Dear Ian,

If we decide to do our ToM paper project together it seems to me useful that we should first agree on an organising question. I think we're quite clear about where we want to end up (there's an interesting distinction between relatively primitive, early developing and more sophisticated, late developing ToM abilities) and also about the relevant data (chimps plus children on a variety of ToM tasks). So what's the question?

Maybe you already have one in mind, but I think it would be useful to mention a couple of possibilities. One question is how to characterise chimp and infant theory of mind abilities. On chimps, this is a pressing question because they appear to have only part of a theory of mind: Tomasello characterises chimps as able to think about perceptions and goals but not beliefs. It's as if chimps have what adult humans have minus one of the attitudes (belief). As a characterisation of chimps' understanding this seems inadequate both because they appear to have some ability to reason about knowledge and also because the attitudes can only be understood as forming an interlocking whole, which means chimps' understanding of perception and goals must differ from adults'. On infants, this is a pressing question because they appear to manifest some awareness of belief on a variety of tasks—violation of expectations, word-learning and deception-related behaviours—while systematically failing a wide variety of standard tests for false belief understanding.

Although organising our paper around the question of how to characterise chimp and infant theory of mind abilities would make it quite straightforward to write, I don't find this approach entirely satisfying because I think the question of how to characterise adult theory of mind is just as pressing and should be asked at the same time. Of

course whatever we do would address the question of how to characterise chimp & infant theory of mind abilities, but I'm not sure it's the best organising question.

My earlier false belief paper is organised around the question, What is awareness of belief? We could take a similar approach and ask, What is it to have a Theory of Mind? I still think this is a good question and hopefully one that our paper will bear on, but I'm not sure it would be ideal as our organising question because it is too open ended.

A third approach, which I think I prefer although you will have a clearer idea of whether this is workable, would be to start by arguing that just as discrepancies on violation-of-expectation vs. reaching tasks creates a 'Paradox of Early Permanence' in the case of object permanence, so also the Garnham/Onishi/Southgate findings leave us with a 'Paradox of Early Metarepresentation'. (I'm not sure this is the right label—it's not obviously a paradox and may well not involve metarepresentation). The organising question is then how to resolve the 'Paradox'. This approach has the advantage of getting us straight in to the interesting data. I also think the ToM case is if anything stronger than the Object Permanence case because in the case of ToM we have not only violation of expectation experiments, which are hard to interpret, but also the word learning, deception and chimp paradigms; plus (if we're being really ambitious) the many Leipzig findings on infants' understanding of ignorance and engagement.

An alternative way to pose the same question would be by comparison with Carey on number instead of by comparison with Early Permanence. Carey's question is why it takes infants so long to get from understanding 3 to understanding 4 (or 4 to 5, must check these numbers); the answer is that very different kinds of competence are

involved in what looks, behaviourally, like an incremental step. Similarly, we could ask why it takes children so long to get from the rich ToM competence they have at around two years or before to passing false belief tasks at four. (I think Tomasello & colleagues ask this question somewhere.)

Suppose, just for the sake of argument, that we take as our organising question how to explain the 'Paradox of Early Metarepresentation' (or whatever it ends up being called). Having summarised findings which support the existence of a paradox we could then play off two diametrically opposed attempts to resolve it against each other. First we have the attempt to resolve the 'Paradox' using executive function or inhibition. Since everyone agrees that executive function is somehow related to ToM at four years but no one as far as I know has anything compelling to say about how, it strikes me as a good project in its own right to think more about this. But for the sake of our paper, it shouldn't be too hard to make the case that while EF/inhibition is surely relevant, it is not currently known to give rise to any straightforward solution to the 'Paradox'. In particular, we can point to the problem you raised of accounting for children's failure on backwards prediction paradigms requiring verbal responses. I also doubt anyone finds it plausible that chimps fail Call & Tomasello's nonverbal false belief task because of EF/inhibition deficits. Turning to the opposing attempt to resolve the 'Paradox', we might next consider Ruffman & Perner's response to Onishi & Baillargeon. This is partly an attempt to resolve the 'Paradox' by debunking the evidence for early ToM competence. One of Ruffman & Perner's ideas is that infants' performance can be explained by supposing that they keep track of where each person last saw each object they encountered. I think this is an interesting idea that we can build on, but by itself it doesn't adequately explain the 'Paradox' because it can't explain failure on Sabbagh et.

al.'s false sign tasks (the sign points to Chester, so infants only need the ability to keep track of <sign,object,location> relations to solve this puzzle). Similarly, the ability to track where someone last saw an object ought also to be sufficient for solving many standard false belief tasks that children fail until around four years. So while it might be true that infants' success on some false belief tasks can be explained by postulating an ability to track relations between subjects, objects and locations, this idea does not explain why they pass those tasks but fail false sign tasks and standard false belief tasks.

I think this will be fairly straightforward. The hard bit is to characterise our positive view and derive some interesting predictions. How should we distinguish two kinds of ToM competence? I think the distinction needs to be considered from various angles …

*Content* We should consider whether adults are aware of beliefs as relations between subjects and propositions, whereas potentially infants are aware of beliefs as relations between subjects and <object, location> pairs. Similarly for desires. Ruffman & Perner suggest this, but I think it is less of a debunking move than they realise because everything depends on how infants understand the relation between subjects and the <object, location> pairs. They might understand this relation as having rich psychological significance. (This seems to be the moral of e.g. Henrike Moll's experiments about how infants use 'last object encountered by that person' flexibly in interpreting adults' excitement.) It also seems to me an interesting question (although probably not one for this project) what other types of content-bearing vehicle we can understand. For example, can we keep track of relations between subjects and some form of abstract map-like entity where the map describes their knowledge of a region of space and the

relations between objects in it? Could adults, children or chimps pass a false map task?

*Process* Most important of all. Your idea that one type of ToM competence might be automatic whereas the other isn't fits here, and this is a good place to push your insight about the importance of understanding the processes involved in ToM. (It's also where the notion of modularity might come in, but much as I would like to invoke it I doubt it fits here both because there is a lack of evidence and also because chimps' and infants' ToM capacities seem highly flexible.)

*Attitude* In my earlier false belief paper I suggested that part of the explanation for the 'Paradox' might involve the fact that infants and adults have different ways of understanding what it is for a representation to be correct. This is closely related to the *Content* issue: just as we might distinguish kinds of competence partly by reference to what is represented—a relation to a proposition or a relation to an <object, location> pair—so we should also distinguish kinds of competence by reference to how the relation is understood. Schematically: to have a belief is to bear relation R to X. The Content question is, What is X (e.g. a proposition, an object, a map, or …). The Attitude question is, What is R?

*Neural specificity* Johnson & Morton proposed that we have two distinct mechanisms for face recognition which have different neurological bases; as far as I know, they continue to be plagued by difficulties in localising the sub-cortical face processing mechanism. Given these problems, no one could reasonably demand that our proposal to have clear neurological implications. But if it's even possible that that there is more than one kind of ToM capacity, claims about the neurological bases of ToM competence need to take this into account.

*Acquisition* How are the different kinds of ToM competence acquired? Is one innate? Can we borrow any illuminating insights on relations between language and ToM in development?

*Behaviour* What behaviours do the two different ToM capacities support? Crudely, primitive ToM supports eye movements and word learning whereas reflective ToM supports verbal judgements. It's not likely to be this simple, but even if it was, we have to explain *why* the two capacities support different behaviours. Furthermore, there's a hard question about whether we think primitive ToM supports purposive action (as we might take the evidence on deception-related behaviours to suggest, e.g. Harris & Pollack on false denials.)

*Interaction in development* How is the early-appearing ToM capacity involved in acquiring the later one? We already discovered that we agree in thinking that the two kinds of ToM capacity are two independent mechanisms: it's not that the early-appearing capacity morphs into the later appearing-capacity.

*Online interaction* When a subject performs a ToM task, how do the two kinds of ToM capacity interact. Garnham's work contrasting eye movements with verbal responses suggests that the interaction is indirect: on a single task, subjects succeed by one measure and fail by another.

I realise this is way too much to cover in one paper. We should focus on getting the basic idea out rather than trying to exhaustively cover the distinction we're after from every angle. The most important angles seem to me to be process and content, but I'm open to considering whichever seem most likely to explain the 'Paradox'. We should also have

something to say about the relation between these two: if we propose (say) an automatic, low-level mechanism which tracks relations between agents and <object, location> pairs and a reflective, slow mechanism which tracks relations between agents and propositions, then we also need to say something about why the low-level mechanism can't handle relations to propositions.

Have I written enough to put you off the idea of writing a paper together? I'm very keen to work on a joint paper and not wedded to any of the details mentioned here, but I can easily imagine that you might prefer to work independently. If you do think a joint paper is a good idea, let's talk about what the question should be and get some of the basics written. Either way I hope we'll also get chance to talk about this when I come to visit you and Dana to talk about the Coltheart stuff.

Best wishes,
Steve