Theory of mind abilities are widespread.

From early in their second year children point to inform others in ways that reflect the others' knowledge or ignorance; and they can predict actions based on false beliefs about the locations of objects;

scrub-jays selectively re-cache their food in ways that deprive competitors of knowledge of its location;

and chimpanzees select routes to approach food which conceal them from a competitor's view and retrieve food using strategies that optimise their return given what a dominant competitor has seen.

Note here that I am talking about theory of mind abilities only. It is essential for my purposes to distinguish theory of mind *abilities* from theory of mind *cognition*. **Theory of mind *abilities* are abilities that exist in part because exercising them brings benefits obtaining which depend on exploiting or influencing facts about others' mental states.** Theory of mind *cognition* is cognition of mental states; **it paradigmatically involves ascribing propositional attitudes such as beliefs, desires and intentions in order to give rationalising causal explanations of thought and action.**

This distinction is important because the facts about other minds which theory of mind abilities exploit are not necessarily the facts which are represented in theory of mind

cognition. To illustrate, it may be possible to exploit facts about what others perceive by tracking their lines of sight and so without representing perceptions as such.

In fact, in some cases theory of mind abilities need not involve any theory of mind cognition at all. For example, some birds may benefit from preening others because doing so causes others to like them and thereby strengthens social bonds (Clayton and Emery 2007). In preening they are exercising a theory of mind *ability*. It doesn't follow, of course, that preening involves theory of mind *cognition*. One might be driven to preen others without understanding that preening is worthwhile because it influences others' attitudes. For example, you might preen because you realise doing so causes others to sing and to preen you back.

I want to suggest that these findings about theory of mind abilities give rise to a puzzle. The puzzle is about how to explain the theory of mind abilities manifested by young

children, chimpanzees and scrub jays. **What do these agents represent that enables them, within limits, to track others' beliefs and other propositional attitudes?**

The most straightforward answer would be to suppose that they represent perceptions, knowledge states, beliefs and other propositional attitudes.

But I don't think this explanation can be right because.[1] this kind of *theory of mind cognition is hard (in two senses).* A body of evidence

---

1 *Lack of systematicity*. Eg two-year-olds generally fail to take into account what their addressee's can see when choosing referring expressions (but three- and four-year-olds choice of referring expressions do take into account what the addressee can see) {Matthews, 2006 #1781}. I wouldn't want to put too much weight on this consideration because there are well-known cases where adults fail to take into account what their addressees can see {e.g \Keysar, 2003 #1782}.

with humans suggests that reasoning about beliefs and other propositional attitudes requires *conceptual sophistication*,

> for it has a protracted developmental course stretching over several years, and
>
> its acquisition is tied to the development of **executive function and language**—things which two-year-olds, scrub jays and chimpanzees are deficient in.
>
> Development of reasoning about beliefs in humans may also be **facilitated by explicit training and environmental influences**, such as siblings.

Ascribing propositional attitudes also appears to be ***cognitively demanding***,

requiring **attention** and **working memory** in fully competent adults.

It makes sense that propositonal attitudes should be conceptually and cognitively demanding. After all, these are states which form **complex causal structures, have arbitrarily nest-able contents, and are individuated by their causal and normative roles in explaining thoughts and actions.** If anything should take years to acquire and consume scarce cognitive resources it is surely states with that combination of properties.

It is sometimes supposed that theory of mind cognition is only hard when false beliefs are

involved. To illustrate, in characterising chimpanzees' less than full-blown theory of mind cognition, Call and Tomasello write:

> "chimpanzees understand … intentions … perception and knowledge … Moreover, they understand how these psychological states work together to produce intentional action" {Call, 2008 #1553@191}.

It is not obvious that this is true. Take knowledge. According to Quassim Cassam,

> "we understand what knowledge is by understanding how we get it or how it comes to be." {Cassam, 2008 #1706}

> "our [typical adult humans'] fundamental conception of what it is to know that P is

> itself an explanatory conception […] we think of S's knowledge that P as something that can properly be explained by reference to what S has perceived or remembered or proved or ..." (2007: 356).

This is an empirical claim. According to Cassam, to understand knowledge in the sense that human adults do involves being able to understand and construct explanations of how individuals know particular facts by reference to what they have perceived (for instance, Tim knows that Luisa scored because he saw her do it).

It is striking that Cassam offers no evidence for this claim. But in some studies Liz Robinson and I did recently, we did find some evidence for Cassam's claim.

There are also further claims about what it is to have a conception of knowledge which philosophers have advanced on *a priori* or introspective grounds. If any of these claims turn out to be correct, they would mark significant discontinuities between human adults' conceptions of knowledge and chimpanzee theory of mind cognition.

This is a problem for Call and Tomasello. It is a mistake to start with concepts like knowledge when our aim is to characterise what chimpanzees—or young children—understand about minds. These concepts are extremely sophisticated and relatively little is known concerning what any groups of subjects, including adult humans, understand of these conceptions.

Of course Tomasello and Call's claim may be useful as a rough, metaphorical guide to chimpanzee thinking {for conflicting views on this, see \Povinelli, 2000 #1391}{Penn, 2010 forthcoming #1785}{Tomasello, 2005 #1786}. But, as they themselves might agree, it does not go very far towards explaining how minds and actions appear to a chimpanzee.

So the puzzle is this:

> Abilities to solve tasks which hinge on facts about what others see, know and believe are widespread
>
> The most straightforward way to explain these abilities would be to suppose that

the agents in question are representing perceptions, knowledge states, beliefs and other propositional attitudes as such.

But abilities to reason about propositional attitudes appear to be require cognitive resources and conceptual sophistication. This may justify caution in supposing that even human adults' theory of mind abilities always depend on representing perceptions, knowledge states and beliefs as such.

The existence of the puzzle shows one respect in which **relatively little is known about what it is to have a theory of mind**. You shouldn't be surprised that I think there are lots more---as a philosopher, not knowing

things is what I do for a living. But I want to focus on this aspect of the puzzle.

So what do infants, chimpanzees and scrub-jays represent that enables them, within limits, to track others' perceptions, beliefs and other propositional attitudes?

To understand how minds appear to infants and other animals, we need to construct notions which resemble propositional attitudes like knowledge states, beliefs, desires and intentions in some respects but representing which is less conceptually and cognitively demanding.

The approach I am proposing is not entirely new …

§ predecessors

On this question there have already been several intuitve suggestions.[2] But on the whole these suggestions imply signature limits which infant theory of mind abilities have since been shown to exceed. In particular, none of the existing suggestions

---

[2] Juan-Carlos Gomez has suggested awareness of primitive intentional relations to objects established by gaze may be important (Gomez 2007: 730), Daniela O'Neil and Martin Doherty have discussed a notion of engagement with objects (O'Neill 1996; Doherty 2006), Josep Call and Michael Tomasello have suggested that chimpanzees track the 'likely target' of others' visual access and understand something about its effects on behaviour (Call and Tomasello 2005: 58), and Andrew Whiten uses the notion of an "intervening variable" to explain primitive theory of mind notions (Whiten 1994, 1996). **Add Bartsch & Wellman (our proposal is also similar to theirs.)

cannot readily explain abilities to track false beliefs.

Minimal theory of mind is an attempt to build on these suggestions in saying what it is that infants, chimpanzees and scrub jays represent that enables them, within limits, to track beliefs and other mental states.

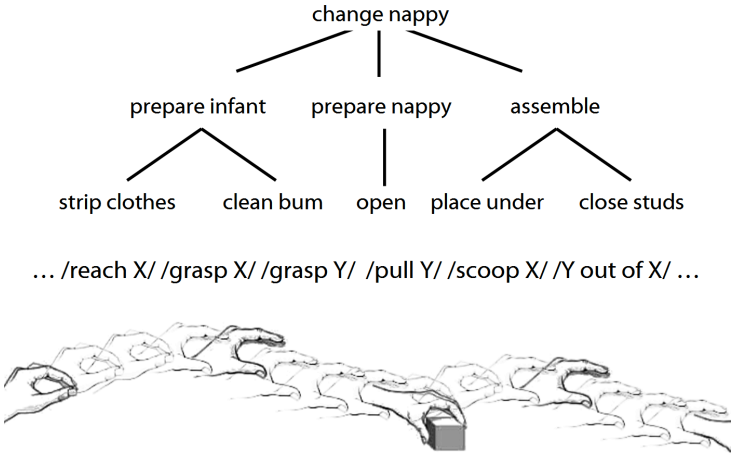
[*acknowledge Ian]

§ behaviour reading

Minimal theory of mind is based on behaviour reading. Because it has received scant attention in theory of mind research, I shall briefly explain what I mean by behaviour reading as a preliminary.

To understand behaviour reading we have to understand how behaviour is structured by plans and goals. As an illustration, take changing a nappy (see Figure 2). This intentional action involves several subplans—preparing the infant, preparing the nappy and assembling them. These subplans can be executed in any order except that the assembly stage needs to come last. Execution of subplans can also be interspersed so that you might start preparing the nappy midway through preparing the infant. Each of the subplans will be realized by one or more goal-directed action. These goal-directed actions are simpler than the planned actions. They do not have components that involve goals or plans, their components typically occur in a fixed order, and they cannot naturally be

interspersed (although they may overlap). Stepping down the hierarchy, each goal-directed actions is realized by a series of simple object-directed actions such as reaching, pulling, tearing and scooping. These are categories of motor-action that abstract from implementation details such as the use of particular types of grasp and particular body parts. The upshot of all this is a continuous stream of bodily movement.



e 2. *The structure of behaviour from bodily movem*
(The lower image is adapted from Rizzolatti, Fogas
Gallese 2001)

The goal of behaviour reading is to recover some or all of this structure by working backwards from the stream of bodily movement.

For present purposes we are interested in pure behaviour reading, that is, behaviour reading not involving knowledge of goals and plans.[3]

[short:] This involves segmenting the movement into chunks and assembling these chunks into larger units on the basis of statistical patterns in the ways they recur and other cues.]

[long:] First bodily movements are segmented into chunks corresponding to categories of motor action; this involves motor emulation (Rizzolatti, Fogassi and Gallese 2001; Byrne 2003). These chunks include simple object-directed actions, for example reaching for a grape or grasping a ball. At the next stage these chunks are

---

[3] For evidence that behaviour reading in human adults is sometimes impure—that is, sometimes influenced by knowledge of goals and intentions—see Zacks (2004).

assembled into units that coincide with goal-directed actions on the basis of sequential probabilities (Baldwin, Baird*, et al.* 2001; Saylor, Baldwin*, et al.* 2007; Baldwin, Andersson*, et al.* 2008). Finally, it may be possible to extract intention-sized complexes on the basis of hierarchal patterns and changes in motion features (Byrne 1999; Zacks, Tversky and Iyer 2001; Hard, Tversky and Lang 2006). By means of these several stages, behaviour readers may discern much of the structure of intentional action without representing goals or intentions.

The stages involved in pure behaviour reading resemble some stages of speech perception, which is also thought to involve motor emulation as well as sensitivity to sequential and hierarchical probabilities. In terms of this parallel, goal-directed actions

correspond to words and planned actions correspond to phrases. To some extent, we can think about behaviour reading as a more general form of speech perception.

Two features of pure behaviour reading are essential for what follows. First**, it does not require representing goals or intentions**. Second, **it does involve an ability to represent object-directed actions**.

Minimal theory of mind is an extension of pure behaviour reading.

§ Minimal theory of mind

I will describe it with a series of concepts and principles.

The first concept is the **field**. An agent's field is a set of objects related to the agent by proximity. The extent of the field depends on things like lighting and the agent's orientation. To say that an agent is **encountering** an object just means that it is in her field. The notion of encountering is a proxy for perception: within limits, an agent perceives an object just when she encounters one. But to think about encountering doesn't require understanding perception. **In particular it doesn't require understanding perceptual modalities, perspectives or appearances.**

The next concept is **goal-directed** action. The term 'goal-directed action' can be used to mean several things. One is intentional action. This notion is no use for constructing

a minimal theory of mind. To represent intentional actions as such you also have to represent intentions and related propositional attitudes such as belief (Davidson 1999). Constructing a minimal theory of mind requires a basic notion of goal-directed action.

[*presented in two phases. First: have goal-directed actions when have sequence of units of object-directed behaviours where the sequence has a function which is not a function of any individual unit.

*The question is then what we mean by function here. *Smuggling* danger. Either teleological or whatever notion of function infants have from tool use.]

We stipulate that for *g* to be the goal of a unit of object-directed behaviours is for two conditions to be met:

> (i) *g* would be the outcome *of the whole unit* if *g* occurred;
>
> and
>
> (ii) units of this type occur in order to bring about *g* (that is, *g* is the function of this unit).

This is not an account of full-blown goal-directed action. It is what someone who has

only a minimal grasp of goal-directed action might understand.[4]

The **first principle** links encountering with goal-direction actions. It says that you can only act on a goal involving an object if you have encountered it. So if you know that someone hasn't encountered an object, you can predict that they won't look for it.[5]

---

4 This teleological approach to characterising goal-directed action has been developed in different ways by several philosophers including Taylor (1964), Wright (1976) and Millikan (1993).

5 (In *this* experiment by Brian Hare and colleagues, a subordinate chimpanzee makes predictions about a dominant chimpanzee's ability to retrieve food. They found that the subordinate's predictions take into account whether the dominant's view was blocked while the food was placed. This could be explained by the First Principle. For the subordinate to predict that the dominant will not be able to recover the food, it is sufficient to think: because the dominant did not encounter the food, she will not be able to retrieve it.)

Equally, one way to prevent someone from being able to act on an object is to prevent her from encountering it. This principle isn't true, but it approximates a truth and is a useful heuristic for predicting action.

[Explains Level-1 perspective taking]

For the next principles we need the concept of **registration**. This is a relation between agents, objects and locations. An agent registers an object at a location just if she last encountered it there. Registrations can be correct or incorrect. A registration is incorrect when an agent registers an object at a location but it is not at that location.

Principle Two says that correct registration is a condition of *successful* action. If you have a goal involving an object, you will only

succeed if you register the object at its actual location. To illustrate, suppose you see someone encountering an object which is then moved while they are not watching it. In this situation, Principle Two allows you to predict that they will not be successful in retrieving that object.

One application of this Principle is to some of Liszkowski and colleagues' pointing studies. [*compressed] In one paper these authors showed that children point in order to provide information to others about the locations of objects needed to perform an action. The adults had previously used the objects so they were not unfamiliar to them; rather, they had recently misplaced them. This could be explained on the hypothesis that the 12- and 18-month-old subjects think of pointing as a way to get others to register

objects and understand this Principle that correctly registering an object as necessary for acting with it.

Principle Two doesn't enable you to pass false belief tasks. This is because when people have incorrect registrations, this Principle doesn't allow you to predict what actions they will perform. To do this you need **Principle Three**. With this Principle, we move from thinking of registration as a *condition* on action to thinking of it as a *cause* of action. Principle Three says that agents will act as if objects were actually at the locations they register them as being. What determines where an agent will look for an object is not its actual location but the location the agent registered it as being at. This Principle is sufficient for passing some false belief tasks.

This, then, is minimal theory of mind. My conjecture is that two-year-olds and perhaps adults too are sometimes able to track beliefs by means of representing encounters and registrations. Just as you might represent weight in order to track mass (within limits), so you can represent registrations in order to track beliefs (within limits).

§ [extensible]

These three Principles can be extended in various ways to explain further theory of mind abilities. Richer notions of goal-directed action could be invoked. Encountering and registration could also be enhanced. These were defined as relations involving agents, objects and locations. It is possible to elaborate these notions by

including other types of property as relata in addition to locations. Further notions, such as a notion of preference, could also be added.

§ Adequacy requirements

mToM must be theoretically plausible and empirically testable.

Theoretically plausible. It must be clear why cognition of encounterings and registrations should be less conceptually and cognitively demanding than cognition of propositional attitudes. [list features of propositonal attitudes …]

Second, must generate testable predictions.

…

§Signature limits

Signature limits of minimal theory of mind. Earlier I emphasised **signature limits**. The idea was that, by analogy with the case of number, any conjecture about how infants, animals and disrupted adults keep track of beliefs ought to generate predictions about the limits of their abilities. Like the three-item limit on number cognition, the limits should be explained by the nature of the mechanism rather than by the nature of number or belief.

What signature limit follows from the minimal theory of mind conjecture? As I said earlier, the essence of minimal theory of mind is the use of objects and their relations to agents, rather than representations of objects, to predict others' behaviours. This means false beliefs involving quantification

or identity cannot be tracked using minimal theory of mind.

[*What follows is presented differently. First note **formal** difference between belief and registration in terms of inferences that each can support. Then point out that this allows us to make **predictions** that distinguish subjects representing registrations only from subjects who are representing beliefs.]

To see why not, consider the following inference:

(1) Mitch believes that Charly is in Baltimore.

(2) Charly is Samantha.

Therefore:

(3) Mitch believes that Samantha is in Baltimore.

On almost any account of belief, this inference is not valid. As human adults typically appreciate (Harlin, 1996), until Mitch realises that Charly is Samantha he is liable to believe things about Charly which he does not believe about Samantha. By contrast, consider the corresponding inference in the case of registration:

(1') Mitch registers <Charly, Baltimore>

(2) Charly is Samantha.

Therefore:

(3') Mitch registers <Samantha, Baltimore>

This inference from (1') and (2) to (3') is logically valid. It is valid because registration is a relation to objects. We can compare registration with other relations like being left of something. If Charly is Samantha (whether you know it or not), then anyone who is left of Charly is left of Samantha; similarly for registering Charly's location.

This formal difference between belief and registration entails a limit on minimal theory of mind cognition. Consider Lucky who tracks beliefs by means of representing registrations only and is unable to represent beliefs. Lucky should have no problem predicting actions based on false beliefs about the locations of objects but she should encounter difficulties in predicting actions based on beliefs essentially involving mistakes about identity. In particular, Lucky

should not be able to understand why, when Mitch registers <Charly, Baltimore>, he continues searching for Samantha.[6] For to register <Charly, Baltimore> is the same thing as registering <Samantha, Baltimore>. And Lucky should be equally at a loss when those she observes mistakenly believe that two distinct people are identical. By contrast, subjects who can represent beliefs as such should have no special problem with false beliefs essentially involving identity (see, e.g., Coen & Coen, 1998). This, in barest outline, is how mistakes about the identities of objects can be used to distinguish minimal

---

6 This assumes that Lucky herself knows that Charly is Samantha. To ease exposition we assume throughout that Lucky has no false beliefs involving identity.

from full-blown theory of mind cognition.[7] Related points about quantification entail further testable distinctions.

This, in barest outline, is how mistakes about the identities of objects can be used to distinguish minimal from full-blown theory of mind cognition.[8]

---

7 Scott and Baillargeon (forthcoming) have conducted experiments designed to test false beliefs about identity. They report that 18-month-olds can predict actions based on false beliefs about the identities of objects. If this conclusion were true it would show that their subjects are not relying on minimal theory of mind. However, these results could be explained by registrations about types of object rather than beliefs involving identity. Part of the point of giving an account of minimal theory of mind is to be able to refine current tests of when and whether infants or children actually ascribe beliefs as such.

8 Scott and Baillargeon {\, forthcoming #1690} have conducted experiments designed to test false beliefs about identity. They

§But is it theory of mind?

we are not saying where theory of mind begins (so not identifying the minimal point): but we are saying that if you follow the construction we give, perhaps extending it in ways sketched, you get to what can genuinely be regarded as theory of mind cognition at some point. So we are ducking the issue of what the minimum is while giving a recipe for constructing it. There are two reasons for reticence: first, we are uncertain that, as presently used, there is a determinate point at which cognition becomes theory of mind cognition; second, our paper is an attempt to get away from the idea that theory of mind cognition is an all or nothing matter.

---

report that 18-month-olds can predict actions based on false beliefs about the identities of objects. If this conclusion were true it would show that their subjects are not relying on minimal theory of mind. However, these results could be explained by registrations about types of object rather than beliefs involving identity. Part of the point of giving an account of minimal theory of mind is to be able to refine current tests of when and whether infants or children actually ascribe beliefs as such.

Conclusion

The motivation for introducing minimal theory of mind was a puzzle about theory of mind abilities. Minimal theory of mind demonstrates that it is possible to succeed on a range of theory of mind tasks without representing perceptions, knowledge states or beliefs as such. Further, minimal theory of mind may be what makes those with limited cognitive resources and little conceptual sophistication, such as infants, chimpanzees or scrub-jays, sensitive to facts about perceptions and beliefs. Or, of course, it might not: the hope is that this is a conjecture worth testing.

While some details are missing and others may be incorrect, what seems certain is that there must be degrees or kinds of theory of mind cognition intermediate between pure

behaviour reading and full-blown propositional attitude psychology. Since our own commonsense psychological concepts do not enable us to understand how there could be degrees or kinds of theory of mind cognition, a constructive approach seems unavoidable. In describing minimal theory of mind I have offered one way of trying to understand how there can be intermediate kinds of theory of mind cognition. *Most important is to show that we need to replace reliance on introspection with more rigorous attempts to model forms of theory of mind cognition. As always the first step is admitting that we don't know what theory of mind cognition *is*.