectives/oclc/767842620.

Crane, T. (2003). *The Mechanical Mind: A Philosophical Introduction To Minds, Machines And Mental Representation*. Taylor and Francis. ISBN: 9780203426319. URL: http://www.worldcat.org/title/mechanical-mind-a-philosophical-introduction-to-minds-machines-and-mental/oclc/437079517.

Daniel, S. (1993). "The Anthropology Of Folk Psychology". In: *Behavioral And Brain Sciences* 16 (01), pp. 38–39. DOI: 10.1017/S0140525X00028752.

Darley, J. M. and C. D. Batson (1973). " "From Jerusalem To Jericho": A Study Of Situational And Dispositional Variables In Helping Behavior". In: *Journal Of Personality And Social Psychology* 27.1, pp. 100–108. DOI: 10.1037/H0034449.

D'Ausilio, A. et al. (2009). "The Motor Somatotopy Of Speech Perception". In: *Current Biology : Cb* 19.5, pp. 381–385. DOI: 10.1016/J.Cub.2009.01.017.

Davidson, D. (1963). "Actions, Reasons, And Causes". In: *Journal Of Philosophy* 60, pp. 685–700. DOI: 10.1093/0199246270.003.0001.

Davies, M. and T. Stone (1995). "Introduction". In: *Folk Psychology: The Theory Of Mind Debate*. Ed. by M. Davies and T. Stone. Blackwell, pp. 1–44. ISBN: 9780631195153. URL: http://www.worldcat.org/title/folk-psychology-theory-of-mind-debate/oclc/301528886.

— (2001). "Mental Simulation, Tacit Theory, And The Threat Of Collapse". In: *Philosophical Topics* 29.1/2, pp. 127–173. DOI: 10.5840/philtopics2001291/212.

Dennett, D. C. (1979). *Brainstorms: Philosophical Essays On Mind And Psychology*. Harvester. ISBN: 9780855275853. URL: http://www.worldcat.org/title/brainstorms-philosophical-essays-on-mind-and-psychology/oclc/5329396.

— (1981). "True Believers: The Intentional Strategy And Why It Works". In: *Mind Design II*. Ed. by J. Haugeland. MIT Press, pp. 57–79. ISBN: 9780262082594. URL:

Http://www.Worldcat.Org/Title/Mind-Design-Ii-Philosophy-Psychology-Artificial-Intelligence/Oclc/42328967.

Dennett, D. C. (2007). *Breaking The Spell: Religion As A Natural Phenomenon*. Penguin Adult. ISBN: 9780141017778. URL: http://www.worldcat.org/title/breaking-the-spell-religion-as-a-natural-phenomenon/oclc/470564789.

Dimaggio, G. et al. (2008). "Know Yourself And You Shall Know The Other... To A Certain Extent: Multiple Paths Of Influence Of Self-Reflection On Mindreading". In: *Consciousness And Cognition* 17.3, pp. 778–789. DOI: 10.1016/J.Concog.2008.02.005.

Doherty, M.J. (2008). *Theory Of Mind: How Children Understand Others' Thoughts And Feelings*. Taylor and Francis. ISBN: 9781135420796. URL: http://www.worldcat.org/title/theory-of-mind-how-children-understand-others-thoughts-and-feelings/oclc/195720264.

Dreyfus, H. L. (2006). "Overcoming The Myth Of The Mental". In: *Topoi* 25.1-2, pp. 43–49. DOI: 10.1007/S11245-006-0006-1.

Editors, The (1992). "Introduction". In: *Mind and Language* 7.1-2, pp. 1–10. DOI: 10.1111/J.1468-0017.1992.Tb00194.X.

Edwards, K. and E. E. Smith (1996). "A Disconfirmation Bias In The Evaluation Of Arguments". In: *Journal Of Personality And Social Psychology* 71.1, pp. 5–24. DOI: 10.1037//0022-3514.71.1.5.

Epley, N. et al. (2004). "Perspective Taking As Egocentric Anchoring And Adjustment". In: *Journal Of Personality And Social Psychology* 87.3, pp. 327–339. DOI: 10.1037/0022-3514.87.3.327.

Evans, J. S. B. T. (1990). *Bias In Human Reasoning: Causes And Consequences*. Lawrence Erlbaum Associates, Incorporated. ISBN: 9780863771569. URL: http://www.worldcat.org/title/bias-in-human-reasoning-causes-and-consequences/oclc/35142289.

Fadiga, L. et al. (2002). "Speech Listening Specifically Modulates The Excitability Of Tongue Muscles". In: *European Journal Of Neuroscience* 15.2, pp. 399–402. DOI: 10.1046/J.0953-816X.2001.01874.X.

Farrant, B. M., J. Fletcher, and M. T. Maybery (2006). "Specific Language Impairment, Theory Of Mind, And Visual Perspective Taking: Evidence For Simulation Theory And The Developmental Role Of Language". In: *Child Development* 77.6, pp. 1842–1853. DOI: 10.1111/J.1467-8624.2006.00977.X.

Fodor, J. A. (1974). "Special Sciences (Or: The Disunity Of Science As A Working Hypothesis)". In: *Synthese* 28.2, pp. 97–115. DOI: 10.1007/Bf00485230.

— (1987). *Psychosemantics: The Problem Of Meaning In The Philosophy Of Mind*. MIT Press. ISBN: 9780262061063. URL: http://www.worldcat.org/title/psychosemantics-the-problem-of-meaning-in-the-philosophy-of-mind/oclc/45844220.

— (1994). "Concepts: A Potboiler". In: *Cognition* 50.1–3, pp. 95–113. DOI: 10.1016/0010-0277(94)90023-X.

— (2008). *LOT 2: The Language Of Thought Revisited*. Oxford University Press. ISBN: 9780191563478. URL: http://www.worldcat.org/title/lot-2-the-language-of-thought-revisited/oclc/299380071.

Fodor, J. A. and E. Lepore (1996). "The Red Herring And The Pet Fish: Why Concepts Still Can't Be Prototypes". In: *Cognition* 58.2, pp. 253–270. DOI: 10.1016/0010-0277(95)00694-X.

Fredrickson, B. L. and T-A. Roberts (1997). "Objectification Theory". In: *Psychology Of Women Quarterly* 21.2, pp. 173–206. DOI: 10.1111/J.1471-6402.1997.Tb00108.X.

Freeman, N. H. (1995). "Theories Of Mind In Collision: Plausibility And Authority". In: *Mental Simulation: Evaluations And Applications*. Ed. by M. Davies and T. Stone. Blackwell, pp. 68–86. ISBN: 9780631198734. URL: http://www.worldcat.org/title/mental-simulation-evaluations-and-applications/oclc/495403648.

Friedman, O. and A. R. Petrashek (2009). "Children Do Not Follow The Rule "Ignorance Means Getting It Wrong"". In: *Journal Of Experimental Child Psychology* 102.1, pp. 114–121. DOI: 10.1016/J.Jecp.2008.07.009.

Fuller, T. (2013). "Is Scientific Theory Change Similar To Early Cognitive Development? Gopnik On Science And Childhood". In: *Philosophical Psychology* 26.1, p. 109. DOI: 10.1080/09515089.2011.625114.

Gallagher, S. and D. Hutto (2008). "Understanding Others Through Primary Interaction And Narrative Practice". In: *The Shared Mind: Perspectives On Intersubjectivity*. 1, pp. 1–18. URL: http://www.worldcat.org/title/shared-mind-perspectives-on-intersubjectivity/oclc/273893638.

Gallese, V. and A. I. Goldman (1998). "Mirror Neurons And The Simulation Theory Of Mind-Reading". In: *Trends In Cognitive Sciences* 2.12, pp. 493–501. DOI: 10.1016/S1364-6613(98)01262-5.

Gallese, V. and C. Sinigaglia (2011). "What Is So Special About Embodied Simulation?" In: *Trends In Cognitive Sciences* 15.11, pp. 512–519. DOI: 10.1016/J.Tics.2011.09.003.

Garson, J. W. (2003). "Simulation And Connectionism: What Is The Connection?" In: *Philosophical Psychology* 16.4, pp. 499–514. DOI: 10.1080/0951508032000121805.

Gilovich, T. (1993). *How We Know What Isn't So*. Free Press. ISBN: 9780029117064. URL:

http://www.worldcat.org/title/how-we-know-what-isnt-so-the-fallibility-of-human-reason-in-everyday-life/oclc/832440458.

Glymour, C. (2000). "Android Epistemology For Babies: Reflections On "Words,
Thoughts And Theories"". In: *Synthese* 122.1/2, pp. 53–68. URL: http://www.jstor.org/stable/20118243.

Godfrey-Smith, P. (2005). "Folk Psychology As A Model". In: *Philosophers' Imprint* 5.6, pp. 1–16. URL: http://hdl.handle.net/2027/spo.3521354.0005.006.

Goldie, P. (1999). "How We Think Of Others' Emotions". In: *Mind And Language* 14.4, pp. 394–423. DOI: 10.1111/1468-0017.00118.

— (2002). *The Emotions*. Clarendon Press. ISBN: 9780199253043. URL: http://www.worldcat.org/title/emotions-a-philosophical-exploration/oclc/807099376.

— (2011). "Grief: A Narrative Account". In: *Ratio* 24.2, pp. 119–137. DOI: 10.1111/J.1467-9329.2011.00488.X.

Goldman, A. I. (1989). "Interpretation Psychologized*". In: *Mind And Language* 4.3, pp. 161–185. DOI: 10.1111/J.1468-0017.1989.Tb00249.X.

— (1992). "In Defense Of The Simulation Theory". In: *Mind And Language* 7.1-2, pp. 104–119. DOI: 10.1111/J.1468-0017.1992.Tb00200.X.

— (1993). "Functionalism, The Theory-Theory And Phenomenology". In: *Behavioral And Brain Sciences* 16 (01), pp. 101–113. DOI: 10.1017/S0140525X00029265.

— (2006). *Simulating Minds: The Philosophy, Psychology, And Neuroscience Of Mindreading*. Oxford University Press, USA. ISBN: 9780198031765. URL: https://www.worldcat.org/title/simulating-

minds-the-philosophy-psychology-and-neuroscience-of-mindreading/oclc/63390792.

Goldman, A. I. (2009). "Replies To Perner And Brandl, Saxe, Vignemont, And Carruthers". In: *Philosophical Studies* 144.3, pp. 477–491. DOI: 10.1007/S11098-009-9358-X.

Goldman, A. I. and N. Sebanz (2005). "Simulation, Mirroring, And A Different Argument From Error". In: *Trends In Cognitive Sciences* 9.7, p. 320. DOI: 10.1016/J.Tics.2005.05.008.

Gopnik, A. (1993). "Theories And Illusions". In: *Behavioral And Brain Sciences* 16 (01), pp. 90–100. DOI: 10.1017/S0140525X00029253.

— (1996). "Theories And Modules, Creation Myths, Developmental Realities And Neurath's Boat". In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 169–183. ISBN: 9780521559164. URL: http://www.worldcat.org/title/theories-of-theories-of-mind/oclc/32311136.

Gopnik, A. and J. W. Astington (1988). "Children's Understanding Of Representational Change And Its Relation To The Understanding Of False Belief And The Appearance-Reality Distinction". In: *Child Development* 59.1, p. 26. DOI: 10.2307/1130386.

Gopnik, A. and H. M. Wellman (1992). "Why The Child's Theory Of Mind Really Is A Theory". In: *Mind And Language* 7.1-2, pp. 145–171. DOI: 10.1111/J.1468-0017.1992.Tb00202.X.

Gordon, R. M. (1986). "Folk Psychology As Simulation". In: *Mind And Language* 1.2, pp. 158–171. DOI: 10.1111/J.1468-0017.1986.Tb00324.X.

— (1992). "The Simulation Theory: Objections And Misconceptions". In: *Mind And Language* 7.1-2, pp. 11–34. DOI: 10.1111/J.1468-0017.1992.Tb00195.X.

— (1995a). "Simulation Without Introspection Or Inference From Me To You". In: *Mental Simulation: Evaluations And Applications*. Ed. by M. Davies and T. Stone. Blackwell, pp. 53–67. ISBN: 9780631198734. URL: http://www.worldcat.org/title/mental-simulation-evaluations-and-applications/oclc/495403648.

— (1995b). "Sympathy, Simulation, And The Impartial Spectator". In: *Ethics* 105.4, pp. 727–742. DOI: 10.1086/293750.

— (2005). "Simulation And Systematic Errors In Prediction". In: *Trends In Cognitive Sciences* 9.8, pp. 361–362. DOI: 10.1016/J.Tics.2005.06.003.

— (2009). "Folk Psychology As Mental Simulation". In: *Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University. URL: http://plato.stanford.edu/entries/folkpsych-simulation/.

Grafton, S. T. (2009). "Embodied Cognition And The Simulation Of Action To Understand Others". In: *Annals Of The New York Academy Of Sciences* 1156, pp. 97–117. DOI: 10.1111/J.1749-6632.2009.04425.X.

Greenwood, J. D. (1999). "Simulation, Theory-Theory And Cognitive Penetration: No 'Instance Of The Fingerpost' ". In: *Mind And Language* 14.1, pp. 32–56. DOI: 10.1111/1468-0017.00102.

Gustafson, D. (1995). "Eighteen Months On The Planet And Already A Psychological Theorist". In: *Philosophical Psychology* 8.2, pp. 125–137. DOI: 10.1080/09515089508573149.

Hale, C. M. and H. Tager-Flusberg (2003). "The Influence Of Language On Theory Of Mind: A Training Study". In: *Developmental Science* 6.3, pp. 346–359. DOI: 10.1111/1467-7687.00289.

Haney, C., C. Banks, and P. Zimbardo (1973). "Interpersonal Dynamics In A Simulated Prison". In: *International Journal Of Criminology And Penology* 1.1, pp. 69–97. URL: http://www.prisonexp.org/pdf/ijcp1973.pdf.

Harris, P. L. (1992). "From Simulation To Folk Psychology: The Case For Development". In: *Mind And Language* 7.1-2, pp. 120–144. DOI: 10.1111/J.1468-0017.1992.Tb00201.X.

— (2009). "Simulation (Mostly) Rules: A Commentary". In: *British Journal Of Developmental Psychology* 27.3, pp. 555–559. DOI: 10.1348/026151009X415484.

Heal, J. (1996). "Simulation, Theory, And Content". In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 75–89.

— (1998). "Understanding Other Minds From The Inside". In: *Royal Institute Of Philosophy Supplement* 43, pp. 83–99. DOI: 10.1017/Cbo9780511615894.005.

— (2000). "The Inaugural Address: Other Minds, Rationality And Analogy". In: *Aristotelian Society Supplementary Volume* 74.1, pp. 1–19. DOI: 10.1111/1467-8349.T01-1-00060.

— (2003). *Mind, Reason And Imagination: Selected Essays In Philosophy Of Mind And Language*. Cambridge University Press. ISBN: 9780521017169. URL: http://www.worldcat.org/title/mind-reason-and-imagination-selected-essays-in-philosophy-of-mind-and-language/oclc/470066271.

Helming, K. A., B. Strickland, and P. Jacob (2014). "Making Sense Of Early False-Belief Understanding". In: *Trends In Cognitive Sciences* 18.4, pp. 167–170. DOI: 10.1016/J.Tics.2014.01.005.

Heyes, C. (2014). "False Belief In Infancy: A Fresh Look". In: *Developmental Science* 17.5, pp. 647–659. DOI: 10.1111/desc.12148.

Hughes, C. et al. (2006). "Cooperation And Conversations About The Mind: A Study Of Individual Differences In 2-year-olds And Their Siblings". In: *British Journal of Developmental Psychology* 24.1, pp. 53–72. DOI: 10.1348/026151005X82893.

Hume, D. (1993). *An Enquiry Concerning Human Understanding: With A Letter From A Gentleman To His Friend In Edinburgh ; And An Abstract Of A Treatise Of Human Nature*. Ed. by E. Steinberg. Hackett Publishing Company Incorporated. ISBN: 9780872202290. URL: http://www.worldcat.org/title/enquiry-concerning-human-understanding-with-a-letter-from-a-gentleman-to-his-friend-in-edinburgh-and-an-abstract-of-a-treatise-of-human-nature/oclc/28339533.

— (2000). *A Treatise Of Human Nature: Being An Attempt To Introduce The Experimental Method Of Reasoning Into Moral Subjects*. Oxford University Press. ISBN: 9780198751724. URL: http://www.worldcat.org/title/treatise-of-human-nature/oclc/41981924.

Igoe, A. R. and H. Sullivan (1993). "Self-Presentation Bias And Continuing Motivation Among Adolescents". In: *The Journal Of Educational Research* 87.1, pp. 18–22. DOI: 10.1080/00220671.1993.9941161.

Ivry, R. B. and T. C. Justus (2001). "A Neural Instantiation Of The Motor Theory Of Speech Perception." In: *Trends Neurosci* 24.9, pp. 513–5. DOI: 10.1016/S0166-2236(00)01897-X.

Jackson, F. (1999). "All That Can Be At Issue In The Theory-Theory Simulation Debate". In: *Philosophical Papers* 28.2, pp. 77–96. DOI: 10.1080/05568649909506593.

Johansson, P. et al. (2006). "How Something Can Be Said About Telling More Than We Can Know: On Choice Blindness And Introspection". In: *Consciousness And Cognition* 15.4, pp. 673–692. DOI: 10.1016/J.Concog.2006.09.004.

Kahneman, D. (2011). *Thinking, Fast And Slow*. Allen Lane. ISBN: 9781846140556. URL: http://www.worldcat.org/title/thinking-fast-and-slow/oclc/751738755.

Kamtekar, R. (2004). "Situationism And Virtue Ethics On The Content Of Our Character". In: *Ethics* 114.3, pp. 458–491. DOI: 10.1086/381696.

Kaplan, J. T. and M. Iacoboni (2006). "Getting A Grip On Other Minds: Mirror Neurons, Intention Understanding, And Cognitive Empathy". In: *Social Neuroscience* 1.3–4, pp. 175–183. DOI: 10.1080/17470910600985605.

Karmiloff-Smith, A. (1998). "Development Itself Is The Key To Understanding Developmental Disorders". In: *Trends in Cognitive Sciences* 2.10, pp. 389–398. DOI: 10.1016/S1364-6613(98)01230-3.

Koelkebeck, K. et al. (2010). "Theory Of Mind In First-Episode Schizophrenia Patients: Correlations With Cognition And Personality Traits". In: *Schizophrenia Research* 119.1-3, pp. 115–123. DOI: 10.1016/J.Schres.2009.12.015.

Kopcha, T. J. and H. Sullivan (2006). "Self-Presentation Bias In Surveys Of Teachers' Educational Technology Practices". In: *Educational Technology Research And Development* 55.6, pp. 627–646. DOI: 10.1007/S11423-006-9011-8.

Kruger, J. and T. Gilovich (1999). " "Naive Cynicism" In Everyday Theories Of Responsibility Assessment: On Biased Assumptions Of Bias." In: *Journal Of Personality And Social Psychology* 76.5, pp. 743–753. DOI: 10.1037/0022-3514.76.5.743.

Kuehberger, A et al. (1995). "Choice Or No Choice: Is The Langer Effect Evidence Against Simulation?" In: *Mind And Language* 10.4, pp. 423–436. DOI: 10.1111/J.1468-0017.1995.Tb00022.X.

Kunda, Z. (1990). "The Case For Motivated Reasoning." In: *Psychological Bulletin* 108.3, pp. 480–498. DOI: 10.1037/0033-2909.108.3.480.

Leslie, A. M. (1987). "Pretense And Representation: The Origins Of "Theory Of Mind"". In: *Psychological Review* 94.4, pp. 412–426. DOI: 10.1037//0033-295X.94.4.412.

Leslie, A. M., O. Friedman, and T. P. German (2004). "Core Mechanisms In "Theory Of Mind"". In: *Trends In Cognitive Sciences* 8.12, pp. 528–33. DOI: 10.1016/J.Tics.2004.10.001.

Leslie, A. M., T. P. German, and F. G. Happe (1993). "Even A Theory-Theory Needs Information Processing: ToMM, An Alternative Theory-Theory Of The Child's Theory Of Mind". In: *Behavioral And Brain Sciences* 16 (01), pp. 56–57. DOI: 10.1017/S0140525X00028934.

Liberman, A. (1985). "The Motor Theory Of Speech Perception Revised". In: *Cognition* 21.1, pp. 1–36. DOI: 10.1016/0010-0277(85)90021-6.

Lohmann, H. and M. Tomasello (2003). "The Role of Language In The Development Of False Belief Understanding: A Training Study". In: *Child Development* 74.4, pp. 1130–1144. DOI: 10.1111/1467-8624.00597.

Luo, Y. (2011). "Do 10-month-old Infants Understand Others' False Beliefs?" In: *Cognition* 121.3, pp. 289–298. DOI: 10.1016/j.cognition.2011.07.011.

McKay, R. T. and D. C. Dennett (2009). "The Evolution Of Misbelief". In: *Behavioral And Brain Sciences* 32 (06), pp. 493–510. DOI: 10.1017/S0140525X09990975.

Mealey, L. and S. Kinner (2002). "The Perception-Action Model Of Empathy And Psychopathic "Cold-Heartedness"". In: *Behavioral And Brain Sciences* 25 (01), pp. 42–43. DOI: 10.1017/S0140525X02460013.

Milgram, S. (1963). "Behavioral Study Of Obedience." In: *The Journal Of Abnormal And Social Psychology* 67.4, p. 371. DOI: 10.1037/H0040525.

— (1974). *Obedience To Authority: An Experimental View*. Harper & Row. ISBN: 9780060129385. URL: http://www.worldcat.org/title/obedience-to-authority-an-experimental-view/oclc/668026.

Miller, D. and R. Ratner (1998). "The Disparity Between The Actual And Assumed Power Of Self-Interest". In: *Journal Of Personality And Social Psychology* 74.1, pp. 53–62. DOI: 10.1037//0022-3514.74.1.53.

Mineka, S. and S. K. Sutton (1992). "Cognitive Biases And The Emotional Disorders". In: *Psychological Science* 3.1, pp. 65–69. DOI: 10.1111/J.1467-9280.1992.Tb00260.X.

Mitchell, J. P. (2005). "The False Dichotomy Between Simulation And Theory-Theory: The Argument's Error". In: *Trends In Cognitive Sciences* 9.8, pp. 363–364. DOI: 10.1016/J.Tics.2005.06.010.

Mitchell, P., G. Currie, and F. Ziegler (2009a). "Is There An Alternative To Simulation And Theory In Understanding The Mind?" In: *British Journal Of Developmental Psychology* 27.3, pp. 561–567. DOI: 10.1348/026151009X441935.

— (2009b). "Two Routes To Perspective: Simulation And Rule-Use As Approaches To Mentalizing". In: *British Journal Of Developmental Psychology* 27.3, pp. 513–543. DOI: 10.1348/026151008X334737.

Moeller, M. P. and B. Schick (2006). "Relations Between Maternal Input And Theory Of Mind Understanding In Deaf Children". In: *Child*

Development 77.3, pp. 751–766. DOI: 10.1111/j.1467-8624.2006.00901.x.

Morin, A. (2007). "Self-Awareness And The Left Hemisphere: The Dark Side Of Selectively Reviewing The Literature". In: *Cortex* 43.8, pp. 1068–1073. DOI: 10.1016/S0010-9452(08)70704-4.

Nagel, J. (2011). "The Psychological Basis Of The Harman-Vogel Paradox". In: *Philosophers' Imprint* 11.5, pp. 1–28. URL: http://hdl.handle.net/2027/spo.3521354.0011.005.

Nestler, S. (2010). "Belief Perseverance". In: *Social Psychology* 41.1, pp. 35–41. DOI: 10.1027/1864-9335/A000006.

Nichols, S. and S. P. Stich (2003). *Mindreading: An Integrated Account Of Pretence, Self-Awareness, And Understanding Other Minds*. Clarendon. ISBN: 9780198236108. DOI: 10.1093/0198236107.001.0001.

Nichols, S., S. P. Stich, and A. M. Leslie (1995). "Choice Effects And The Ineffectiveness Of Simulation: Response To Kuehberger Et Al." In: *Mind And Language* 10.4, pp. 437–445. DOI: 10.1111/J.1468-0017.1995.Tb00023.X.

Nichols, S., S. P. Stich, A. M. Leslie, and D. Klein (1996). "Varieties Of Off-Line Simulation". In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 39–74. ISBN: 9780521559164. URL: http://www.worldcat.org/title/theories-of-theories-of-mind/oclc/32311136.

Nico, D. and E. Daprati (2009). "The Egocentric Reference For Visual Exploration And Orientation". In: *Brain And Cognition* 69.2, pp. 227–235. DOI: 10.1016/J.Bandc.2008.07.011.

Nisbett, R. E. and N. Bellows (1977). "Verbal Reports About Causal Influences On Social Judgments: Private Access Versus Public Theories".

In: *Journal Of Personality And Social Psychology* 35.9, pp. 613–624. DOI: 10.1037//0022-3514.35.9.613.

Nisbett, R. E. and T. D. Wilson (1977). "Telling More Than We Can Know: Verbal Reports On Mental Processes". In: *Psychological Review* 84.3, pp. 231–59. DOI: 10.1037//0033-295X.84.3.231.

Oberman, L. M. and V. S. Ramachandran (2007). "The Simulating Social Mind: The Role Of The Mirror Neuron System And Simulation In The Social And Communicative Deficits Of Autism Spectrum Disorders". In: *Psychological Bulletin* 133.2, pp. 310–327. DOI: 10.1037/0033-2909.133.2.310.

Onishi, K. H. and R. Baillargeon (2005). "Do 15-Month-Old Infants Understand False Beliefs?" In: *Science* 308.5719, pp. 255–258. DOI: 10.1126/Science.1107621.

O'Shaughnessy, B. (1991). "The Anatomy Of Consciousness". In: *Philosophical Issues* 1, pp. 135–177. DOI: 10.2307/1522927.

Perner, J. (2000). "About + Belief + Counterfactual". In: *Children's Reasoning And The Mind*. Ed. by P. Mitchell and K. J. Riggs. Psychology Press. ISBN: 9781317715221. URL: http://www.worldcat.org/title/childrens-reasoning-and-the-mind/oclc/41958990.

Peterson, C. and V. Slaughter (2003). "Opening Windows Into The Mind: Mothers' Preferences For Mental State Explanations And Children's Theory Of Mind". In: *Cognitive Development* 18.3, pp. 399–429. DOI: 10.1016/S0885-2014(03)00041-8.

Peterson, D. M. and K. J. Riggs (1999). "Adaptive Modelling And Mindreading". In: *Mind And Language* 14.1, pp. 80–112. DOI: 10.1111/1468-0017.00104.

Plotkin, H. (2011). "Human Nature, Cultural Diversity And Evolutionary Theory". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1563, pp. 454–463. DOI: 10.1098/rstb.2010.0160.

Pratt, I. (1993). "Matching And Mental-State Ascription". In: *Behavioral And Brain Sciences* 16 (01), pp. 71–72. DOI: 10.1017/S0140525X00029071.

Premack, D. and G. Woodruff (1978). "Does The Chimpanzee Have A Theory Of Mind?" In: *Behavioral And Brain Sciences* 1 (04), pp. 515–526. DOI: 10.1017/S0140525X00076512.

Prentice, R. A (2007). "Ethical Decision Making: More Needed Than Good Intentions". In: *Financial Analysts Journal* 63.6, pp. 17–30. DOI: 10.2469/Faj.V63.N6.4923.

Preston, S. D. and F. B. M. De Waal (2002). "Empathy: Its Ultimate And Proximate Bases". In: *Behavioral And Brain Sciences* 25 (01), pp. 1–20. DOI: 10.1017/S0140525X02000018.

Prinz, J. J. (2006). "The Emotional Basis Of Moral Judgments". In: *Philosophical Explorations* 9.1, pp. 29–43. DOI: 10.1080/13869790500492466.

— (2011). "Is Empathy Necessary For Morality". In: *Empathy: Philosophical And Psychological Perspectives*. Ed. by A. Coplan and P. Goldie. Oxford University Press, pp. 211–229. ISBN: 9780199539956. URL: http://www.worldcat.org/title/empathy-philosophical-and-psychological-perspectives/oclc/767842620.

Pronin, E., T. Gilovich, and L. Ross (2004). "Objectivity In The Eye Of The Beholder: Divergent Perceptions Of Bias In Self Versus Others". In: *Psychological Review* 111.3, pp. 781–799. DOI: 10.1037/0033-295X.111.3.781.

Pronin, E., C. Puccio, and L. Ross (2002). "Understanding Misunderstanding: Social Psychological Perspectives". In: *Heuristics And Biases: The Psychology Of Intuitive Judgment*. Ed. by T. Gilovich, D. Griffin, and D. Kahneman. Cambridge University Press. ISBN: 9780521796798. URL: http://www.worldcat.org/title/heuristics-and-biases-the-psychology-of-intuitive-judgement/oclc/47364085.

Purshouse, L. (2001). "Embarrassment: A Philosophical Analysis". In: *Philosophy* 76.298, pp. 515–540. DOI: 10.1017/S0031819101000547.

Rey, G. (2013). "We Are Not All 'Self-Blind': A Defense Of A Modest Introspectionism". In: *Mind And Language* 28.3, pp. 259–285. DOI: 10.1111/Mila.12018.

Riggs, K. J. and D. M. Peterson (2000). "Counterfactual Thinking In Preschool Children: Mental State And Causal Inferences". In: *Children's Reasoning And The Mind*. Ed. by P. Mitchell and K. J. Riggs. Psychology Press. ISBN: 9780863778551. URL: http://www.worldcat.org/title/childrens-reasoning-and-the-mind/oclc/41958990.

Rochat, P. (2002). "Various Kinds Of Empathy As Revealed By The Developing Child, Not The Monkey's Brain". In: *Behavioral And Brain Sciences* 25 (01), pp. 45–46. DOI: 10.1017/S0140525X02490012.

Ross, L., T. M. Amabile, and J. L. Steinmetz (1977). "Social Roles, Social Control, And Biases In Social-Perception Processes". In: *Journal Of Personality And Social Psychology* 35.7, pp. 485–494. DOI: 10.1037//0022-3514.35.7.485.

Ross, L., D. Greene, and P. House (1977). "The "False Consensus Effect": An Egocentric Bias In Social Perception And Attribution Processes". In:

Journal Of Experimental Social Psychology 13.3, pp. 279–301. DOI: 10.1016/0022-1031(77)90049-X.

Ross, L., M. R. Lepper, and M. Hubbard (1975). "Perseverance In Self Perception And Social Perception: Biased Attributional Processes In The Debeliefing Paradigm". In: *Journal Of Personality And Social Psychology* 32.5, pp. 880–892. DOI: 10.1037//0022-3514.32.5.880.

Ruffman, T. (1996). "Do Children Understand The Mind By Means Of Simulation Or A Theory? Evidence From Their Understanding Of Inference". In: *Mind And Language* 11.4, pp. 388–414. DOI: 10.1111/J.1468-0017.1996.Tb00053.X.

Ruffman, T. et al. (1998). "Older (But Not Younger) Siblings Facilitate False Belief Understanding". In: *Developmental Psychology* 34.1, pp. 161–174. DOI: 10.1037/0012-1649.34.1.161.

Ryle, G. (2009). *The Concept Of Mind*. Routledge. ISBN: 9780415485470. URL: http://www.worldcat.org/title/concept-of-mind/oclc/297405035.

Samson, D. et al. (2010). "Seeing It Their Way: Evidence For Rapid And Involuntary Computation Of What Other People See". In: *Journal of Experimental Psychology: Human Perception and Performance* 36.5, pp. 1255–1266. DOI: 10.1037/a0018729.

Saxe, R. (2005a). "Against Simulation: The Argument From Error". In: *Trends In Cognitive Sciences* 9.4, pp. 174–179. DOI: 10.1016/J.Tics.2005.01.012.

— (2005b). "Hybrid Vigour: Reply To Mitchell". In: *Trends In Cognitive Sciences* 9.8, p. 364. DOI: 10.1016/J.Tics.2005.06.017.

— (2005c). "On Ignorance And Being Wrong: Reply To Gordon". In: *Trends In Cognitive Sciences* 9.8, pp. 362–363. DOI: 10.1016/J.Tics.2005.06.002.

Saxe, R. (2005d). "Tuning Forks In The Mind: Reply To Goldman And Sebanz". In: *Trends In Cognitive Sciences* 9.7, p. 321. DOI: 10.1016/J.Tics.2005.05.011.

— (2009). "The Happiness Of The Fish: Evidence For A Common Theory Of One's Own And Others' Actions". In: *Handbook Of Imagination And Mental Simulation*. Ed. by K. D. Markman, W. M. P. Klein, and J. A. Suhr. Psychology Press. ISBN: 9781841698878. URL: http://www.worldcat.org/title/handbook-of-imagination-and-mental-simulation/oclc/222134918.

Scholl, B. J. and A. M. Leslie (1999). "Modularity, Development And 'Theory Of Mind' ". In: *Mind And Language* 14.1, pp. 131–153. DOI: 10.1111/1468-0017.00106.

— (2001). "Minds, Modules, And Meta-Analysis". In: *Child Development* 72.3, pp. 696–701. DOI: 10.1111/1467-8624.00308.

Segal, G. M. A. (1996). "The Modularity Of Theories Of Mind". In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 141–157. ISBN: 9780521559164. URL: http://www.worldcat.org/title/theories-of-theories-of-mind/oclc/32311136.

Sevdalis, N. and N. Harvey (2007). "Biased Forecasting Of Postdecisional Affect". In: *Psychological Science* 18.8, pp. 678–681. DOI: 10.1111/j.1467-9280.2007.01958.x.

Shanahan, M (2009). "The Frame Problem". In: *Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University. URL: http://plato.stanford.edu/archives/win2009/entries/frame-problem/.

Short, T. L. (1992). "The Design Of The ZEUS Regional First Level Trigger Box And Associated Trigger Studies". PhD Thesis. University of Bristol. URL: http://discovery.ucl.ac.uk/1354624/.

— (2012). "Nietzsche On Memory". MPhilStud Thesis. UCL. URL: http://discovery.ucl.ac.uk/1421265/.

— (2014). "How Can We Reconcile The Following Apparent Truths: 'Sherlock Holmes Doesn't Exist' And 'Sherlock Holmes Was Created By Conan Doyle'?" In: *Opticon1826* 16.8, pp. 1–9. DOI: 10.5334/Opt.Bs.

— (2015). *Simulation Theory: A Psychological And Philosophical Consideration*. Taylor and Francis. ISBN: 9781138816053. URL: http://www.routledge.com/books/details/9781138816053/.

Slaughter, V. and A. Gopnik (1996). "Conceptual Coherence In The Child's Theory Of Mind: Training Children To Understand Belief". In: *Child Development* 67.6, pp. 2967–2988. DOI: 10.2307/1131762.

Sloman, S. A. (1996). "The Empirical Case For Two Systems Of Reasoning". In: *Psychological Bulletin* 119, pp. 3–22. DOI: 10.1037//0033-2909.119.1.3.

Soteriou, M. (2013). *The Mind's Construction: The Ontology Of Mind And Mental Action*. OUP Oxford. ISBN: 9780199678457. URL: http://www.worldcat.org/title/minds-construction-the-ontology-of-mind-and-mental-action/oclc/829743933.

Stanley, J. (2011). *Know How*. OUP Oxford. ISBN: 9780199695362. URL: http://www.worldcat.org/title/know-how/oclc/706025101.

Stich, S. P. and S. Nichols (1992). "Folk Psychology: Simulation Or Tacit Theory?" In: *Mind And Language* 7.1-2, pp. 35–71. DOI: 10.1111/J.1468-0017.1992.Tb00196.X.

— (1995a). "Folk Psychology: Simulation Or Tacit Theory?" In: *Folk Psychology: The Theory Of Mind Debate*. Ed. by M. Davies and T. Stone.

Blackwell, pp. 123–158. ISBN: 9780631195153. URL: http://www.worldcat.org/title/folk-psychology-theory-of-mind-debate/oclc/301528886.

Stich, S. P. and S. Nichols (1995b). "Second Thoughts On Simulation". In: *Mental Simulation: Evaluations And Applications*. Ed. by M. Davies and T. Stone. Blackwell, pp. 87–108. ISBN: 9780631198734. URL: http://www.worldcat.org/title/mental-simulation-evaluations-and-applications/oclc/495403648.

— (1998). "Theory Theory To The Max". In: *Mind And Language* 13.3, pp. 421–449. DOI: 10.1111/1468-0017.00085.

— (2002). "Folk Psychology". In: *Blackwell Guide To Philosophy Of Mind*. Ed. by S. P. Stich and T. A. Warfield. Vol. 7. 1-2. Blackwell, pp. 35–71. ISBN: 9780470998762. DOI: 10.1002/9780470998762.Ch10.

Stone, T. and M. Davies (1996). "The Mental Simulation Debate: A Progress Report". In: *Theories Of Theories Of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge University Press, pp. 119–137. ISBN: 9780521559164. URL: http://www.worldcat.org/title/theories-of-theories-of-mind/oclc/32311136.

Strijbos, D. W. and L. C. De Bruin (2013). "Universal Belief-Desire Psychology? A Dilemma For Theory Theory And Simulation Theory". In: *Philosophical Psychology* 26.5, p. 744. DOI: 10.1080/09515089.2012.711034.

Taleb, N. N. (2007). *Fooled By Randomness: The Hidden Role Of Chance In Life And In The Markets*. Penguin Books Limited. ISBN: 9780141930237. URL: http://www.worldcat.org/title/fooled-by-randomness-the-hidden-role-of-chance-in-life-and-in-the-markets/oclc/827950242.

— (2008). *The Black Swan: The Impact Of The Highly Improbable*. Penguin Books Limited. ISBN: 9780141034591. URL: http://www.worldcat.org/title/black-swan-the-impact-of-the-highly-improbable/oclc/175283761.

Tversky, A. and D. Kahneman (1973). "Availability: A Heuristic For Judging Frequency And Probability". In: *Cognitive Psychology* 5.2, pp. 207–232. DOI: 10.1016/0010-0285(73)90033-9.

— (1974). "Judgment under Uncertainty: Heuristics and Biases." In: *Science (New York, N.Y.)* 185.4157, pp. 1124–1131. ISSN: 0036-8075. DOI: 10.1126/science.185.4157.1124.

— (1983). "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy In Probability Judgment". In: *Psychological Review* 90.4, pp. 293–315. DOI: 10.1037/0033-295X.90.4.293.

Wellman, H. M. and C. Peterson (2013). "Deafness, Thought Bubbles, And Theory-Of-Mind Development". In: *Developmental Psychology* 49.12, pp. 2357–2367. DOI: 10.1037/a0032419.

White, P. A. (1988). "Knowing More About What We Can Tell: 'Introspective Access' And Causal Report Accuracy 10 Years Later". In: *British Journal Of Psychology* 79.1, pp. 13–45. DOI: 10.1111/J.2044-8295.1988.Tb02271.X.

Wilkerson, W. S. (2001). "Simulation, Theory, And The Frame Problem: The Interpretive Moment". In: *Philosophical Psychology* 14.2, pp. 141–153. DOI: 10.1080/09515080120051535.

Wilkinson, M. R. and L. J. Ball (2012). "Why Studies Of Autism Spectrum Disorders Have Failed To Resolve The Theory Theory Versus Simulation Theory Debate". In: *Review Of Philosophy And Psychology* 3.2, pp. 263–291. DOI: 10.1007/S13164-012-0097-0.

Wilkinson, M. R., L. J. Ball, and R. Cooper (2010). "Arbitrating Between Theory-Theory And Simulation Theory: Evidence From A Think-Aloud Study Of Counterfactual Reasoning". In: *Proceedings Of The Thirty Second Annual Conference Of The Cognitive Science Society*, pp. 1008–1013. URL: http://Csjarchive.Cogsci.Rpi.Edu/Proceedings/2010/Papers/0314/Paper0314.Pdf.

Wimmer, H. and J. Perner (1983). "Beliefs About Beliefs: Representation And Constraining Function Of Wrong Beliefs In Young Children's Understanding Of Deception". In: *Cognition* 13.1, pp. 103–128. DOI: 10.1016/0010-0277(83)90004-5.

Wittgenstein, L. (2001). *Philosophical Investigations: The German Text, With A Revised English Translation*. Wiley. ISBN: 9780631231271. URL: http://www.worldcat.org/title/philosophical-investigations/oclc/496575871.

Zahavi, D. (2010). "Shame and the Exposed Self". In: *Reading Sartre: On Phenomenology and Existentialism*. Ed. by J. Webber. Routledge. ISBN: 9780415550956. URL: http://www.worldcat.org/title/reading-sartre-on-phenomenology-and-existentialism/oclc/551722193.