Factive Theory of Mind

Jonathan Phillips

Department of Psychology, Harvard University

Aaron Norby

Department of Philosophy, Yale University

Abstract

Research on theory of mind has primarily focused on demonstrating and understanding the ability to represent others' non-factive mental states, e.g., others' beliefs in the false belief task. The motivation behind this focus has been that the representation of false beliefs are the best way to unequivocally demonstrate a genuine capacity for theory of mind, since they ensure that subjects' responses cannot depend on their own representation of the world. Here, argue that the false belief requirement confuses the ability to represent a particular kind of (non-factive) content with the more general capacity to represent another agent's understanding of the world. We then provide a way of correcting this error. We first offer a simple and theoretically motivated account on which *tracking* another agent's understanding of the world and keeping that representation *separate* from one's own are the essential features of a capacity for theory of mind. This account provides a straightforward way of understanding when factive representations, e.g., representations of what others know, provide evidence for a theory of mind capacity. We then develop a new test, the 'diverse-knowledge task', which shows how these criteria can be operationalized in an experimental paradigm. Finally, we turn to a number of existing examples from theory of mind research and illustrate (1) how to decide when a behavior does not demonstrate theory of mind, (2) why some existing research falls short of demonstrating the capacity for theory of mind (even when it focuses on false beliefs), and (3) where we've missed good evidence for theory of mind in a number of surprising places.

Keywords: theory of mind, knowledge, belief, factivity, false belief task

Factive Theory of Mind

On an island in Peurto Rico, sometimes called in 'Monkey Island', a group of experimenters approached free-ranging Rhesus macaques and offered them a single chance to steal food (Santos et al., 2006). While the monkey was watching, two translucent containers were placed on the ground and a grape was placed in each. Both of the boxes had jingle bells glued to their lids, but the ringers had been removed on one of the boxes, making one of the boxes silent and the other noisy. After placing the boxes on the ground, the experimenter retreated a couple of meters, and put his head between his knees so that he could not see the monkey or the boxes. Of the fourteen monkeys that attempted to steal food from the experimenter, twelve of them attempted to steal the food from the box without ringers, avoiding alerting the experimenter. Sometimes though, instead of putting his head down, the experimenter looked straight at the monkey as it tried to steal food. When the experimenter was watching, the preference for taking food from the silent box reversed itself; eleven of the sixteen monkeys tried to steal the grape from the noisy box with ringers. They no longer seemed to care whether or not the sound would alert the experimenter.

*

As early as 12 months after birth, prelinguistic infants begin to direct others' attention by pointing. They use this ability not only to draw adults' attention to things that they themselves find interesting, but also to direct adults' attention to things that they believe adults are interested in (Liszkowski et al., 2006). In one study, infants watched an experimenter

who displayed an interest in one of two 'adult' objects (e.g., a stapler and a hole-puncher). Both of these objects were then transferred to different locations that were out of the experimenter's view, and the experimenter then began searching for an object. 12-month-old infants consistently pointed to the location of the object the experimenter had previously expressed an interest in. A further set of studies employed a similar paradigm but also varied whether the experimenter had watched the objects be displaced (Liszkowski et al., 2008). When the experimenter had seen the object move to a new location, infants did not preferentially point to the object, despite the experimenter's unsuccessful attempt at searching. Yet once again, when the experimenter had not seen the object be displaced, and thus was ignorant of the location of the object, infants preferentially pointed to the location of the object.

*

One good thing about secrets is that they have a way of making almost any social situation more interesting. Whether they involve illicit affairs or just surprise birthday parties, secrets are primarily interesting because they're something that others don't know—not because they are one more thing that you *do* know. When you find out that someone is going to have a surprise birthday party, for example, it's certainly not the fact that they will have a birthday party that one finds interesting; it's that *they don't know* that they are having a birthday party. And then there's the intrigue that goes along with knowing a secret: not letting on that you know something that others don't; or trying to secretly find out if someone else already knows;

or subtly indicating to others that you too know. When things become most interesting is when the others find out that you know a secret. They know *that* they don't know, even if they don't know *what* they don't know. What they do know, though, is that you know.

1.1 What's compelling about these cases

What's compelling about each of these cases is that they illustrate the way in which different subjects seem to keep track of something about what other agents know or understand and then use that representation to achieve their own goals. Moreover, they all are cases where the subject takes advantage of the fact that others understand the world in a way that differs from the way that they do. In short, these seem to be the kind of paradigmatic instances of representing others' minds that researchers working on theory of mind should be interested in. Perplexingly though, examples like these aren't often taken to be all that important. According to traditional wisdom, these cases don't demonstrate the core ability required for a genuine theory of mind because they don't require the ability to represent *false beliefs*. In the view of received wisdom, these cases may be thought of as suggestive or interesting, but the idea that they are clear examples of the essential abilities for a theory of mind is generally taken to be misguided.

We must have taken a wrong turn somewhere. There are cases of clever behavior that do not demonstrate a genuine ability for theory of mind (such as gaze-following), but these cases aren't like that. In some of them, our own experience tells us that they involve representing others' minds; in others, researchers have gone to great lengths to document the way that these sub-

jects can flexibly make use of others' mental states in a way that is unlikely to be explained by any set of behavioral tricks. Despite this, the standing sentiment in theory of mind research is that these sorts of cases belong on the periphery and not at the center of theory of mind research. This seems odd, and it's worth taking a moment to look back and consider how we got here.

1.2 A brief history of false beliefs

The rise of false beliefs in research on theory of mind is surprisingly easy to track. In 1978, Premack and Woodruff published an article in *Behavioral and Brain Sciences* called "Does the Chimpanzee have a theory of mind?"(Premack and Woodruff, 1978). They argued, based on evidence that chimpanzees could identify the solution to problems that other agents faced, that chimpanzees could recognize other agents' goals or intentions, and thus had a theory of mind. In the commentary to Premack and Woodruff's article, three or four philosophers (depending on how one counts philosophers) argued that this kind of evidence was not sufficient to demonstrate a capacity for theory of mind (Dennett, 1978; Bennett, 1978; Harman, 1978; Pylyshyn, 1978). The problem, they pointed out, was that the experiments did not dissociate chimpanzees' representations of others' mental states from chimpanzees' own representations of the world. The chimpanzees' behavior could be explained just in terms of their own understanding of the world – their knowledge about which solutions solved which kinds of problems – and they need not have actually represented another agent's goals at all.

An alternative test was proposed: researchers should examine whether

chimpanzees could represent *false* beliefs. The reasoning was that representing a false belief requires that you represent a belief that differs from your own. Otherwise you wouldn't treat it as false. False beliefs guarantee the necessary dissociation between representing the world itself and representing another agents' mental states. Thus was born the gold standard for determining whether a subject has a genuine theory of mind.

Soon after the commentaries were published, Wimmer and Perner (1983) began testing children's ability to correctly predict others' actions based on their false beliefs. Shortly after that, Baron-Cohen and colleagues developed and published the now classic Sally-Anne version of the false belief task in their article, "Does the autistic child have a 'theory of mind?'" (Baron-Cohen et al., 1985). Both of these articles cite the commentaries in *Behavioral and Brain Sciences* when explaining why false belief representation was the appropriate test of the capacity for theory of mind. And now, thirty eight years later, we often find ourselves teaching introductory students that having (or developing) a theory of mind requires the capacity to pass a false belief test, and we've collectively written roughly 8,000 papers employing or discussing one version or another of this test.[1]

1.3 The trouble with false beliefs

The trouble with false beliefs is that they don't carve the world at its joints. There's more to theory of mind than false beliefs, and much of the time we spend representing or reasoning about others' minds, we actually aren't

[1] According to a highly informal Google Scholar search for the term "'false belief task' OR 'false belief test'" completed by the first author around the end of August, 2017.

concerned with what others falsely believe. More often, we're just interested in keeping track of what they *know* (or whether they too *saw* something, or whether they *remember* that time when, or if they *recognize* the person you're pointing out). Borrowing a term from linguistics, we can call all of these ways of representing and reasoning about others' minds, '*factive* theory of mind'.[2] These representations of others minds are factive because they're directly tied to the way we take the world to be. When you know the Queen of England is in Scotland, you can't represent someone as having *seen* her in Paris. That's just not the way that representations of seeing work. There are no false seeings. The same is true for knowing, recognizing, realizing, and so on. There are false beliefs though. You can also imagine things that are false, make a wrong guess, and think incorrectly. Representations of belief, imagining, guessing, thinking, and so on are all 'non-factive' — the great thing about them is that they aren't tied to the way you think the world actually is.

We seem stuck. On the one hand, we know that there are a lot of phenomena that certainly *seem* to involve keeping track of what others take the world to be like, and then using that knowledge to predict or explain others' behavior. Yet, for all the reasons pointed out thirty nine years ago, we're wary that these kinds of *factive* theory of mind won't pass muster. They may not be good evidence for genuine theory of mind representations, since these representations are, by definition, confounded with the way one actu-

[2]While these details won't concern us too much, the distinction between factive and non-factive attitudes is roughly that factive attitude ascriptions, e.g., those of the form S knows that p, presuppose that the complement p is true, while non-factive attitude ascriptions, e.g., those of the form S believes that p, do not presuppose p is true (Kiparsky and Kiparsky, 1970).

ally takes the world to be. On the other hand, we have 'non-factive theory of mind' which we are pretty certain is good evidence for genuine theory of mind representations, but focusing on these as the core ability of theory of mind ignores a large majority of what we do when we represent and reason about others' minds.

Up til now, most people have tended to agree we're stuck, and so we've gone with the sure bet. Having a theory of mind has required non-factive representations, and this has been the standard for how we decide, for example, which species actually have a genuine capacity for theory of mind (Heyes, 1998; Call and Tomasello, 2008; Drayton and Santos, 2016; Martin and Santos, 2016), when in the course of the human lifespan a capacity for theory of mind develops (Onishi and Baillargeon, 2005; Wimmer and Perner, 1983; Kovács et al., 2010; Baron-Cohen et al., 1985), and even which brain regions are responsible for representing and reasoning about others' minds (Saxe and Kanwisher, 2003; Gallagher and Frith, 2003; Gweon et al., 2012; Frith and Frith, 2012; Koster-Hale and Saxe, 2013).

But we don't agree that we're stuck, we just think that we took a wrong turn, and that moving forward is going to require backtracking a bit. We think there's a principled way of showing when and how factive representations are genuine theory of mind representations. We also think there are good reasons to believe that having a capacity for theory of mind doesn't and shouldn't require the ability to represent false beliefs, or even beliefs at all. And we also think that getting clear on this will significantly reorganize the way we think about both the past and the future of theory of mind research. That's where we're going, but we're going to start by going back

to the basics.

2 Back to the basics

There is still a great deal we don't know about theory of mind. For example, there's some evidence that infants have the ability to succeed on non-verbal false-belief tasks (Onishi and Baillargeon, 2005), and thus that they have some theory of mind ability. But very little is known about the principles by which infants are able to do this, even at the most abstract level of description. Similar points can be made about theory of mind in non-human primates, and in many respects, about theory of mind in human adults as well. This is not to say that past forty years of theory of mind research have not been both impressive and productive; they have. The point is rather that theory of mind research has moved forward to an impressive degree without working out a well-motivated understanding of what an ability for theory of mind, at bottom, is.

A natural objection to this starting point would be to argue that, although there is no agreed upon framework for understanding theory of mind, theory-theory, simulation theory, and hybrid accounts all are aimed at understanding how theory of mind "works" and what it is. Simulation theory tells us theory of mind involves simulating the mental states of others and running them off-line (Goldman, 2006; Gordon, 1986). Theory-theory tells us that theory of mind involves a scientific-theory-like set of principles about how the mental states of other agents are formed and influence their behavior (Gopnik and Wellman, 1992). Hybrid theories tell us that theory of mind

involves some combination of both of these (Nichols and Stich, 2003).

But these theories all presuppose an account of the basic functional or computational features of theory of mind, rather than providing one. It shouldn't be too hard to see that this is the case. Assume for the moment that the way theory of mind representations are produced and manipulated is broadly via a simulationist architecture. Even if this is the case, we still seem to be left with a basic unresolved question: "Which of the representations produced by this sort of simulationist architecture are actually *theory of mind* representations and which are instead some other kind of simulation-based representations?" Other types of representations can of course be produced by the same sort of simulation (e.g., what part of the room I would be in if I were in another person's position). The unresolved issue we are pointing out concerns what the computational role of these representations must be for them to be *theory of mind* representations in particular. It's also important to notice that the answer to this question is not going to be as simple as, "When the representations support prediction and explanation of other agents' behavior," since many of the representations that support the representations of others' minds are not themselves theory of mind representations (e.g., visual representations of other agents). Even more problematically, it is widely recognized that in some cases explanations and predictions of other agents' behavior can be accomplished without involving any genuine theory of mind abilities (Andrews, 2012; Gergely and Csibra, 2003; Penn and Povinelli, 2007; Perner and Ruffman, 2005; Butterfill and Apperly, 2013). The issue here is in no way specific to simulation theory. Similar questions can be raised with respect to theory-theory, e.g.,

which representations produced and manipulated by theory-like structures are *theory of mind* representations and which are not? Combining the two, like hybrid theories do, obviously won't solve the problem either.

These are all theories about the kind of processes that are responsible for producing and manipulating theory of mind representations. Accordingly, they don't actually provide an account of what makes a representation a *theory of mind* representation, but rather assume that there is a worked out way of determining which representations are theory of mind representations, and then proceed to argue over which kind of processes underwrite *those* representations. The problem with this assumption is that we haven't actually spent much time working out an account of what the functional or computational features a mental representation would have to have for it to count as a genuine theory of mind representation (or really even exactly what ability one has when one has the capacity for theory of mind).[3]

Researchers working on theory of mind cannot simply set this question to one side to be resolved later; the answers to many central questions in theory of mind are going to depend on how this issue ends up being resolved. As we pointed out earlier, consider the debate about when during human development theory of mind emerges (Carruthers, 2013; Onishi and Baillargeon, 2005; Surian et al., 2007; Fabricius et al., 2010). The answer to this question is going to depend straightforwardly on what turns out to be required for a mental representation to be a theory of mind representation.

[3]One attempt that we are aware of dates back to 1987 (Leslie, 1987). Like much of the empirical work since the commentaries on Premack and Woodruff's article on Chimpanzee theory of mind (1978), this account simply assumes that genuine theory of mind representations must be non-factive, and characterizes the computational features of non-factive representations. We return to the way our account differs in more detail in § 4.2.

The same is true of the debate over whether theory of mind is a uniquely human ability (Call and Tomasello, 2008; Martin and Santos, 2014; Penn and Povinelli, 2007; Lurz, 2011). And the same is true for which brain regions are responsible for representing and reasoning about others' minds (Saxe and Kanwisher, 2003; Gallagher and Frith, 2003; Gweon et al., 2012; Frith and Frith, 2012; Koster-Hale and Saxe, 2013). In all of these cases, we suspect that at least part of the disagreement over these issues comes from the fact that we have been proceeding without a more worked out idea of what it is exactly we are looking for when we're looking for theory of mind.

3 Mental Representations

In restarting, it's going to be helpful to set aside theory of mind and focus on a broader question: How do we discover whether a subject is able to represent some particular property? Suppose for example, that we want to know whether human infants have the ability to represent *number*. How is it that we go about determining whether or not they can do this?

One obvious place to look is the experimental work that's been done to answer this question. Typically, the stimuli used in these studies are arrays of dots, and the question asked is whether infants represent the number of dots in these arrays. This is standardly determined by testing whether infants' behavior indicates that they differentiate between arrays with different numbers of dots (for reviews, see Carey (2009); Feigenson et al. (2004)). For instance, infants might be habituated to arrays of seven dots in varying positions and then on a test trial be shown either a new array containing

seven differently arranged dots or an array that instead contains three dots. What's then measured is whether infants dishabituate and look longer at the three-dot array than at the new seven-dot array. But even if the infants do look longer at the three-dot array, what has to be shown is that infants' behavior is specifically sensitive to the number of dots rather than something else that naturally covaries with number. Thus, experiments are designed to vary number while holding fixed properties like density and area of the array, the size of individual dots, and so on (Dehaene and Changeux, 1993; Gallistel and Gelman, 1992; McCrink and Wynn, 2007; Xu et al., 2005).

The purpose of this methodology is clear: we want to show that the subjects are tracking *number* and not something else. And if it can be shown that their behavior is tracking number and not anything else, then we can be reasonably satisfied that, in some way or another, infants are capable of *representing* number in one way or another. Moreover, it is often claimed that this tracking methodology can not only reveal *that* number is represented but can also provide information about *how* number is represented. For example, there is now substantial evidence which suggests that there are multiple systems for representing number, one of which precisely represents small numbers of objects and one of which approximately represents larger numbers (Carey, 2009; Feigenson et al., 2004).

Number cognition is only one example, and although the picture we've given is simplified in various ways, the methodology here is representative of much of the work that goes on in cognitive and developmental psychology, in vision science, and in related fields. Whether we call it 'tracking' or 'sensitivity' or 'attunement', the basic idea is that the ability to respond dif-

ferentially to some particular feature of the world (and not something that simply covaries with it) is directly tied to the ability to represent that thing itself.

Returning now to research on theory of mind, however, what is odd is that this does not seem to be the prevailing methodology. Rather, what is often demanded is not so much a sensitivity to the mental states of others, but instead an ability to represent a particular feature of some mental states, namely, non-factivity. Requiring that one be able to represent the falseness of beliefs is more similar to requiring that one be able to represent the oddness of number. Sure – if you could do it, no one would doubt you could represent number, but it'd hardly be a reasonable test of whether you could had the capacity for representing number in the first place.

Our suggestion here is not meant to be radical: it is merely that we ought to frame our understanding of representation in the theory of mind domain in the same way that we frame our understanding of representation in other domains. And throughout much of psychology and philosophy, tracking is taken to be the basis, if not the whole, of representational capacities. In our view, this is right. Being able to *track* the minds of others is the first step in being able to think any kind of complex thought about others' minds. It is the *basis* of an ability for theory of mind.

4 What is essential for theory of mind

We've pointed to the close connection between tracking a property or object and representing that same property or object, and here we want to turn

this into a full-fledged theory of theory of mind ability. The theory has two key requirements. The first is that an organism be able to *track* the contents of other agents' representations of the world; the second is that the organism be able to keep the outputs of its tracking mechanism *separate* from its own representation of the world. If both of these are satisfied, then the organism has the capacity for theory of mind. Our claim is that this is the core of theory of mind ability, around which more and less sophisticated capacities are built.

The first condition of the theory—that mindreaders be able to track others' representations of the world (their 'perspective' for short)—is clear enough. Once we recognize that representation is a matter of tracking or sensitivity to a feature (in the sense described in § 3), then it's obvious that the perspectives of other agents are the things that need to be tracked if one is to represent those perspectives. This makes sense not only theoretically but also from a practical research perspective. The place to start when trying to determine whether some species or other has theory of mind ability is to find out whether they are sensitive to changes in others' representations of the world. The first step in answering the question, "Do X's have theory of mind?" is asking, "Are X's sensitive to changes in the other agents' understanding of the situation?"

However, having a theory of mind ability is a matter not only of tracking and thus having some way of representing the content of another agent's perspective, but also of *separating* those contents to that other agent. In other words, in order to utilize theory of mind, I have to be able to predict your behavior specifically on the basis of how *you* represent the world, not on

what I think the world really is like. So if I simply attribute my perspective to you (or confuse your perspective with my own), then I won't have utilized any theory of mind capacity. Predicting what you will do based on my own representation of the world is not sufficient; predicting what you will do based on your understanding of the world is.

Thus, in order to represent the minds of other agents in the sense important to theory of mind research, a subject has to be able not only to keep track of others' perspectives, but also keep those representations *separate* from the subject's own perspective. Theory of mind, then, is a matter of both tracking and separation. Tracking demonstrates that you have a representation of another agent's perspective. Separation demonstrates that this representation is a specifically theory of mind representation, rather than just a representation of what the extra-mental world is like. This is the core ability one has when one has a capacity for theory of mind.

4.1 A rough analogy: Separate maps

To get an intuitive grasp on the sort of picture we're advocating, it may be helpful to temporarily conceive of the suggestion with the aid of a more concrete analogy. Try thinking of subjects' representations as *maps*. My own map is just my representation of the world, the way that I take the world to be. Other agents have their own maps, each of which captures what the world appears to be like from that agent's perspective. On this rough analogy, theory of mind consists in tracking aspects of another agent's map (representing a second map of the world in addition to my own), and keeping that map separate from my own.

To see what this sort of representation could allow one to do, consider a simple example. Imagine that you can see an opaque box on the floor, and that you know that there is a banana inside the box. In the maps analogy, this is to say that your map represents there being a banana inside of a box. Now imagine that a second person comes along and can also see the box.

First, let's suppose that you are incapable of constructing multiple maps - all you have is your own map of what the world is like; that is, you have no capacity for theory of mind. In a case like this, you can still go some way toward predicting, manipulating, and understanding the behavior of this other person. You can do this by (tacitly) treating the other person as though they have the same map as you. Supposing for example that the person is looking for bananas, you could predict that the person will go after the banana in the box. What you won't be able to do, however, is represent the person as being ignorant of the fact that there is a banana in the box.

Suppose now that you also have an altercentric map that tracks some aspects of the other agent's representation of the world and is functionally distinct from your own map. Let's even suppose that it's an incredibly simple kind of altercentric map – one that is incapable of representing anything that is inconsistent with your own map. All you can do is either represent the other agent as realizing the way the world actually is, or else as not being aware of certain small pieces of it. To extend the separate maps analogy, we can imagine that you can construct altercentric maps by "removing" parts of your own map and replacing them with question marks and then using this altered map to predict the other person's behavior. With this ability, you'd now be able to successfully represent the other person as ignorant of

the fact that the banana is in the box.[4] Supposing that the person is still looking for bananas, you'd have the capacity, for example, to realize that the person won't look for the banana in the box. You'd also have the capacity to know that you should wait until the person leaves the room before retrieving the banana from the box. These later sorts of abilities, though, require two functionally separate maps. Here's why: if you remove pieces from your own map to represent the ignorance of someone else, the removed information must be somehow retained so that you yourself do not become ignorant; and that requires representing two separate perspectives of the same situation, which is to say that it requires two functionally distinct maps.

4.2 Factive Theory of Mind

A straightforward consequence of the account offered here is that a capacity for theory of mind can be had without the ability to represent *non-factive* mental states. One can, for example, have a capacity for theory of mind without being able to represent S's *belief* that p. If the core capacities of theory of mind are tracking and separation, then the representation and attribution of factive attitudes (S's knowing that p or S's not knowing that p) is sufficient for theory of mind. All that is required is a demonstration that these factive attitudes both *track* another agent's perspective and are kept *separate* from the agent's own representation of the world.

While this suggestion may seem straightforward enough, if it's correct, it should radically change how we think about what is and is not evidence

[4]This is similar in some sense to the way that Subsystem 1 is suggested to function in Scott and Baillargeon's discussion of infant theory of mind (Scott and Baillargeon, 2009))

for a genuine capacity for theory of mind. Perhaps the best way to make this difference clear is to apply it to one of the controversial cases with which this article began. Consider the study of Rhesus macaque theory of mind by Santos et al. (2006). In this study, the animals faced a human competitor and two visually identical boxes, each containing food. One of the boxes would make noise if opened while the other would not. When the experimenter positioned himself so that he could not see the monkey or the boxes, but the monkey could see the experimenter, 12 of 14 monkeys attempted to steal food from the silent rather than from the noisy box. In fact, this is something that the animals figured out on their very first attempt after being presented with the boxes. This latter fact is important, because it strongly suggests that the animals were making inferences about what to do based on what the experimenter knew, rather than simply applying some more simple form of associative learning that would link silent boxes with being able to retrieve food.

What representations then do the monkeys need in order to succeed as they do on this task? An obvious answer is that they must infer that they are more likely to get food if they attempt to steal from the silent as opposed to the noisy box. But how do they figure *that* out? Going after the silent box looks like a good strategy only if one realizes that one's approach will not be part of the competitor's representation of the world. That is, only if one realizes that the competitor's representation of things will not include one's opening the box and taking the food.

Thus, the animal must be able to *track*, in part prospectively, what things are like from the competitor's perspective, recognizing that its manipulation

of the silent box will not lead to the competitor updating his map of the world to include the attempt to take the food. Moreover, the animal's own map of the world must be kept separate from this representation of the competitor's understanding. Even though the animal's picture will include the approach and retrieval attempt, this representation must be separately maintained, because it needs to simultaneously know both that it is approaching the silent box and that the experimenter doesn't understand that this is the case.

In other words, success on this task seems to require both *tracking* and *separation*. It requires maintaining a representation of what things are like from another's perspective, and using that representation to predict and manipulate the other's behavior while not treating that agent's representation as an accurate picture of the world (i.e., keeping it functionally separate from the way the monkey takes the world to actually be).

4.3 Factive vs. non-factive theory of mind

The ability that we've argued these monkeys are exhibiting—the ability to represent an agent's understanding of the situation as including some facts but not others, while simultaneously maintaining their own separate representation of situation—would only provide a capacity for *factive* theory of mind. This ability allows one to represent which things others know and which things they don't, but there would still be things one couldn't do. One couldn't, for example, pass a standard false-belief task.

It's not hard to see why. Returning to the earlier example, suppose that the other person originally saw the banana being put into the box, but then

didn't see the same banana being moved to another location. Factive theory of mind would allow one to represent the person as *not knowing* where the banana is. But that won't help you pass the false belief task; all that would allow you to do is have no idea where the person would look for the banana. Representations of *knowing* won't help either. You don't take the banana to actually be in the box, so you can't represent the other agent as *knowing* that. Neither will it help to represent the agent as knowing where the banana actually is, since this would lead you to make precisely the wrong prediction about where the agent will look. To be able to correctly represent the agent's understanding of the situation, you'd need to construct an altercentric maps that directly contradicts the way you take the world to be. Your own map of the world would need to represent the banana not being in the box while you simultaneously construct and maintain an altercentric map in which the banana is in the box. That is to say, you'd need a capacity for *non-factive* theory of mind. Without it, you couldn't represent the person as falsely believing that the banana is still in the box, and you wouldn't be able to predict that this is where they'd look for the banana.

It's worth getting clear on exactly what the difference between factive and non-factive theory of mind boils down to. In a certain light, the two can seem quite similar. In one case, you're systematically tracking what the person represents as *not* being the case; in the other, you're systematically tracking what else the person takes as being the case. When understood this way, it's easy to wonder what the real difference is supposed to be and why it would matter so much. We suspect that a lack of clarity on this issue is in large part what has led us down the wrong track in theory of mind research,

and that once we get clear on what the difference actually is, a number of other important (and missed) distinctions in theory of mind will also begin to come into focus as well.

4.4 What the difference is

To be somewhat precise in characterizing the difference between factive theory of mind and non-factive theory of mind, it'll be important to formalize things a little bit. Wherever possible though, we'll stick to a high-level description of the formal details, but we'll try to keep the details nearby for those who are interested.

Let's start by taking your own map to be the way you take the relevant part of the world (call it the situation) to actually be. Not too much will hang on it, but for the purposes of simplicity, let's suppose that we can represent each of the various things you take to be true about the situation as propositions. We can now think of your map as just the set of these propositions. To illustrate, suppose you take there to be a banana in the box. We'll take the proposition that there is a banana in the box, and assume that your map will include that proposition, along with all of the other things you take to be true about the situation. We can also think of the map that represents the other agent's understanding of the situation in the same way.

Within this slightly more formal way of characterizing maps, we can now restate what the core abilities of theory of mind are. *Tracking* requires that one dynamically update the other agent's map in a way that reflects changes in the way that the agent takes the situation to be. *Separation* additionally requires that the other agent's map is maintained and updated

independently of your own map. One must be able to update one's own map in a way that does not demand a corresponding change to the altercentric map, and vice versa. With just these pieces, we now have everything we need to characterize the difference between (1) *not* having a capacity for theory of mind, (2) having a capacity for *factive* theory of mind, and (3) having a capacity for *non-factive* theory of mind.

If one does not have a capacity for theory of mind (either because one does not track the other agent's understanding of the situation, or because one does not keep that map separate from one's own), then the other agent's map will simply be identical to one's own map. To the extent that one predicts or explains others' behavior, these predictions and explanations will simply have to draw on a single map which represents both one's own understanding of the situation and the understanding attributed to others.[5]

By contrast, factive theory of mind allows one to construct and update an altercentric map that is *not identical* to one's own. For example, the other's map be a proper subset of one's own. In this case, your map may contain the proposition that the banana is in the box, but the altercentric map may not. This is what it means to represent another person as *not knowing* that the banana is in the box. This capacity to represent two non-identical maps of the same situation, however, would not by itself allow for your own map to be *inconsistent* with the other's map. As long as your understanding of the situation itself does not involve contradictions, no subset of the things you

[5]That is, let M_S be the set of propositions $\{p_1, p_2 \dots p_n\}$ that you take to be the case, and M_O be the set of propositions $\{p_1, p_2 \dots p_n\}$ that you represent the agent as taking to be the case. If one does not have any capacity for theory of mind then $\forall p : p \in M_O \iff p \in M_S$.

take to be true of the situation will be inconsistent with your understanding of the situation. Thus, when you take some thing to be the case, factive theory of mind allows you to represent someone as *not representing* that thing, but it does not allow you to represent someone as representing that thing *not being the case*.[6]

To be able to do that, one must have the capacity for non-factive theory of mind. With this ability, the other agent's map can *both* be non-identical to your own and can also be inconsistent with it. Non-factive theory of mind allows for it to be the case that your own map contains some proposition and simultaneously, the other agent's map contains the negation of that proposition (or some set of proposition that are jointly inconsistent with the proposition your map contains). The upshot of having non-factive theory of mind is that you can both represent the other as not representing something you take to be the case and also as representing that this thing is not the case.[7]

This slightly more precise way of thinking about factive and non-factive theory of mind should make it easy to see what the difference between them is. It is not a matter of whether one is tracking another's understanding of the world, or even whether one is keeping that representation separate from your own; both factive and non-factive theory of mind require this. The difference is just just a matter of whether one can construct and maintain a particular kind of representation: a non-factive representation. That is, one

[6] If one is exercising *only* one's capacity for *factive* theory of mind, then $\exists p : p \in M_S \wedge p \notin M_O$. However, $\forall p : p \in M_S$, $\neg p \notin M_O$ and $\forall p : p \in M_O$, $\neg p \notin M_S$.

[7] If one is exercising the capacity for non-factive theory of mind, then $\exists p : p \in M_O \wedge \neg p \in M_S$.

that is *inconsistent* with the way you take the world to actually be.

4.5 What the difference is not

When seen for what it is, it should be clear that the difference between factive and non-factive theory of mind is simply not a difference of *whether or not* one represents others' minds. Rather, it's a difference that concerns what kind of content one has the ability to represent, in general.

An easy to way to see that the ability that allows for non-factive theory of mind representations is not essentially about theory of mind is to notice that the same ability also allows for other completely non-social representations. Consider, for example, the difference between hypothetical and counterfactual reasoning. When reasoning hypothetically, you consider what would happen *if* a certain state of affairs were to occur; when reasoning counterfactually, you consider what would have happened if a different state of affairs had occurred *rather than* what actually happened. In both cases, one makes predictions about what would happen when conditions are different from the way you take them to actually be. However, only counterfactual reasoning requires constructing and maintaining a representation that is inconsistent with your own understanding of the world.

To succeed at counterfactual reasoning, you have to consider what would have happened if something had been *different than it actually was* — you have to reason counter-to-the-facts. Doing so requires a non-factive representation in precisely the same way as success on the false belief task. In both, you must construct and update a representation of the situation that is inconsistent with the way you take the situation to actually be and then

use that representation to make predictions about future states of affairs.[8] By contrast, reasoning hypothetically does not require representing a situation that is inconsistent with your own. One can reason about what would happen if p were the case while simply not having any belief as to whether or not p is the case. Just as with factive theory of mind, hypothetical reasoning requires a representation that is *different* from the way you take the situation to actually be, but it does not require that one represent a situation that is inconsistent with your own understanding.[9]

The key point here is that the ability that allows one to move from reasoning hypothetically to reasoning counterfactually is precisely the same kind of ability that allows one to move from factive to non-factive theory of mind. In both, it is the ability to represent a particular type of content: non-factive states of affairs. Thus it should not be surprising that if you can't represent states of affairs that are inconsistent with the way you take the world to be, then you can't represent another person representing states of affairs that are inconsistent with the way you take the world to be. Perhaps then, it also should not so surprising that young children who cannot pass simple verbal counterfactual reasoning tasks also cannot pass the verbal false

[8] Formalizing this makes the similarity perfectly clear. Once again, take your representation of the actual situation, M_S, to be the set of propositions that you take to have actually occurred, $\{p_1, p_2 \dots p_n\}$. Now suppose you are asked what would have happened if something else had occurred instead of what in fact did. To do this, we would need to construct a separate representation of the situation, M_{CF}, which involves some particular thing being the case even though it in fact was not, $\exists p : p \in M_{CF} \wedge \neg p \in M_S$. Similar to the false belief task, one then makes predictions based on M_{CF} rather than M_S. See (Kratzer, 2012; Peterson and Bowler, 2000; Grant et al., 2004) for related discussions.

[9] Reasoning hypothetically requires that you construct and update a representation, M_H that differs from your own, M_S, such that the hypothetical situation involves some particular component, p, that is not part of your representation of the actual situation (e.g., $\exists p : p \in M_H \wedge p \notin M_S$). While these two ways of representing the situation are not identical, neither are they inconsistent with one another, i.e., $M_S \cup M_H \neq \varnothing$.

belief task, though they can pass strikingly similar hypothetical reasoning tasks (Rafetseder et al., 2010; Riggs and Peterson, 2000; Peterson and Riggs, 1999; Riggs et al., 1998). In much the same way, it also should not be surprising that people with Austism Spectrum Disorder have difficulty not only with false belief reasoning but also with counterfactual reasoning (e.g., Peterson and Bowler 2000), while they have very little trouble with theory of mind tasks that do not require representing non-factive content (e.g., Tan and Harris 1991).

The suggestion here isn't that the ability that allows for both non-factive theory of mind and counterfactual reasoning isn't a critically important ability; it clearly is an incredibly productive ability. Rather, the argument is that it is simply not an ability that is essentially concerned with theory of mind representations, and that any test that makes it essential would not be a very good test of theory of mind.

4.6 Altercentric vs. egocentric ignorance: A missed distinction

We've illustrated that factive theory of mind allows for you to track other agents' understanding of the world even when it is not identical to your own. Excluding the representation of inconsistent maps though, there are still two different ways in which others' maps could differ from yours. One way would be for you to represent the other agent as being ignorant of something you know (call this 'altercentric ignorance'). Another would be for you to represent the other agent as knowing something you are ignorant

of (this would then be a representation of 'egocentric ignorance').[10] The ability to represent altercentric ignorance allows you to represent another agent as not knowing where you placed a banana while they were out of the room. The ability to represent egocentric ignorance, on the other hand, allows you to represent the other person as knowing where they placed the banana while you were out of the room.

A full-fledged capacity for factive theory of mind should allow one to represent both the agent as knowing more *and* less than you, since neither of these representations will be inconsistent with yours. However, there's still an important distinction between these two kinds of ignorance. It turns out that representations of egocentric ignorance are necessarily more complex than representations of altercentric ignorance. Here's why: to represent altercentric ignorance, you can simply take your map, remove the parts the other agent doesn't know, and then attribute that map to the other agent. But what about egocentric ignorance? What kind of map do you attribute to someone when they know more than you do? It'd be nice if you just take your own map and then add the propositions that the other agent knows and then attribute that map to the other agent. But obviously, you can't do that. You have no way of knowing which propositions the agent knows (if you did, you wouldn't be ignorant of them). So you can't construct representations of egocentric ignorance in the same way as representations of altercentric ignorance. But, if that's right, then how do you do it? The solution to this problem requires that you instead construct and attribute a more complex

[10] That is, $\exists p : p \in M_O \wedge p \notin M_S$. Recall, by comparison that altercentric ignorance of factive content is instead captured by the fact that $\exists p : p \in M_S \wedge p \notin M_O$.

kind of representation that involves a set of maps. Seeing why may be easier if we return to our running example.

Let's suppose that you don't know where the banana is, but you know that the other agent knows where it is. If your capacity for factive theory of mind is working correctly, you should not be surprised if, on the very first try, the agent looks for the banana where it actually is. Say, the agent looks for the banana behind the door and finds it there. It seems that the reason you are not surprised by this is that you understand that the agent had a map in which the banana was behind the door.

Attributing this map to the other agent is a step in the right direction, but by itself, it isn't yet enough to fully capture your representation of the other agent's knowledge. After all, you also wouldn't have been surprised if the agent looked for the banana in one of the boxes instead and found it there. So you must have also been representing it as possible that the agent's map was one in which the banana was in that box. In fact, the same thing is true for any location in the room where the banana might be hidden, since you don't know where the banana actually is. Generalizing then, what we are left with is a set of all of the different maps that you think the agent might have had. We could think about this as a map of all of the different possible maps. And this is precisely the kind of more complex representation that you have to use when representing *egocentric* ignorance. What you track and update is not a single map, but a map of maps.[11]

[11] In the case of altercentric ignorance, $\exists p : p \in M_S \wedge p \notin M_O$. By contrast, *egocentric* ignorance requires the following: Let P be the set of propositions $\{p_1, p_2, p_3, \ldots\}$ that are not part of your understanding of the situation (i.e., $P \cap M_S = \varnothing$) but which, for all you know, may be true of the situation (i.e., $\forall p : p \in P, p \cap (\bigcap M_S) \neq \varnothing$) and which are not inconsistent with M_O (i.e., $\forall p : p \in P, p \cap (\bigcap M_O) \neq \varnothing$). Let $\mathcal{M}_O$ be the set of

The basic insight is that in cases egocentric ignorance, you know *that* the agent knows something you don't, but you don't know *what* they know, so you have to represent their knowledge by representing a map of all the different maps they might have. This kind of representation is necessarily more complex than representing altercentric ignorance, since that just involves attributing a single map to the other agent. Neither of these, however, require representing or attributing a map that is inconsistent with your own.

At the same time, though, it is worth pointing out that this distinction is actually not specific to factive theory of mind; it also straightforwardly applies to non-factive representations. Here's how to see it. First, to ensure we've got a case of non-factive theory of mind, we'll use an instance of false belief. Start with a simple case in which the agent put the banana in the box, but then left the room and you moved the banana to a different location. In this case, you can represent the agent as having a false belief, and you know precisely what the false belief is. In other cases though, you can know *that* the agent has a false belief, but you don't actually know *what* the false belief is. Suppose that the agent put the banana somewhere in the room but you don't know where. Then, another person removed the banana, and you saw her put it behind the door. Similar to the previous case

all maps $\{M_{O_1}, M_{O_2}, M_{O_3}, \ldots\}$, such that $M_{O_1} = M_O \cup \{p_1\}$, $M_{O_2} = M_O \cup \{p_2\}$, and so on. Accordingly, $\cap\mathcal{M}_O = M_O$, and in cases where there is no altercentric ignorance or false beliefs, $\cap\mathcal{M}_O = M_S$. Of course, when this model is actually implemented in a representation of an agent's knowledge, it is likely to be simplified in a number of ways. For example, the size of P will presumably be restricted by the number of propositions that you represent as likely candidates for expansions of the agent's knowledge (and which are relevant to the question at hand). Critically though, one feature that will not changes is that representing an agent as knowing more than you will involve representing the agent's understanding of the situation as a map of maps, $\mathcal{M}_\mathcal{O}$, rather than some particular single map, M_O.

of egocentric ignorance, you once again cannot represent what the agent's false belief actually is, since you don't know where they hid the banana originally. Instead, you have to represent a map of the different possible non-factive maps the agent might have, and attribute that map of maps to the agent.

Unlike the distinction between attributing true and false beliefs, the distinction between attributing a single map and a map of maps remains almost completely un-studied. And yet, once we've seen why the focus on false beliefs was misplaced from the beginning, it should also be apparent that this distinction is at least as fundamental as the one between factive and non-factive theory of mind. In fact, much like the representation of false belief, the representation of egocentric ignorance guarantees that the subject is not making predictions based on their own understanding of the world. In both cases, these are important distinctions in the different kinds of representations one is able to maintain and update, and consequently they are distinctions we can make within a more general capacity for tracking others' representations of the world and keeping them separate from one's one. That is, they are both distinctions one can make within a more general capacity for theory of mind, but what they are not, are distinctions that concern whether or not one has a theory of mind.

With this theoretical foundation on the table, what remains to be spelled out is how we should move forward empirically. We think the right way to start is by setting all of these distinctions to one side, and developing a way to test for the core abilities of theory of mind: the capacity to track another's understanding of the world and keep it separate from your own.

In the next section, we lay out what we think the minimal version of such a test looks like. With this test in hand, we then return to these distinctions and show that it's easy enough to extend the test in ways that allow for it to distinguish between the different kinds of maps one can construct and attribute to others.

4.7 How to test for theory of mind

As we laid out at the beginning of the paper, the current state of things is that the only universally-accepted test for theory of mind is the false belief task. However, given that the false belief task tests not only for an ability for theory of mind, but also for an ability to represent a certain type of content, we want to propose an alternative task, which tests for the core abilities of theory of mind without also making the representation of a particular kind of (non-factive) content essential. Since it seems to be helpful to give these things a name, we'll call it the *diverse-knowledge task*, for reasons that will become obvious. As we discuss in § 5.3, there are many different ways of operationalizing this test in different experimental paradigms, but we'll illustrate the proposal by focusing on one simple paradigm, which we hope will help to clarify how to test for theory of mind without relying on false beliefs.

Given that demonstrating a genuine ability for theory of mind requires, at a minimum, demonstrating a capacity for both tracking and separation, the diverse-knowledge task requires two responses from subjects: one which provide evidence that they are tracking the other agent's understanding of the situation, and another which provides evidence that they are keeping it

separate from their own representation of the situation.

One way to implement this is illustrated in Figure 1. First, subjects are introduced to two empty boxes, one of which is the more likely location for the desired object (in this case, a banana). Perhaps the more likely location is the kind of place the desired object would usually be, or perhaps it has some other indication that it contains a desired object. After subjects are shown that both locations do not contain the desired object, the desired object is placed in the less likely of the two locations (in 1, the unmarked box). With this setup, two tests are then required. The first is one that tests whether the subject predicts that other agents will approach the location that is more likely to have the desired object; the second is one that asks whether the subject herself will go to the location which actually has the desired object.

Passing this or a similar version of the diverse-knowledge task would provide prima facie evidence for tracking and separation, and thus for a capacity for theory of mind. Of course, just as with other experiments, we may also want additional conditions that ensure subjects' responses are being driven by the hypothesized representations and not something that covaries with them. One obvious way to augment this test would be to also include a condition in which the agent is present when the desired object is placed in the less likely location. If subjects are *tracking* the agent's perspective, then they should now predict that the agent will go to location that actually contains the desired object, ruling out alternative explanations for predicting that the agent will go to the more likely location (e.g., a simple preference for, or habit of, going to that location).

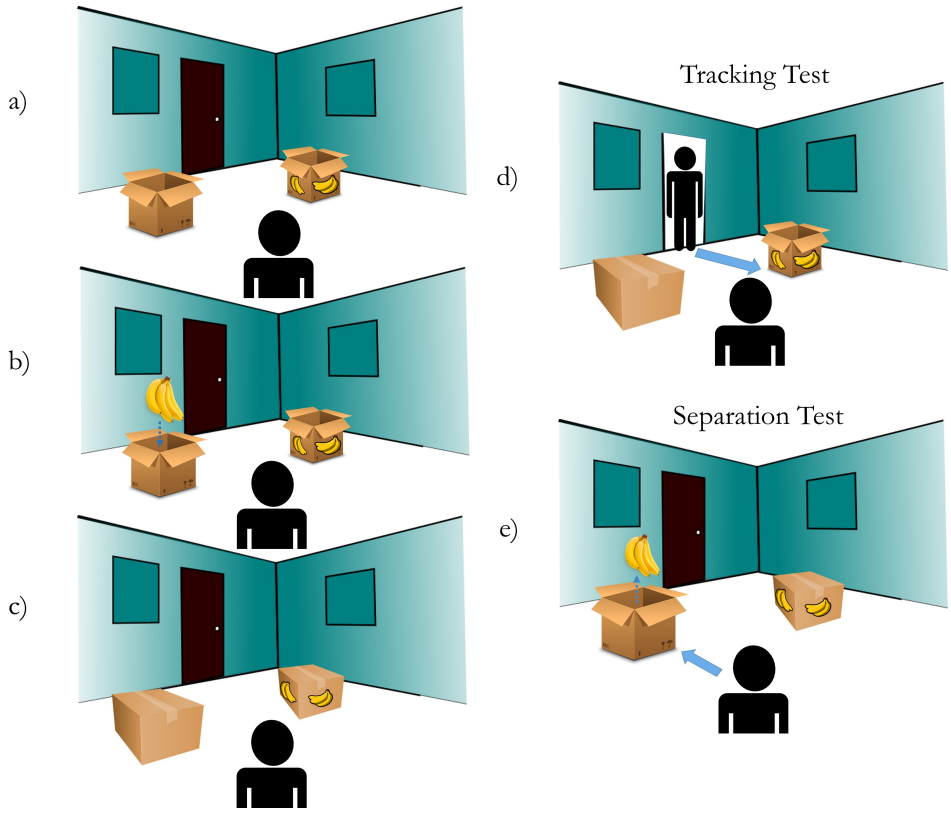

Figure 1: Schematic of a diverse-knowledge test

It may be helpful to contrast the key difference between this task and the original false belief task. The false belief task was proposed as a way of testing whether representations of other agents' beliefs were separate from subjects' representations of the world (Dennett, 1978; Bennett, 1978; Harman, 1978; Pylyshyn, 1978). We've proposed an alternative way of satisfying this criterion without requiring the representation of non-factive content. The trade-off is this: while the false belief task requires evidence for a single complex representation (i.e., a non-factive representation with content that is inconsistent with the agent's own beliefs), we simply require that there be

two non-identical representations of the world, one that the subject uses to guide her own actions and one that the subject uses to predict the actions of others.

On the one hand, the diverse-knowledge task demands more from theory-of-mind researchers, since they are required to show the agent simultaneously has access to two non-identical representations, which are differentially used to predict others' actions and guide subjects' own behavior. On the other hand, however, this test demands less from subjects in these studies, since they are no longer required to have a single complex representation which satisfies both criteria at once. We think this trade-off is one that researchers should be willing to make, as it allows for a much more sensitive test for the capacity for theory of mind, without introducing unrelated constraints on the kind of content of the theory of mind representation.

Once we have the most basic form of this paradigm in hand, though, it's not hard to see how it can be extended to test for the other distinctions we've discussed. First, consider the distinction between altercentric and egocentric ignorance. The previous test is sufficient for establishing the capacity to represent altercentric ignorance of factive content (which is sufficient for factive theory of mind). To test for the additional capacity to represent *egocentric* ignorance of factive content, one would simply invert the initial setup. The subject should not be able to see where the desired object is placed, but instead only sees that the other agent can see where it is placed. If the subject is able to represent altercentric ignorance of factive content, they should not be surprised when the agent retrieves the desired object from an unlikely location (or any location), but the subject herself should

be surprised by finding the desired object in an unlikely location. Passing this task would provide prima facie evidence for the capacity to represent egocentric ignorance.

The paradigm can obviously also be extend to test for the capacity to represent non-factive content by having the agent be in the room during the first placement of the desired object, but out of the room when the location of the object is switched. This is the equivalent of a standard false belief task. If the subject has the capacity for non-factive theory of mind, then she should be able to predict that the agent will look for the desired object in the location where it was originally placed, while the subject herself will look for the object in its actual location.

And finally, to test if the subject is able to represent egocentric ignorance of non-factive content, one would first have the subject see only that the agent can see where the desired object placed. Then, after the agent has left the room, the object would be removed from that location, and the subject would now be able to observe the new location where it is placed. If the subject has the capacity to represent egocentric ignorance of non-factive content, then she should be surprised if the agent initially searches for the object where it actually is, but should not be surprised if the agent looks for the object in any other location. She herself, should obviously just search for the object in its actual location.

While these ways of extending the diverse-knowledge task may help determine the kind of content that subjects can represent in theory of mind tasks, it bears reemphasizing that these are not tests of whether or not the subject has a genuine capacity for theory of mind. For that, all one needs

is the capacity to track others' understanding of the situation and to keep that separate from your own understanding of the situation. Evidence for that capacity only requires passing the diverse knowledge task itself.

5 Reconsidering the evidence for theory of mind

While the previous sections have argued that the false belief task is poorly suited to discovering whether a particular species or subject has theory of mind, the vast majority of research to date has been conducted with this standard in mind. As such, it is worth returning to the existing theory of mind research with a clearer account of what the core capacity for theory of mind actually is. With these new standards in mind, we can reexamine where it is that we do (or do not) find a genuine capacity for theory of mind.

While we go on to argue that the application of the diverse-knowledge task provides reason for thinking that we have overlooked important instances of genuine theory of mind ability, it's important to note at the outset that this conclusion was not arrived at by simply relaxing the standards for demonstrating an ability for theory of mind. As we go on to show, rather than requiring less evidence, the proposal we are making requires different (and in some ways stronger) evidence to demonstrate a genuine ability for theory of mind. As a result, there are good reasons to doubt that some widely accepted findings actually provide evidence for a genuine capacity for theory of mind.

5.1 When and how to decide that some behavior is *not* evidence for ToM

Let's begin by admitting that there are many instances of behaviors that appear to involve theory of mind but which can be explained much more simply without any appeal to its involvement. This is an important consequence of our view. Theory of mind is a relatively rare achievement; it requires both *tracking* another agent's understanding of the world, and keeping it *separate* from your own, so we shouldn't be too quick to see it where it may not be.

Here's one example. A species of a genus of spider, *Portia*, use "aggressive mimicry" in order to deceive and prey upon other spiders. The way they do this is by going on to the edge of the prey spider's web and then trying out a series of vibration patterns. They work through different patterns until they reach one that elicits a response from the prey spider, whereupon *Portia* continues to repeat the successful vibration. Generally, the vibration mimics either the vibrations of a trapped insect or of a potential mate. As a result, *Portia* "calls" the prey-spider closer and closer, until it is less than a body-length away, at which point *Portia* will attack and eat their prey (Jackson and Wilcox, 1993). They also use wind to "smokescreen" their approaches on to prey spiders' webs, and plan circuitous routes that allow them to drop into a web from above in order to attack spiders that they must approach from behind in order to subdue (Jackson and Wilcox, 1998). Members of *Portia* also sometimes use a "monotonous web signal that keeps the victim calm" while it gets close enough to strike.

All of this is behavior that looks like intentional trickery. It's not difficult

to take the intentional stance in this case, and explain what the spider is doing by saying that it's trying to make the prey-spider think that there's a prey-insect or mate in the web and that it starts out by looking for a vibration that will induce this false belief in the prey spider. Indeed, we might even take this to show that *Portia* can pass a false-belief task: it expects the prey spider to search where it falsely believes a trapped insect to be, and not elsewhere.

Of course, most researchers do not take this to demonstrate theory of mind ability, but some may be tempted to think that a view like ours might be unable to avoid counting this as an instance of theory of mind, so it's worth explaining why it does not. The central problem is that there simply isn't any evidence that *Portia* is actually tracking the other spider's understanding of the situation. When *Portia* goes after a victim, it starts out by generating vibrations from a pre-set repertoire and then, when those do not work, generating a wide array of new vibrations until it finds one that works (Jackson and Wilcox, 1998). It zeros in on a vibration pattern by responding to simple feedback from the prey-spider — specifically, movements from the prey spider that appear to be contingent on one of the *Portia*-induced vibrations (Jackson and Wilcox, 1993). Thus, they are actually following a rule such as: "Cycle through the repertoire of vibration patterns. When there's a response, repeat the most recent vibration pattern until prey is within reach." The rules required to describe the actual range and flexibility of *Portia*'s aggressive mimicry are much more complicated than this, but the representational resources required (i.e., what it has to represent) are not.

In short, the lesson here should be familiar one. Just like any other area of cognition (e.g., work on number representation in infants, Carey 2009; Feigenson et al. 2004), when doing theory of mind research, one has to show that what is being tracked is the property in question and not something that simply covaries with it. In the case of *Portia*, careful work revealed that what was tracked was something that covaried with prey-spiders' understanding of the situation (the prey-spider's movement), and not their actual understanding of the situation. Thus, thus this work does not demonstrate that *Portia* have a genuine capacity for theory of mind, and our account which emphasizes tracking, provides a principled way of understanding why it does not.

5.2 Where and how existing theory of mind research falls short

The difference in the requirements proposed by the diverse-knowledge task and the false-belief task can be demonstrated by reexamining a few examples of the research that has previously been taken to demonstrate an ability for theory of mind. We will review two different groups of research which, while clearly suggestive, have not yet provided the evidence required by the diverse-knowledge test. To be clear, the argument here is not that participants in these studies were not employing theory of mind, rather the argument is that these studies are missing a critical piece of evidence required to *demonstrate* that this is really the case. Accordingly, we both lay out precisely what these studies lack and also suggest relatively simple modifications of the experimental paradigms used which would allow them

to provide the requisite evidence needed to demonstrate theory of mind by passing the equivalent of a diverse-knowledge task.

5.2.1 Example 1: Misleading appearances

Demonstrating that a participant is tracking an agent's understanding of the situation simply requires evidence that the participant is sensitive to information that is represented therein. Demonstrating separation, however, requires more than this; it also requires evidence that this information is kept separate from the participant's own representation of the world. Such evidence has not always been provided.

Consider, for instance, the intriguing work on infants' reasoning about others' false perceptions. In one study (Song and Baillargeon, 2008), 14.5 month old infants were first familiarized with a female agent who demonstrated a preference for one toy (a toy doll with blue hair) over another (a skunk) by continually reaching for that toy when given a choice. Subsequently, when this agent was no longer in the room, the experimenter introduced two opaque boxes with lids. One of these boxes had a tuft of blue hair that resembled the doll's attached to the lid of the box (as the experimenter demonstrated). Then, while the agent was still out of the room, the skunk was placed inside the box with the tuft of blue hair, and the doll was placed inside the other box. The critical test was which box the infants expected the agent to reach for when she returned to the room, as measured by looking time. The infants looked longer when the agent reached for the box where the doll actually was (the box without the blue hair on the lid) suggesting that the infants expected the agent to act on the false belief that

the doll was in the box with the blue hair attached to the lid.

This evidence, while suggestive, is not sufficient to demonstrate the capacity for theory of mind because it does not demonstrate that the infants represented the agent's understanding of the situation as separate from their own. The easiest way to see this is to realize that it is possible that after the boxes were closed, the infants themselves represented the doll as being in the box with the blue hair attached to the lid. If this were the case, then infants would not need to track the agent's understanding of the situation to predict that the agent would reach for the box with the blue hair attached to the lid; this expectation could have been derived by merely referencing their own understanding of the situation. Of course, it would not be difficult to modify the paradigm to demonstrate that these two representations were kept separate. All that would be required would be additional conditions to demonstrate that infants' own expectations are violated when the doll is taken from the box with the blue hair than when the skunk is taken from the same box.

Abstracting away from the details of this particular experimental paradigm, the general lesson is that the diverse-knowledge task puts a strong emphasis on the *separateness* of the representations, and evidence for this separateness cannot be had without directly testing the participant's own representation. When it comes to the core features of a genuine theory of mind, the participant's own perspective is just as important as the agent's. We take this to be one of the key lessons that can be drawn from this proposal for how to rethink what the capacity for theory of mind actually is.

5.2.2 Example 2: Automatic theory of mind

A separate line of research has focused on testing for automatic theory of mind by using participants' response times (or error rates) as a measure of the extent to which they are tracking another agent's beliefs. According to the criteria we've laid out for the core capacity for theory of mind, the evidence provided by these tasks is also insufficient. The issue can again be most easily seen by considering one of the paradigms used in this research.

Consider the impressive body of work that has demonstrated automatic Level-1 perspective-taking (Samson et al. 2010; Surtees and Apperly 2012; Surtees et al. 2016b,a, see Apperly 2010 for review). In one of the most well-known studies, participants were shown a room with an avatar standing in the middle, facing one of two walls (Samson et al., 2010). On each trial, a number of dots could be displayed both on the wall that the avatar could see and on the wall that was behind the avatar (which the avatar couldn't see). On half of these trials, participants' task was to indicate how many dots the avatar could see, while on the other half of the trials, participants task was simply to indicate the number of dots that they themselves saw. The striking finding was that when participants were asked to simply report the number of dots they could see, they were slower (and made more errors) if the avatar in the scene saw fewer dots than they did. That is, even though the avatar was strictly irrelevant to the question at hand, participants seemed to be automatically representing how many dots the avatar saw, and this representation was interfering with their judgment of what they themselves saw.

The evidence provided by this line of research (and other research employing similar methods) is not well-suited to demonstrate a capacity for theory of mind because it is not designed to provide evidence for the separation between the participants' representation of the situation and the representation attributed to the avatar. In fact, what is most striking about this research is that it demonstrates a *lack* of separation between the avatar's representation of the ball's location and the participants'. If this weren't the case, then how would participants' own responses be affected by what the avatar saw? In fact, further evidence for a lack of separation between participants' own representations and that of other avatar comes from the fact that participants also make *egocentric* errors. That is, on trials in which participants were asked to indicate how many dots the avatar saw, they were also slower and made more errors when they themselves saw more dots than the avatar did.

This critique naturally extends to many other research paradigms that have investigated whether participants' own responses are influenced by the automatic representation of what others take to be the case (e.g., Kovács et al. 2010; van der Wel et al. 2014). Of course, our argument here is not that these research tasks do not involve participants engaging in theory of mind, but rather that this kind of paradigm does not good provide good evidence for it. To demonstrate a genuine capacity for theory of mind, it is critical not only to show that participants are sensitive to other agents' understanding of the situation but also that participants' own understanding is kept *separate* from other agents'.

5.3 Surprising places where we have missed evidence for theory of mind

Finally, we want to return to a couple of examples where we believe that studies have already provided provocative evidence that the observed behavior reveals a capacity for factive theory of mind by demonstrating both tracking and separation. In each example, further research may reveal this is not in fact the best explanation of the observed behavior. However, to the extent that alternative explanations are tested and rejected, the research discussed should be taken to provide clear evidence for the core abilities of a theory of mind.

5.3.1 Example 1: How helpful infants can be

As described earlier, infants as young as 12 months will help an adult find an "adult" object (e.g., a stapler) that has been hidden while the adult was gone, but not when the object was hidden in the presence of the adult (Liszkowski et al., 2006, 2008). Setting aside the debate over whether infants represent others' false beliefs (Onishi and Baillargeon, 2005; Southgate and Vernetti, 2014), it's worth pointing out that even these comparatively simple helping studies provide evidence of for a genuine theory of mind ability. Moreover, they do so without involving the representation of beliefs whatsoever. If infants did not represent the fact that the adult is ignorant of the location of the object (tracking the agent's understanding of the situation), and at the same time keep track of where that object really is (thus keeping their own representation of the situation separate), they would be unable

to succeed on this task. Accordingly, these studies functionally implement a version of a modified diverse-knowledge task.

5.3.2 Example 2: What non-human primates know you know

A second example of an extant paradigm that captures the essence of the diverse-knowledge task (even if not under that description), is the work on Rhesus macaque theory of mind described earlier (Santos et al., 2006). Since we've already described the study and explained how the monkeys' behavior requires the core abilities of a theory of mind in § 4, we won't describe it again here. Instead we'll just reemphasize that the most natural interpretation of the monkeys' behavior is that they were both tracking the experimenter's understanding of the situation and keeping it separate from their own.

In addition, similar studies have also been done with other primate species. For example, a study done by Hare et al. (2006) uses a conceptually similar experimental paradigm to study chimpanzees understanding of what other agents have visual access to. Like the Rhesus macaques, the chimpanzees were able to choose the action that would not alert the experimenter, even on their first attempt. This provides evidence for both a calculation of which parts of the situation the experimenter did not understand, and some way of keeping that understanding separate from their own. Moreover, using a slightly different paradigm, researchers found that chimpanzees chose actions that advantageously manipulated both the auditory and visual evidence the experimenter had access to, and that this behavioral pattern was directly contingent on what the experimenter could see or hear

(Melis et al., 2006), again providing evidence for tracking and separation.

It's worth pointing out that in each of these cases, great apes and even monkeys are succeeding at these tasks by exercising a capacity for theory of mind that is completely *factive*. What makes this success particularly interesting, is that there is currently no good evidence to date that monkeys have any capacity for non-factive representations (Martin and Santos, 2014, 2016), and only very mixed evidence in the case of great apes (Kaminski et al., 2008; Krachun et al., 2009; Krupenye et al., 2016). Yet, despite this lack of an ability for non-factive representations, these studies provide clear examples of a capacity for both tracking what another agent understands and keeping that representation separate from their own understanding of the world. In other words, these tasks seem to provide clear evidence for a genuine capacity for theory of mind in non-human primates.

Accordingly, an important upshot of our proposal is that, contrary to the prevailing view, there *is* evidence that some non-human primates have a genuine capacity for theory of mind (see Martin and Santos (2014) and Call and Tomasello (2008) for recent counterposing reviews). Of course, our claim is not that these monkeys understand what mental states are in any kind of complex or reflective way, or that they have the same concept of mental states that adult humans have. Rather, our claim is that there is good evidence that they can track others' representations of the world and keep those representations separate from their own. And if this isn't what is essential for theory of mind, then we are not sure what is.

6 Conclusion

Stepping back from the details of our discussion, it's not hard to see that a very different understanding of theory of mind has emerged. We began by laying out why the long-standing emphasis on false beliefs has led us down the wrong track: it makes the capacity for non-factive representations essential when testing for theory of mind, which confounds the kind of content that can be represented with the capacity to represent another agent's understanding of the world. In place of false beliefs, we've argued that a more principled set of criteria are simply (1) tracking (the same criteria used throughout cognitive science), and (2) the ability to keep this representation separate from one' own understanding of the world. With these criteria in hand, we developed the diverse-knowledge task. Critically, this task provides a way of testing for the comparatively simple capacity for *factive theory of mind*, while additionally allowing for simple modifications that make it possible to instead test for the ability to represent altercentric ignorance or non-factive content. Lastly, we turned to a number of existing examples of theory of mind research with this account of theory of mind in hand. We illustrated how to decide when some behavior is not theory of mind, how some existing research falls short of demonstrating the capacity for theory of mind, and how we've missed good evidence for theory of mind in a number of surprising places. We hope that this picture inspires a new generation of work on theory of mind.

7 Acknowledgements

Among many other people, we would like to thank Matthew Mandelkern, Jessie Munton, Alia Martin, Laurie Santos, Enoch Lambert, the Yale Cognitive Science Reading Group and the SHAME writing group.

References

Andrews, K. (2012). *Do apes read minds?* MIT Press.

Apperly, I. (2010). *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21:37–46.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, 1(4):557–60.

Butterfill, S. A. and Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, 28(5):606–637.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–92.

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

Carruthers, P. (2013). Mindreading in infancy. *Mind and Language*, 28(2):141–72.

Dehaene, S. and Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience*, 5(4):390–407.

Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1(4):568–70.

Drayton, L. A. and Santos, L. R. (2016). A decade of theory of mind research on cayo santiago: insights into rhesus macaque social cognition. *American journal of primatology*, 78(1):106–116.

Fabricius, W. V., Boyer, T. W., Weimer, A. A., and Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology*, 46(6):1402–16.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, 8(7):307–314.

Frith, C. D. and Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, 63:287–313.

Gallagher, H. L. and Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, 7(2):77–83.

Gallistel, C. and Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, 44:43–74.

Gergely, G. and Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–92.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.

Gopnik, A. and Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language*, 7:145–71.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, 1:158–71.

Grant, C. M., Riggs, K. J., and Boucher, J. (2004). Counterfactual and mental state reasoning in children with autism. *Journal of autism and developmental disorders*, 34(2):177–188.

Gweon, H., Dodell-Feder, D., Bedny, M., and Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child development*, 83(6):1853–1868.

Hare, B., Call, J., and Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, 101:495–514.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, 1(4):576–77.

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21(01):101–114.

Jackson, R. R. and Wilcox, R. S. (1993). Spider flexibly chooses aggressive mimicry signals for different prey by trial and error. *Behaviour*, 127(1/2):21–36.

Jackson, R. R. and Wilcox, R. S. (1998). Spider-eating spiders. *American Scientist*, 86(4):350–7.

Kaminski, J., Call, J., and Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109(2):224–234.

Kiparsky, P. and Kiparsky, C. (1970). Fact. In Bierwisch, M. and Heidolph, K. E., editors, *Progress in linguistics: a collection of papers*. Walter de Gruyter GmbH & Co. KG.

Koster-Hale, J. and Saxe, R. (2013). Functional neuroimaging of theory of mind. *Understanding Other Minds: Perspectives from developmental social neuroscience*, pages 132–163.

Kovács, A. M., Téglás, E., and Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330:1830–34.

Krachun, C., Carpenter, M., Call, J., and Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, 12(4):521–535.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives*, volume 36. Oxford University Press.

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308):110–114.

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind.". *Psychological review*, 94(4):412.

Liszkowski, U., Carpenter, M., Striano, T., and Tomasello, M. (2006). 12- and 18-month-olds point to provide information for others. *Journal of Cognition and Development*, 7(2):173–187.

Liszkowski, U., Carpenter, M., and Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, 108(3):732–739.

Lurz, R. W. (2011). *Mindreading animals: the debate over what animals know about other minds*. MIT Press.

Martin, A. and Santos, L. R. (2014). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. *Cognition*, 130:300–8.

Martin, A. and Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, 20(5):375–382.

McCrink, K. and Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, 18(8):740–5.

Melis, A. P., Call, J., and Tomasello, M. (2006). Chimpanzees (pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology*, 120(2):154.

Nichols, S. and Stich, S. (2003). *Mindreading: An integrated account of pre-*

tence, self-awareness, and understanding of other minds. Oxford: Oxford University Press.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308:255–8.

Penn, D. C. and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B*, 362:731–44.

Perner, J. and Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308(214):214–6.

Peterson, D. M. and Bowler, D. M. (2000). Counterfactual reasoning and false belief understanding in children with autism. *Autism*, 4(4):391–405.

Peterson, D. M. and Riggs, K. J. (1999). Adaptive modelling and mindreading. *Mind & Language*, 14(1):80–112.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–26.

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified?[p&w]. *Behavioral and brain sciences*, 1(04):592–593.

Rafetseder, E., Cristi-Vargas, R., and Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child development*, 81(1):376–389.

Riggs, K. J. and Peterson, D. M. (2000). Counterfactual thinking in preschool children: Mental state and causal inferences. *Children's reasoning and the mind*, pages 87–99.

Riggs, K. J., Peterson, D. M., Robinson, E. J., and Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, 13(1):73–90.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., and Scott, S. E. B. (2010). Seeing it their way: Evidence for rapid and involuntary computations of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5):1255–66.

Santos, L. R., Nissen, A. G., and Ferrugia, J. A. (2006). Rhesus monkeys, *Macaca mulatta*, know what others can and cannot hear. *Animal Behaviour*, 71(5):1175–1181.

Saxe, R. and Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4):1835–1842.

Scott, R. M. and Baillargeon, R. (2009). Which penguin is this? attributing false beliefs about object identity at 18 months. *Child Development*, 80(4):1172–96.

Song, H. and Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, 44(6):1789–95.

Southgate, V. and Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, 130:1–10.

Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18(7):580–86.

Surtees, A., Apperly, I., and Samson, D. (2016a). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition*, 150:43–52.

Surtees, A., Samson, D., and Apperly, I. (2016b). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, 148:97–105.

Surtees, A. D. and Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child Development*, 83(2):452–60.

Tan, J. and Harris, P. L. (1991). Autistic children understand seeing and wanting. *Development and psychopathology*, 3(02):163–174.

van der Wel, R. P., Sebanz, N., and Knoblich, G. (2014). Do people automatically track others' beliefs? evidence from a continuous measure. *Cognition*, 130(1):128–133.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–28.

Xu, F., Spelke, E. S., and Goddard, S. (2005). Number sense in human infants. *Developmental science*, 8(1):88–101.