Defending simulation theory against the argument from error

Timothy L. Short

Department of Philosophy, University College London

Kevin J. Riggs

Department of Psychology, University of Hull

Acknowledgement: We gratefully acknowledge Ian Phillips, and Steve Butterfill for feedback on an earlier draft

Address for correspondence: Timothy L Short, Department of Philosophy, University College London, Gower Street, London WC1E 6BT.

Email: t.short@ucl.ac.uk.

Word count: 7575

Abstract

We defend the Simulation Theory of Mind against a challenge from the Theory Theory of Mind. The challenge is that while Simulation Theory can account for Theory of Mind errors, it cannot account for their systematic nature. There are Theory of Mind errors seen in social psychological research with adults where persons are either overly generous or overly cynical in how rational they expect others to be. There are also Theory of Mind errors observable in developmental data drawn from Maxi-type false belief tests. We provide novel responses to several examples showing that Simulation Theory can answer these challenges.

Saxe (2005) challenges Simulation Theory of Mind (ST) by noting occasional systematic error when persons assess the mental states of others. On her view, these errors are more easily explained by Theory Theory of Mind (TT) than by ST. She allows that ST could explain the *existence* of errors, but denies that it can explain the different *types* of error that reliably occur in different circumstances. Her question is a good one, and we will be taking it seriously in this paper.

We agree that while mindreading abilities have many successes, there are also many failures of prediction. We also agree that these errors in mindreading are systematic in nature, with many participants making the same errors in the same circumstances. For example, adults often exaggerate the extent to which people hold the same beliefs that they do, and 4-year-old children seem to assimilate ignorance to error. Saxe does allow the 'wrong inputs' defence of these errors – the claim that error in the simulation comes about because the inputs to the simulation were wrong. Saxe's challenge though is that the wrong inputs defence cannot account for the *systematic nature* of the errors. The fact that children systematically say that someone who does not know something will have a false belief about it, rather than not know it, is held by Saxe to comport more happily with a TT line. The line is that children use a false theoretical lemma – 'ignorance means you get it wrong'. That line, according to Saxe, is not open to ST – one might think that the errors should be random or at least less one-sided if simulation abilities grounded mindreading capacities.

The lack of any substantial response from the ST side to Saxe's clear and comprehensive paper has partly driven an emerging consensus for a hybrid view involving both ST and TT. For example, Apperly writes that 'many authors now argue for a hybrid account in which both Simulation and Theory play a role' (Apperly, 2008). Apperly supports his view by noting that 'cases where people make systematic errors [...] are seen by many as good evidence' for TT, and gives only two citations, of which Saxe (2005) is one. Similarly, Morin (p. 1069, 2007) writes that 'ST is largely accepted in the literature (but see Saxe, 2005)'. In sum, as Doherty (p. 47, 2008) points out, the ' "argument from error" is one of the most powerful arguments against' ST. The lack of a response to Saxe (2005), then, has driven this emerging hybrid consensus. In this paper, we will supply this response.

Saxe uses evidence from both adults and children to make her case. The adult data comes from decades of research in social psychology, where participants systematically make errors in assessing and evaluating the mental states of others. Our response will be that *bias mismatch* can supply the missing element to ST which can explain the systematic nature of these errors. It is well-known from psychological research that persons exhibit many errors in their reasoning due to cognitive biases, such as Confirmation Bias and the use of the Availability Heuristic. There are several types of Confirmation Bias; for example, one might seek only information tending to confirm a hypothesis. The Availability Heuristic is applied when someone 'evaluates the frequency of classes or the probability of events by availability, i.e., by the ease with which relevant instances come to mind' (Tversky and Kahneman, 1972). We propose that simulation of others does not include simulation of their cognitive biases, which is why there can be systematic errors in mindreading. One reason they do not include these biases in their simulations is that they are less exposed to the emotional impact of the situation than the persons being simulated. Another possibility is that they are using different systems of reasoning. It is widely accepted that there are two systems of reasoning; one is `quick and dirty' while the other is slower and more rational (Sloman, 1996). If the person being simulated and the person doing the simulation are using different systems, there will likely be ToM errors. So in this way also, we respond to Saxe's challenge by employing a variant of the wrong inputs defence.

Saxe's second source of evidence is developmental, where the data suggest that young children do not differentiate between 'not knowing' and 'getting it wrong' (Ruffman, 1996). Our response here is different to the one we will use with the adult data. Rather than propose a problem with modelling bias, we show that the Ruffman data can be explained by a simulationist account coupled with development in processing capacity.

In what follows, we first outline some of the situations cited by Saxe (2005) where there is systematic error in theory of mind and explain how failure to model these cognitive biases explains the errors. Saxe suggests a number of relevant circumstances. In some situations, persons are not cynical enough about the reasoning capacities of others and in other situations we are too cynical. In other words, persons sometimes expect too much in the way of rationality and logic from others and sometimes too little. We will show how bias mismatch driven by emotional or affective mismatch explains the errors made. In the second part of the paper, we will consider Saxe's developmental claim that the way that children confuse ignorance with error is difficult to explain on ST. Throughout the paper, we will use the term 'S' to refer to the subject who is using Theory of Mind abilities to mind-read a target individual 'O', who is the object of Theory of Mind abilities.

1: Theory of Mind: Not Cynical Enough

Saxe (2005) claims that adults believe that they and others are more rational and logical than they are; they are ‘not cynical enough’ when they mindread. One useful citation for Saxe is the notorious Milgram (1963) study. She writes: ‘if we could accurately simulate other minds, half a century of social psychology would lose much of its power to shock and thrill. [...] The experiments of Milgram [...] are famous because there is a specific, and vivid, mismatch between what we confidently expect, and what the [Os] actually do’ (Milgram, 1963, p. 177). Similarly, she claims that ‘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’ (Saxe, 2005, p. 176). This then, is an error of mindreading – systematically over-rating how closely our behaviour, and that of others, reflects optimal reasoning.

Saxe backs this claim by citing Gilovich (1993). Gilovich’s work is primarily about first order errors: errors of logic and reasoning that people make and the reasons they make those errors. Saxe’s question is about errors in mindreading – the second order errors that people make *about* the first order errors that other people make. However, since Gilovich’s whole thrust is that the prevalence of first order errors that people make when they hold ‘questionable’ beliefs is surprising, some discussion of his work is useful here. The basic project of the book is to ask why ‘questionable and erroneous beliefs are learned, and how they are maintained’ (Gilovich, 1993, pp. 9-10). The maintenance of these false beliefs is of interest to us, since one might expect confrontation with evidence to remove them. To the extent that this does not happen, while persons predict that it will, then we have a second order error of mindreading that we want to consider. We will not be able to explain away these second order errors by positing unexpected lack of cognitive ability or exposure to evidence as explanations of the first order errors: as Gilovich goes on to note: ‘[e]rroneous beliefs plague both experienced professionals and less informed laypeople alike’ (p. 10).

We are also interested in the biases that Gilovich cites as explanations of the false beliefs that people hold, because our proposal is that absence of specifically those biases in S *at the time of the simulation and as part of the simulation* is what accounts for the surprise and the errors in mindreading. Naturally we do not claim that S is free of the biases displayed by O; merely that the relevant biases are not triggered in S or used as part of the simulation because S is not actually in O's situation with its affective import. Many first order errors, Gilovich writes, 'can be traced to imperfections in our capacities to process information and draw conclusions' (p. 10). We will argue that these imperfections are not simulated. In what follows we will outline the studies used by Saxe to support her argument, but additionally we also focus on data relevant to second order errors. That is, errors of mindreading.

1.1 Unexpected Willingness to Shock

Milgram's (1963) experiment involved deceiving participants (the Os) into thinking that they were aiding the experimenter in testing how well students learnt word pairs. The Os were told to apply an electric shock to the student if the student made a mistake. The students were in fact confederates of the experimenter and did not receive any shocks. However, participants believed they were administering shocks ranging from 'moderate' through 'intense' to 'danger: severe shock' and beyond to the mysterious 'XXX' category. The various pretended shocks were accompanied by increasing signs of distress from the student confederates. The surprising results were that 26 of the 40 participants obeyed the orders of the experimenter to the end, applying the strongest shock available.

We now come to the evidence relevant to errors in mindreading. Milgram 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 [Os]. [...] 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' (p. 375).

The defence of ST that we propose must now explain this failure to predict actual performance. The answer, we suggest lies in the substantial affective mismatch between S and O. The Ss, whether ourselves or Yale seniors, when considering the question as to how much participants would be prepared to shock are in a relatively calm, reflective state – we/they are able to 'reflect carefully'. The Ss are not this instant under pressure from an authority figure in a white coat, issuing stringent instructions. We may even imagine that the 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.

In addition, note that Ss are given the salient facts and asked to opine on them rationally, which differs from the position of the Os, who simply experience the world without the important, salient or significant facts being given to them as such. Nothing about the calmness and lack of involvement of the Ss is true of the Os. As Milgram writes, many of the Os exhibited extreme affect: ‘the degree of tension reached extremes that are rarely seen in socio-psychological laboratory studies. [...] Fourteen of the 40 [Os] showed definite signs of nervous laughter and smiling [...] Full-blown, uncontrollable seizures were observed for 3 [Os]’ (p. 375).

There is then a very clear affective mismatch between S and O. This, we suggest, explains the absence of the appropriate bias in the simulation which, in turn, explains the systematic failure of mindreading. Milgram even provides this interpretation: ‘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’ (pp. 375-376). Our explanation however must in addition account for a further intriguing element of failure of ST also described. Milgram had hidden Ss, final year psychology undergraduates, watching the experiment and fully aware that the shocks were pretend. Still, these Ss ‘often uttered expressions of disbelief upon seeing an experimental participant administer more powerful shocks to the victim’ even though the Ss ‘had a full acquaintance with the details of the situation’ (p. 377). These hidden Ss were still not subject to the bias towards obeying the instructions of the experimenter to administer the ‘shocks’. They faced much less pressure and affect than the experimental participants, the Os. The affective mismatch between S and O creates a bias mismatch between S and O. This explains the simulation failure by S in relation to O.

The particular bias involved here is the Conformity Bias, also known as the Asch Effect (Asch, 1952). Conformity Bias, put simply, is our tendency to do what we are told. More formally, Prentice (2007, p. 18) gives the following definition: ``conformity bias strongly pushes people to conform their judgments to the judgments of their reference group''. Here, the relevant `reference group' for O is the authoritative figure of the experimenter who is urging the O to give the electric shocks. By so urging, the experimenter suggests to the O that the giving of the shocks is `normal behaviour' and that the acceptability of giving the shocks is the judgment of the reference group. O demonstrates Conformity

Bias in agreeing to apply the shocks. S does not share the affective involvement of O, and therefore omits the Conformity Bias of O in his simulation of O, leading to systematic ToM error.

1.2 Unexpected Belief in Agreement

Gilovich (1993) also includes some evidence of the 'false consensus effect' which Saxe uses to support her position. 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 points out that there is a 'systematic defect in our ability to estimate the beliefs and attitudes of others' whereby persons 'often exaggerate the extent to which other people hold the same beliefs that [they] do' (pp. 112-113). This is evidence for a failure in mindreading because if people simulated accurately, they would not predict the presence of this support where it is absent.

Gilovich (and Saxe) cite data from Ross *et al.* (1977). Students were asked if they would be prepared to wear a large sign around campus bearing the legend 'REPENT'. Many agreed. These students were then asked the critical question as to what percentage of their fellow students they thought would also agree to do so. It transpires that students thought that their peers would decide roughly as they had. Gilovich (1993) reports that '[t]hose who agreed to wear the sign thought that 60% would do so, whereas those who refused thought that only 27% would agree to wear it' (p. 114). This is an error of mindreading of exactly the sort used by Saxe to support her argument. However, we would argue that such data in fact speak against Saxe's position. The error is basically that of thinking that others are more like oneself than they are. What could be stronger evidence for ST? Furthermore, of special interest to us is Gilovich's explanation for the false consensus effect. He claims that a basic desire is to 'maintain a positive assessment of our own judgment' which is particularly likely to play a part when we 'have an emotional investment in the belief' (p. 114). That is, persons wish to believe that they are highly skilled at making judgments of all kinds and they will as a consequence tend to think that have penetrated to the truth of a matter; this truth then will be what others hold as well in their view. This is exactly what one would predict according to our view: since the S has no affective investment in the beliefs of O, S commits a mindreading error by not simulating the resultant false consensus effect and the consequent errors of O.

At this point, it might be objected that the Ross *et al* (1977) data could as well be explained by TT as by ST. The TT proponent could postulate an incorrect theoretical axiom in ToM, to the effect that `others (mostly) believe what I believe'. Our initial response to this is dialectical to the effect that even if true, our explanation of the data has value as a `weak defence'. As far as we know, ST proponents have offered no response at all in print to the charge that ST cannot account for such systematic errors as exemplified in experiments on the false consensus effect and so we feel it is valuable that we have provided one.

Beyond this, however, we can offer a `strong defence' with the import that ST provides not just *an* explanation of the data, but a *better* explanation that TT. One argument for this relies on exposing what we call the `setting the bar too low error'. This error involves making the false assumption that if a ToM exercise can be described in theoretical terms, then it perforce was in fact conducted theoretically. This is like assuming that because I get the same answer every time I simulate my feelings on hearing of the rejection of my paper, I must have a theoretical axiom relating paper rejection and emotional states.

Secondly, we note that ST provides a more parsimonious explanation of the false consensus effect than TT. In fact, ST barely needs to explain the effect at all since it is a natural outcome of simulation that there is `default belief attribution' viz., since S starts from his own state in simulating O, the simulation begins with O held as believing everything S believes. TT may have an axiom to perform default belief attribution, but the simplicity of the axiom would not be paralleled by simplicity of implementation. In fact the implementation of such an axiom would be very complicated. Nichols and Stich (2003, p. 107) have ably pressed this objection against TT. They note that it would ``generate chaos if the model also contained most of [S's] own beliefs, since some of those beliefs will be incompatible with the discrepant beliefs of [O]''. For more on both of these arguments, see Short (forthcoming).

A further objection suggests that that our account over-generalises. The objection starts from the idea that people want to keep beliefs which are central for their self-image. We agree that this is the case and that there are good arguments around self-serving biases for self-deception to this effect (Mele, 2001). The claim is that we do not exercise such biased belief retention all the time but only quite selectively. Thus, the objection runs, our account tends to generalize too much some phenomena which are reported in the literature but are not universal.

Our response is to note that we are merely accounting for literature that the TT camp has cited in support of its position, and we need not solely in virtue thereof be committed to any further claims about other circumstances. We seek only to provide ST with a response to the charge that it cannot explain such examples of systematic error, and by appealing to a bias mismatch, we have done so. Whether the effect is seen frequently or infrequently, it is described in the literature and we explain the ToM errors seen in that literature, no more and no less. S does not simulate the false consensus effect in O and therefore makes a ToM error about O only when the effect operates in O. We explain on a simulationist basis what happens whenever this particular error occurs. It is worth noting here that this objection equally applies to TT. The TT explanation of ToM error is to say that a false axiom has been employed, and that *whenever* such a false axiom is employed, ToM errors will result. TT is more prone to over-generalisation than the ST account.

1.3 Unexpected Belief Perseverance

Saxe cites Stich and Nichols (1995) who describe an experiment by Ross *et al.* (1975) on what we may term the Belief Perseverance Bias. In their paper, Ross *et al.* note that 'once formed, impressions are remarkably perseverant and unresponsive to new input' (p. 880). Os were given a test which suggested that they were unusually good (or bad) at a certain task. Later it was explained to them that the test was bogus. The surprising finding is that Os continued to believe that they were good or bad at the task even when the evidence for that belief had been dismissed. Following on from this study, Stich and Nichols formed a body of Ss from among their students and asked them to predict the results of the study. They found that the Ss predictions were more often wrong than right – students often assumed that people's beliefs would change in the light of the bogus test data. Thus there is evidence for systematic error (or failure) in mindreading.

Stich and Nichols adduce this failure as evidence against ST by noting that the students would have exhibited the belief perseverance effect had they taken the test as opposed to being asked to predict its outcome. Had they been simulating, they would not have made the error, according to Stich and Nichols. This is of course easily explained on the view we are proposing. The S's simulations failed because they failed to include the belief perseverance effect in their simulation. They failed to include that effect because they were not in the situation faced by the Os, who had an affective component resulting from being told something about their competencies which may have been pleasing or displeasing. The experimental task in the Ross *et al.* study was to detect which is real when presented with one real and one fictitious suicide note 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 Os. In addition, the Os were randomly assigned to three groups – success, fail, average – and we can conjecture that membership of at least two of these groups will have had some influence on self-esteem and the associated affective component. This is confirmed by Ross *et al.* who note that 'subjects in the success condition reported having felt more satisfaction than subjects in the average condition' (p. 883) who in turn felt more satisfaction than the subjects assigned to the fail condition.

If the situation of the Ss is made sufficiently similar to that of the Os, then they will exhibit the same bias. This is what Ross *et al.* did when they recruited additional experimental subjects who were engaged in observing and listening to a whole experiment through a one-way mirror. They also exhibited the Belief Perseverance Effect about the ability of the Os. That is, they continued to believe that the Os in the success group were better at the task even after they had learned that the Os did not really succeed. When Stich and Nichols find that their students are 'more often wrong than right', they are polling a group of Ss in a very different affective scenario. The effective aim of Stich and Nichols's student poll may be paraphrased as asking the students whether they would continue to believe a claim about themselves after evidence for it was removed. People would feel quite silly affirming that.

2 Theory of Mind: Too Cynical

Saxe cites experimental data to show that spouses believe that their partners will be *more* self-serving than they are. That is, spouses are 'too cynical' about how self-serving their partners will be. This opposed direction of error from the 'not cynical enough' error discussed above illustrates Saxe's challenge: defenders of ST must explain this directionality of error as well as the possibility of error. Saxe (2005) cites work by Kruger and Gilovich (1995). Each member of a married couple were asked, separately, to rate how often he or she was responsible for common desirable and undesirable events in the marriage. They were also asked to predict how their spouse would assign responsibility. Each predicted that their spouse would be self-serving – taking more responsibility for good events, and less responsibility for bad events. Saxe then concludes that 'whereas reasoning about reasoning is usually characterised by overly optimistic expectations about people's rationality, in specific circumstances [...] Ss are overly pessimistic, an effect dubbed 'naive cynicism' (Saxe, 2005, p. 177).

The data actually reported by Kruger and Gilovich are more complex than this and Saxe is over-simplifying to say that the partners were self-serving. In fact, Kruger and Gilovich canvass an alternate explanation, based on the surprising observation that Os also overstate their own contribution to *negative* events as well as positive ones. The alternate explanation is the well-known Availability Heuristic whereby salient explanations are preferred to other ones not because they are a better fit to the explananda, but simply because they are more available. This factor, when added to the fact that people remember their own activities more easily than those of others, means that people are likely to claim more responsibility for both positive *and* negative events than is warranted. As Kruger and Gilovich point out this should lead to accuracy and error of different types in assessment of the bias of others: Ss 'may be surprised to find that others often claim too much responsibility for [negative] activities as well' (p. 744). This can clearly be explained though on the proposal for which we have been arguing. The Ss failed to model accurately the effects of the Availability Heuristic. The affective mismatch which drives the inaccurate modelling of the Availability Heuristic is more difficult to spot here, though it might be driven by the underlying emotional impact of being asked to comment on one's marriage. That situation may well be emotionally involving and call to mind both difficult times and easier ones.

Kruger and Gilovich also find that undergraduates, asked to predict the results of the above study, exhibit the same effect: i.e., they also expect the married couples to be self-serving in their allocations of responsibility. Importantly, the undergraduates were asked *why* they gave the assessments they did. The explanation they gave were more consistent with a self-serving account than an Availability Heuristic account. For example, '(83%) articulated a theory of motivated bias for at least one of the [positive or negative] activities' (Kruger and Gilovich, 1999, p. 746). Holding a theory of motivated bias means that the undergraduates said they expected people to make statements in the direction that would make those people look good. One interpretation of this is that the undergraduate Ss were not simulating the Availability Heuristic in the Os. The participants in the marriages were subject to one bias – the Availability Heuristic – and this was not modelled adequately by the Ss. It will also be clear that the affective mismatch explanation for this missing bias is available here as it was previously, since commenting on a marriage in which one is a participant has higher emotional impact than commenting on one not involving oneself. Many events within a marriage will take on a higher level of emotional importance than would be predicted by Ss or even perhaps by the Os themselves when they reflect on their behaviour in a calmer later period.

Kruger and Gilovich consider the exact question Saxe raises: when do people make the switch from naive cynicism to naive realism? In other words, what principled account can be given to explain when people make this type of error and when they do not? The authors conclude: 'the naive cynicism we have documented here applies to people's intuitions about the judgments of individuals who are seen as having a vested interest in the matter at hand' (p. 752). So again we have affect mismatch between S and O: O is seen as emotionally committed while S is not.

2.1 Interim Summary of the adult data

Our task in this paper has been to supply ST with a response to the claims that it cannot account for some literature detailing systematic ToM error. We have identified a bias in each case, suggested that it is not simulated leading to a bias mismatch *in these specific circumstances*. Before going on to deal with Saxe's claims based on developmental data, we first address some questions raised by one of the anonymous reviewers.

It might be objected that there can be ToM errors in circumstances where there does not appear to be a great deal of affective difference between S and O. We agree, but can only briefly sketch here a second reason for bias mismatch. [See Short (2015) for much more on this.] The idea is that there can also be system mismatch. Sloman (1996) and others have proposed that there are two systems of reasoning. System 1 is quick and dirty, and may be termed intuitive and the slower and more rational System 2 is reflective. The basic idea of system mismatch is that if S and O are using different systems of reasoning then there will likely be ToM error. For example, if S reflectively and carefully simulates O's reasoning in a scenario where O has employed quick System 1 reasoning, S will likely make a wrong prediction about O's behaviour. By contrast, if S and O both employ the same system of reasoning, then there are good prospects of S avoiding ToM error and making a successful prediction of O's behaviour.

A further objection might be based on the idea that our view denies that S's sometimes account for the relevant informational basis of O's as basis of the understanding of O's. In fact, we do not deny this. Such informational changes to S's belief set to account for the different information held by O will often be essential to successful simulation. The simulationist account of this would be roughly as follows. If S is simulating an O with no relevant informational differences to S, as far as S knows, then S can safely predict O's behaviour on the basis of S's simulated behaviour. Making changes to account for O's different information, where relevant to the particular behaviour in question, is an important part of getting the simulation right. In fact, the objection to TT of Stich and Nichols (2003) which we mentioned earlier bears on this point. Stich and Nichols argue persuasively that specifying a set of axioms under TT to account for belief differences between S and O would be difficult. Under ST, S can start from assigning O the same belief set as S has and then adjust it. ToM questions such as ‘would I enter the coffee shop because I have a desire for coffee even though I believe that I have no money?’ become easily answerable. S is thus enabled to predict a circumstance in which O will not enter the coffee shop, even when O desires coffee, and even when S knows that S has money.

It might be also objected that our account entails that no biases are ever simulated. This would be a serious objection were we committed to it. We are not however. Consider an example from Michel and Newen (2010) about the egocentric bias in belief formation, whereby individuals engage in self-deception to bolster their self-esteem. We agree that if S's never accounted for such biases, ToM would rarely be successful, which is clearly not the case. It also looks implausible just from everyday experience: many of us are familiar with the experience of imagining the reaction to some success of a rather pompous academic colleague who has an unrealistic and inflated sense of his own capabilities. A select few may even have some insight into their own tendencies in this direction.

Beyond this, we also point out bias matching can lead to successful simulation and successful ToM. We can form a picture of this in the egocentric bias case which provided the basis of this objection. Imagine that S knows he is rather prone to self-deception in relation to his abilities in chess. [More plausibly perhaps, imagine S knows of someone else who has this proneness.] This will assist S in simulating the pompous academic O, even if S is himself free of that particular instantiation of egocentric bias. S can simply predict that O will be somewhat overly positive about himself *in relation to that topic* and not necessarily about other topics.

Again, it is important to avoid the ‘setting the bar too low’ error here as always. Neither the TT or the hybrid causes are supported by finding an episode of ToM operation which can be described in theoretical terms, since the fact that a process *can* be described using an axiom does not mean that that axiom or any other *was* employed in conduct of that process.

3 Children's Assimilation of Ignorance to Error

Saxe (2005) also cites experimental data showing a mindreading error in children. We will begin by describing her challenge before using the approach to mindreading called adaptive modelling proposed by Peterson and Riggs (1999) to respond to it.

Saxe reports an experiment by Ruffman (1996) suggesting that four-year-olds do not differentiate 'not knowing' from 'getting it wrong'. In one experiment, a child and a doll (A) are seated in front of two dishes of sweets. The round dish contains red and green sweets, but the square dish contains only yellow sweets. The child and the doll watch while a sweet from the round dish is moved into an opaque bag. Although the child knows that the chosen sweet was green, the doll does not. Children are then asked 'what colour does the doll think the sweet in the bag is?' The correct answer is that the doll does not know or that the doll thinks it is either red or green. Children report that the doll thinks the sweet is red. Ruffman concludes that children found it easier to ascribe a false belief (the sweet is red) than to ascribe a true belief (the sweet is green) and this is contrary to the predictions of ST. Thus on his view, the child simulates using his or her own beliefs as a basis for prediction. If the child believes, correctly, that the sweet in the box is green, then using their own belief state as an input they should find it easier to ascribe true belief to the doll, since the child has that same true belief.

Saxe concludes that 'the actual result is best explained by an inaccurate generalisation in the child's developing theory of mind: 'ignorance means you get it wrong'. Because A is ignorant of which sweet was chosen from the round dish, A must think that it was the wrong colour, a red one' (p. 175). Saxe claims that the data can explained by TT but not by ST since it is an example of the application of a false theoretical lemma 'if X is the case and one does not know whether X is the case, then one believes not-X'. Saxe's general challenge here is why the children fail to model the doll correctly, viz., why they wrongly simulate the doll as being wrong rather than ignorant i.e., with a chance of being fortuitously right.

3.1 Unadapted Modelling Process

Saxe's challenge does have a response though, using the model proposed by Peterson and Riggs (1999). This model was put forward to explain mindreading performance of 3 and 4 year olds in the false belief task. Their approach was to devise models (or databases) of the key events / facts concerning the false belief scenario; one model for the child, and one model simulated by the child of the protagonist.

With regard to the Ruffman task, we can represent the child's model or database as follows;

Query: What colour is the sweet?

Fact: The sweet came from the round bowl (either it is red or it is green)

Fact: The sweet was green

When asked what colour is the sweet the child can consult her model and read off the answer 'green' to the query.

3.2 Adapted Modelling Process

We now set out how the database needs to be modified for the child to simulate the doll's database. Recall that Saxe's challenge here is to explain why children *systematically* assimilate ignorance to error. That is, why the child will usually say that the doll will say that it is red, when a better answer would be to say that the doll does not know.

The key point of the Peterson and Riggs proposal is that mind-reading in a false belief task is accomplished by i) identifying the thing which the protagonist does not know and then ii) implementing this as an 'ignore' instruction in one's own system. The doll is the protagonist. What does the doll not know? It does not know that the sweet is green. This gives us the following simulated database (or adapted model).

Query: What colour is the sweet?

Fact: The sweet was from the round bowl (either it is red or it is green)

Fact: The sweet was green

Ignore (Fact: The sweet was green)

To produce this database the child has had to inhibit the fact that the sweet was green. In other words cognitive control was required to simulate the doll's model. Now we come to the question asked of the child – what colour does the doll think the sweet is? To answer this query the child consults the modified database. However, this does not give a single solution, because the doll only knows only that the sweet came from the round bowl – that it is either red or green. There are now three options – to say that the doll thinks the sweet is red or to say that the doll thinks the sweet is green, or to say that the doll thinks the sweet is either red or green. But here is where a second consequence of the ignore instruction comes into play, and it does so deleteriously from the perspective of mindreading capacities. The child has already suppressed the fact that the sweet is green. Therefore, the only option available is to respond that the sweet was red. Thus we explain why the child wrongly answers that the doll thinks the sweet is red without appealing to the theoretical lemma 'ignorance means you get it wrong'.

Based on this analysis we can then hypothesise that as mindreading abilities develop, the ignore instruction can be imposed selectively. It should be used to simulate the doll's knowledge base. But, if the ignore instruction is also used as a response inhibitor to the question 'what colour does the doll think the sweet is?', then the child makes the error Ruffman reports.

Ruffman also makes a point on inputs which is helpful for our case. It brings out that our defence can be seen as not a wrong inputs defence, but a wrong processing defence. Our claim is that children improve in processing ability, not the ability to process the right inputs. Ruffman claims that the observation that children assimilate error to ignorance is more congenial to TT than ST because 'children were generally good at 'inputting' or taking account of the relevant information (i.e. the [...] knowledge of the sweets' colours'. (p. 388). This exposes a useful ambiguity in what is meant by an 'input'. By 'input', Ruffman means raw data gathered from the senses, such as might be expressed in the proposition 'the sweet is green'. In contrast, by 'input', we understand 'input to the simulation', which of course is what matters for understanding the results of that simulation. Our view allows that simulation takes place by changing the inputs, changing the processing on those inputs, changing the outputs, or some combination of those three factors.

An additional merit of our account is that we can now understand why the developing ability to simulate will handle ascriptions of true belief and false belief before it can handle ascriptions of ignorance. Both TT and ST allow, as they must, for improving Theory of Mind abilities as development progresses. They predict different sorts of development however. For Saxe (2005), development would consist in abandoning the mistaken assimilation of ignorance to error. On our ST view by contrast, the change would be due to an improved ability to take the perspective of another, including the ability to ascribe 'ignorance' to them, driven by changes in cognitive control. Note how difficult this might be. Children can more easily simulate true and false belief because these beliefs derive from their own perspective – they have first-hand experience of their false beliefs. To ascribe ignorance, they need to become aware, in a meta-belief, that there are propositions about which they do not have a true or a false belief – or about which they have no belief. In other words, they are ignorant of some data which would give warrant to asserting or denying some proposition. This is more complex than the assessment of truth or falsity, and will therefore result in a developmental lag.

In conclusion: we have shown that ST can explain systematic Theory of Mind errors in both adults and children, contrary to Saxe's claims.

Timothy L. Short
Department of Philosophy
University College London
Kevin J. Riggs
Department of Psychology
University of Hull

References

Apperly I, 2008: Beyond simulation-theory and theory-theory: why social cognitive neuroscience should use its own concepts to study 'theory of mind'. *Cognition*, 107, 1, 266-283. DOI: 10.1016/j.cognition.2007.07.019

Asch S E, 1952: *Social Psychology.* Upper Saddle River: Prentice-Hall. URL: http://www.worldcat.org/title/social-psychology/oclc/254969

Davies M and T Stone, 1995: *Mental Simulation: Evaluations and Applications - Readings in Mind and Language.* Hoboken: Wiley. URL: http://www.worldcat.org/title/mental-simulation-evaluations-and-applications/oclc/31970746

Doherty, M J, 2008: *Theory Of Mind: How Children Understand Others' Thoughts and Feelings*. Abingdon: Taylor and Francis. URL: http://www.worldcat.org/title/theory-of-mind-how-children-understand-others-thoughts-and-feelings/oclc/195720264.

Gilovich T, 1993: *How We Know What Isn't So*. New York: The Free Press. URL: http://www.worldcat.org/title/how-we-know-what-isnt-so-the-fallibility-of-human-reason-in-everyday-life/oclc/22956975

Gordon R M, 1986: Folk psychology as simulation. *Mind and Language*, 1, 158-171. DOI: 10.1111/j.1468-0017.1986.tb00324.x

Harris P F, 1992: From simulation to folk psychology: the case for development. *Mind and Language*, 7, 120-144. DOI: 10.1111/j.1468-0017.1992.tb00201.x

Kruger J and T Gilovich, 1999: 'Naive cynicism' in everyday theories of responsibility assessment: on biased assumptions of bias. *Journal of Personality and Social Psychology*, 76, 743-753. DOI: 10.1037//0022-3514.76.5.743

Mele A R, 2001: *Self-Deception Unmasked*. Princeton: Princeton University Press. URL: http://www.worldcat.org/title/self-deception-unmasked/oclc/52522503

Milgram, S, 1963: Behavioural study of obedience. *The Journal of Abnormal and Social Psychology*, 67, 4, 371-378. DOI: 10.1037/h0040525

Morin, A, 2007: Self-awareness and the left hemisphere: the dark side of selectively reviewing the literature. *Cortex* 43, 8, 1068-1073. DOI: 10.1016/S0010-9452(08)70704-4

Nichols S and S P Stich, 2003: *Mindreading: an Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Clarendon. URL: http://www.worldcat.org/title/mindreading-an-integrated-account-of-pretence-self-awareness-and-understanding-other-minds/oclc/52485349

Peterson D M and K J Riggs, 1999: Adaptive modelling and mindreading. *Mind and Language,* 14, 80-112. DOI: 10.1111/1468-0017.00104

Prentice R A, 2007: Ethical decision making: more needed than good intentions. *Financial Analysts Journal* 63.6, 17–30. DOI: 10.2469/faj.v63.n6.4923

Ross L, M R Lepper and M Hubbard, 1975: Perseverance in self perception and social perception: biased attributional processes in the debeliefing paradigm. *Journal of Personality and Social Psychology,* 32, 5, 880-892. URL: http://www.ncbi.nlm.nih.gov/pubmed/1185517

Ross L, D Greene and P House, 1977: The 'false consensus effect': an egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology,* 13, 3, 279-301. DOI: 10.1016/0022-1031(77)90049-X

Ruffman T, 1996: Do children understand the mind by means of a simulation or a theory? evidence from their understanding of inference. *Mind and Language*, 11, 388-414. DOI: 10.1111/j.1468-0017.1996.tb00053.x

Saxe R, 2005: Against simulation: the argument from error. *Trends in Cognitive Sciences*, 9, 174-179. DOI: 10.1016/j.tics.2005.01.012

Short T L, 2015: *Simulation Theory, a Psychological and Philosophical Consideration.* Abingdon: Taylor and Francis. URL: http://www.routledge.com/books/details/9781138816053/

Sloman S A, 1996: The empirical case for two systems of reasoning. *Psychological Bulletin* 119, 3–22. DOI: 10.1037//0033-2909.119.1.3

Tversky, A and D Kahneman, 1973: Availability: a heuristic for judging frequency and probability. *Cognitive Psychology,* 5, 2, 207-232. DOI: 10.1016/0010-0285(73)90033-9