Rita Astuti

Some after dinner thoughts on theory of mind

Article (Published version)

SOME AFTER DINNER THOUGHTS ON THEORY OF MIND

RITA ASTUTI

[*Editor's note: This is, literally, the transcript of an after dinner talk given by Rita Astuti on the occasion of a conference held at Stanford University in September 2011. The conference, organised by Tanya Luhrmann, was entitled "Towards an anthropological Theory of Mind". For readers unfamiliar with the relevant debates in psychology and philosophy, it will help to know - at least roughly - what "Theory of Mind" refers to. To put it simply: "Theory of Mind" is the human ability to attribute mental states to oneself and to other people (this is one reason it has been called a* theory*: because it is about phenomena that are not directly observable). We use Theory of Mind to attribute knowledge and ignorance, emotions and thoughts, intentions and desire to others, and to predict and explain their behaviour (this is the second reason it has been called a* theory: *because it is used to make predictions). The awareness that other people may have false beliefs about the world has been regarded as the ultimate proof that one has Theory of Mind. Whether this awareness is something we are born with or something that emerges in the course of cognitive development is a hotly debated issue. In the literature, Theory of Mind is often abbreviated as ToM.*]

The only other time I've been at a conference with a scheduled after dinner talk was a few years ago at Trinity College, Cambridge. The conference was on belief, and it had been funded by the Perrott-Warrick Fund, a Fund set up in the 1930s to scientifically prove the existence of the paranormal and the afterlife.

Although this unusual source of funding did not appear to affect the proceedings during the day, by the time darkness fell magic was allowed to take over. At dinner, we were treated to a magic show, delivered by a professional magician who was also a professionally trained psychologist. He first enchanted us with his tricks, and then revealed the psychological principles underlying them – basically, he made us aware of how he had manipulated our minds to make the magic work.

For example, he performed a simple – and yet so seductive – disappearing trick, and we all fell for it. Then he explained that, apart from the undeniable dexterity of his fingers, the magic worked because of our – the audience's – mind reading abilities.

He counted on the fact that if he intently looked OVER HERE, at his left hand, the audience would FOLLOW HIS GAZE and concentrate on what he was looking at. And while we were focusing on his left hand, the real trick was happening OVER THERE, in his right hand – but nobody noticed.

Of course, the reason I'm telling you this story – which I chose to do before realizing that Graham Jones was going to be here[1] – is that gaze following is one of the basic components of our mind reading abilities: a building block of Theory of Mind, the topic of *this* conference.

You follow my gaze because you want to know what I'm seeing (you know that looking is seeing) and what state of knowledge or ignorance I'm in (you know that seeing is knowing).

By following each other's gaze, we can coordinate our focus of attention (we can both see and know the same thing) and monitor whether we are paying attention to each other – for example, monitoring where you are looking at is a pretty good way of finding out whether you are following what I'm saying or whether you are bored or distracted; although of course, like the magician, you can easily deceive me by pretending to focus your attention on me while thinking about something else…

In saying all of this, I might be taking a big gamble – I'm assuming that you have a mind! This assumption, and what follows from it – that you have knowledge, desires, intentions, emotions, beliefs and that it is your knowledge, desires, intentions, emotions and beliefs that explain your actions or lack thereof – is what Theory of Mind is all about.

In essence, having a Theory of Mind amounts to having a non-behaviourist way of looking at the world.

When you see someone running, you don't just see a physical body in acceleration – you see the intention or the desire to catch the bus or win a medal; when you see a hand reaching for an object, you don't just see a trajectory through space – you see the goal of getting that object; and so on.

Having a Theory of Mind, in other words, means having the capacity to go beyond the surface, beyond the behaviour and the actions to the intentions, the desires, the beliefs that motivate them. From this "deeper" perspective, the world is not just made up of arms, legs and eyes that move in a coordinated fashion; the world is also made up of a host of mental states – your own and those of others – that direct and animate what those arms, legs and eyes do.

Now, as I understand it, the broad aim of this conference is to find ways of establishing whether there is any foundation to what Tanya Luhrmann has referred to as the "anthropological suspicion" that there are significant cross-cultural variations in ToM.

Before we set out to work on this task, bringing together our expertise as anthropologists, linguists, and psychologists, I want to explore what I take to be two very crucial evidential questions.

The first one is: what can count as evidence that others have Theory of Mind – that they understand other people in terms of their mental states?

And the second one is: what can count as evidence that there are cross-cultural variations in Theory of Mind?

The two questions are obviously linked, but let me take them in turn.

The first question raises at least two tricky evidential problems:

One is that agents can behave *as if* they have a ToM, while in fact all they need is an understanding of the regularity of certain patterned behaviours. For example, the fact that I can predict that someone who is stretching her arm towards an object has the goal of retrieving it, can be explained by the fact that I have seen many times that people who stretch their arms towards an object, complete their action by grasping and retrieving that object; I can predict what the person will do next and I might even help them complete the sequence, without imputing any goals at all.

This has been, and continues to be, a thorny issue in the quest to determine whether chimpanzees have a ToM – as in the famous 1987 Premack and Woodruff paper which launched the research on ToM.

Those who are sceptical, most notably Povinelli and his colleagues (Povinelli & Vonk 2003; Povinelli 2004), argue that all the attempts that have been made to show that chimpanzees attribute mental states to others – either humans or conspecifics – can be interpreted more parsimoniously in terms of their sophisticated understanding of the rules that govern behaviour. On this view, chimps are smart but are not mentalizers (and indeed, Povinelli accuses those who say that chimpanzees have a ToM of anthropocentricism – we are so dependent on our ToM to understand action and behaviour that we can't help but see it in the actions of creatures that don't have it).

At the methodological level, this scepticism explains why experiments that aim to establish the existence of ToM, whether in human infants or chimpanzees, try to come up with entirely novel situations whose solution could not have been predicted from experience, but which depend on novel inferences that can only be made on the basis of the attribution of mental states.

The second methodological problem in assessing the existence of ToM competence has to do with whether the various diagnostic tasks designed to reveal it really require the reading of the mind of the other. As pointed out when Premack and Woodruff claimed that their chimp Sarah had a ToM because she seemed able to predict what a human actor would do next in order to achieve his goal, all that Sarah needed in order to succeed was 1) to be very smart, which she clearly was, and 2) to imagine what *she* would do if she were in that sequence of events. Again, she could have solved the task without ever imputing a goal to the mind of another actor.

This problem, of course, is at the origin of the false belief task (Dennett 1987; Winner & Perner 1983), which has come to dominate the research in ToM. As you all know, the task is designed to find out whether a child is able to predict the behaviour of another person based on what that other person believes, which happens to be different from what the child believes and knows to be the case. Thus, to succeed in this task, you truly have to enter the mind of the other and understand that other people can have mental states different from your own *and* that they can have false beliefs, i.e. that they can fail to see how the world really is. The point of the task is that it requires what Paul Harris (n.d.) calls a "self-abnegating prediction."

I'm sure that in the next couple of days we will have many opportunities to come back to the false belief task and to the exciting new findings that have emerged in the past few years in relation to it (e.g., Baillargeon et al. 2010; Onishi & Baillargeon 2005; Southgate et al. 2007; Surian et al. 2007) but for the moment all I want to stress is the great care taken by those that attribute ToM to children or chimpanzees – the evidential tests are extremely stringent, as researchers are well aware that it is far too easy to over-interpret the evidence and over-attribute mind-reading abilities to others.

What then of the evidence used to assess whether there are cross-cultural variations in ToM?

Angeline Lillard, herself a developmental psychologist, has argued that there are significant variations by making extensive use of the anthropological literature on the way people talk about minds, persons, emotions, and so on (Lillard 1998). Here's one notable response she got from Scholl and Leslie – supporters of the view that ToM is grounded in a cognitive module that develops along universal lines:

> The cross-cultural differences catalogued by Lillard explicitly include differences in religious beliefs, and beliefs in phenomena such as witchcraft, magic and karma. As such, her view of cross-cultural ToM differences pertains only to *the inessential fluorescence of mature ToM competence*, rather than to *its essential character in early acquisition*… in general, Lillard seems to be looking at differences in *specific beliefs*, rather than at *the concept of belief*… even specific beliefs *about* the concept of belief are not necessarily relevant: the concept of belief could be universally grounded in a module [as Scholl & Leslie argue] even though most cultures do not recognize the "modular" account in their own folk psychology! (Scholl & Leslie 1999: 137)

Think of the following analogy. People might have all sorts of different theories about the way their vision works – e.g. they might have an emission theory of vision, as apparently 50% of US college students have. But whatever theory they have, it is unlikely to make a difference to the way their retina works, to what they see and what they don't. By analogy, this is what School and Leslie argue is the case with Theory of Mind: whatever explicit folk theory about the mind people have, it is going to make no difference to their mind-reading abilities.

I hasten to say that I don't find this position entirely plausible, at least in its extreme form, but I think that it is useful to keep Scholl and Leslie's criticism in mind when working towards the aim of this conference. Like those who work with young infants and chimpanzees, we need to be very careful in considering (and hopefully agreeing) on *the kind of evidence that would count as evidence* for the fact that, to paraphrase Tanya Luhrmann once again, cross-cultural variations in the ways people imagine the mind have consequences for mental experience.

To use one of Lillard's examples, does the fact that people explain the behaviours of others by invoking witchcraft or astrology count as evidence that they have a different Theory of Mind? Or does the fact that people have a concept of mind that diminishes what happens inside individual minds and emphasizes instead what happens between minds count as evidence that they have a different Theory of Mind?

To begin to approach these kinds of questions, we might need to start by acknowledging that, in fact, Lillard, on the one hand, and Scholl and Leslie on the other, are really talking about different things, or rather, about different levels of the same thing – one is the level of

conscious reflection about the mind, what we might call *Explicit Theories* of Theory of Mind, and the other is the level of mindreading that happens largely outside conscious reflection and probably conscious control. Ethnographic methods are of course well suited to record the former, while experimental methods are best suited to tap into the latter.
Questions about the way these two levels may be linked to each other (and it seems, prima facie, sensible to expect that they must be linked somehow) and questions about how they might influence each other (again, it seems sensible to expect that they must do somehow) can only be asked if we recognize that what people say about the mind, as a result of personal and culturally mediated reflection, is not the same kind of phenomenon as what people do when, largely automatically and unconsciously, they follow somebody's gaze, they infer intentions or beliefs to predict what some one will do next, and so on.

To illustrate this point, let me tell you about an experiment, which is part of a series of studies, designed at the Department of Cognitive Science at the Central European University in Budapest, which have come to be known as the SMURF studies (Kovács et al. 2010)

The experiment asks participants, recruited in Trieste, Italy, to watch one of four different versions of a movie which involves a ball that first rolls behind a screen and then either stays there or rolls away out of sight. Participants are told that at the end of the movie, the screen will be removed and that if they see the ball after this happens, they have to press a button as fast as they can.

This sounds like a daft reaction task, were it not for the surreptitious role played by a Smurf, in all its blue glory! The Smurf is just a bystander who comes on the scene at the start of each movie. From where it stands, on the side of the screen, the Smurf has the same visual access to the ball as the participants who are watching and reacting to the film. Except that, in all four versions of the movie, but at different moments in the sequence of events, the Smurf leaves and then re-enters the scene.

Now, the point of the experiment is to manipulate the expectation that you – the participant – have formed about the location of the ball and the expectation of the Smurf who, because of its exits from the scene, does not always share the same expectation as you.

Thus, when the screen is removed and, in half of the trials, the ball is there to be detected – only these are test trials, since reaction times can only be measured when the ball is behind the screen – you and the Smurf can both be "right" in your expectation that the ball is behind the screen, or you can both be "wrong" (neither of you expecting the ball to be there), or one of you can be "right" and the other one "wrong" (where "right" and "wrong" are measured against the test trials' actual outcome).

So, this is how it works:

FILM 1

The Smurf enters

The ball rolls behind the screen

The Smurf leaves – nothing happens

The Smurf returns

In this case, both you and the Smurf rightly expect that the ball is behind the screen;

FILM 2

The Smurf enters

The ball rolls behind the screen

The ball rolls away

The Smurf leaves – nothing happens

The Smurf returns

In this case, both you and the Smurf wrongly expect that the ball is not behind the screen;

FILM 3

The Smurf enters

The ball rolls behind the screen

The Smurf leaves – balls rolls away

The Smurf returns

In this case, the Smurf rightly expects that the ball is behind the screen, while you wrongly expect that it is not;

FILM 4

The Smurf enters

The ball rolls behind the screen

The ball rolls away

The Smurf leaves – balls returns

The Smurf returns

In this case, you rightly expect that the ball is behind the screen, while the Smurf wrongly expects that it is not.

Because of these manipulations, the question that this cleverly designed study allows us to ask is whether the expectation of the Smurf makes any difference to the reaction time of the participants.

Note that the participants were never told anything about the Smurf, they were not instructed to pay attention to its presence, its movements or its expectations about the presence of the ball, and indeed the Smurf's expectations are totally irrelevant to the participants' perceptual task of detecting the presence of the tennis ball.

And yet, what the reaction times show is that, irrelevant as it may be, participants automatically and unconsciously represented to themselves and were affected by where *the Smurf* expected the ball to be.

Compared to the baseline reaction time – when both the participants and the Smurf had the same incorrect expectation that the ball was *not* behind the screen and were thus at their slowest in detecting the ball – the results indicate that:

- Unsurprisingly, participants were faster than in the baseline when both they and the Smurf correctly expected the ball to be behind the screen;
- That, reassuringly, participants were faster than in the baseline when they correctly expected the ball to be behind the screen and the Smurf didn't – the Smurf's incorrect expectation did not affect them;
- And that, somewhat astonishingly, participants were faster when*they* did not expect the ball to be behind the screen but the Smurf did!

In other words, participants computed the Smurf's expectation – that the ball is behind the screen – and this influenced their behaviour even though it was inconsistent with their own expectation that the ball was not there. Indeed, their reaction times were not significantly different in the two conditions when the expectation that the ball was behind the screen was their own or that of the Smurf.[2]

To me, this study provides a perfect illustration of the kind of automatic, unconscious, low-level processes that make up our mind-reading abilities. Here we have participants whose very simple task is to press a button when they detect a ball, and what we find is that even when performing such a simple task, they cannot help entering somebody else's mind and becoming entangled in what somebody else knows and expects about the world.

Now, let me put this Smurf study side by side with the fascinating accounts by Joel Robbins, Bambi Schieffelin and the other contributors to the special issue of *Anthropological Quarterly* on Mind Opacity in the Pacific Region (Rumsey & Robbins 2008). They report that their informants insist that they cannot know what is inside other people's minds, and recoil at the idea that one might dare read other people's intentions.
Is such a "doctrine" – as Robbins refers to it – just *inessential fluorescence*as Scholl and Leslie would have it?

Well, I don't think that we actually know the answer to this kind of question yet, but the way forward, in my view, would be to take something like the Smurf study to people who insist that they cannot know the content of the minds of others – and who teach their children not to discuss other people's intentions and not to interpret their behaviours with reference to their elusive and unknowable mental states – in order to see whether their doctrine and the social practices that are shaped by it make a difference to their reaction times.

I'm prepared to believe that they might, although my uneducated guess – not having done fieldwork with people who subscribe to this doctrine – is that probably, at this level of fast, automatic, unconscious interpenetration of minds, we would find no difference.

If this were to be the case, we might proceed by using the same sort of task, but by first priming participants to think about their opacity doctrine and see whether this makes a difference, such that their doctrine, when explicitly activated, interferes with the computation of the content of somebody else's mind. One could, for example, manipulate the identity of the bystander, making it into a more or less powerful figure, or changing its religious affiliation; or perhaps one could get people to experience a typical situation in which they are reminded of their "opacity" doctrine before they take the task.

And if this doesn't work, one might design a number of tasks that make mind reading gradually more explicit to see at what point – and at what age – people refuse to enter the mind of another agent.

Consider that back in Trieste, using a simplified version of the study, the Smurf mind-reading effect was found among 7 month old infants.[3] We might thus predict that younger children in the Pacific region might similarly be inclined to automatically compute other people's beliefs and expectations – just like infants in Trieste – but that, as they grow older, they might gradually learn to abide by the (culturally specific) "opacity of mind" doctrine, at least in some contexts.

These are just a few examples of what I see as the task ahead.

The fundamental task is to explore in great detail, through stringent evidential procedures that take time and a great deal of care, the points of articulation between people's explicit theories about the mind and the mental processes that, largely automatically and unconsciously, take them beyond the surface, well inside the minds of others, and, who knows, might make them all susceptible to the magician's tricks.

References

Baillargeon, R., Scott R.S., & He, Z (2010) False-belief understanding in infants. *Trends in Cognitive Sciences*, 14, 3: 110-18

Dennett, D. (1978) Beliefs about Beliefs (commentary on Premack, et al.).*Behavioral and Brain Sciences* 1 (1978): 568-70.

Harris, P. L. (n.d.) One or two theories of mind. Lecture notes, Harvard Graduate School of Education

Jones, G. (2011). *Trade of the Tricks: Inside the Magician's Craft*. Berkeley: University of California Press.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*,*330*(6012), 1830 -1834

Lillard, A. (1998). Ethnopsychologies: cultural variations in theories of mind.*Psychological Bulletin*, *123*(1), 3-32.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255.

Povinelli, D. (2004). Behind the ape's appearance: escaping anthropocentricism in the study of other minds. *Daedalus*, *133*(1), 29-41.

Povinelli, D., & Vonk, J. (2003). Chimpanzee minds: suspiciously human?*Trends in Cognitive Sciences*, *7*(4), 157-160.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain sciences*, *1*(4), 515–526.

Rumsey, A. & Robbins, J. (2008) Social Thought and Commentary Section: Anthropology and the Opacity of Other Minds, *Anthropological Quarterly*, *81*(2), 407-494

Scholl, B. J., & Leslie, A. M. (1999). Modularity, Development and "Theory of Mind." *Mind and Language*, *14*(1), 131-153.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of Beliefs by 13-Month-Old Infants. *Psychological Science (Wiley-Blackwell)*, *18*(7), 580-586

Wimmer Josef, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, *13*(1), 103–128.

1. In his book *Trade of the tricks (2011),* Graham Jones discusses the crucial role of what he calls a “working theory of mind” in the practice of magicians in contemporary France. ↩

2. And the effect did not go away even when, in a follow up study, the Smurf did not reappear on the scene to witness the lowering of the screen, suggesting that the representation of the Smurf’s expectation endures in the mind of the participants, even in the Smurf’s absence. By contrast, the effect goes away when participants witnessed the exact same ball movements, but in the presence of a stack of blocks, instead of a Smurf. ↩

3. The study measured the infants’ surprise (as measured by their looking times) at the fact that the ball was *not* behind the screen when they expected it to be there. As with adults, infants also computed the expectation of the Smurf, looking longer when they themselves did not expect the ball to be behind the screen but the Smurf did. ↩