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 ensure that subjects' responses cannot depend on their own representation of the world. This requirement confuses the ability to represent a particular kind of non-factive content (e.g., a false belief) with the more general capacity to represent others' understanding of the world even when it differs from one's own. We provide a way of correcting this. 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 see, hear, or know, provide evidence for a genuine theory of mind. We then show how these criteria can be operationalized in a new experimental paradigm: the 'diverse-knowledge task'. Finally, we illustrate how this account should reorient our understanding of the past and the future of theory of mind research.

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

Factive Theory of Mind

On an island in Peurto Rico, sometimes called '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 from the bells on one of the boxes, making one of the boxes silent and the other noisy as they were handled by the experimenter. The experimenter then 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 who attempted to steal food from the experimenter, twelve of them went to the box without ringers, avoiding alerting the experimenter. Other times 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 or a hole-puncher). Both 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. However, if the experimenter had seen the object move to a new location, infants did not preferentially point to the object, despite the experimenter's unsuccessful attempts at searching for an object (Liszkowski et al., 2008).

*

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. And when things become most interesting is if someone finds 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 very different subjects seem to keep track of something about what other agents know or understand and then use that information 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 informative. 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* (for classic examples of false belief tasks, see, Wimmer & Perner, 1983 or Baron-Cohen et al., 1985). In the view of received wisdom, the cases we started with may be thought of as suggestive or intriguing, but the idea that they are clear demonstrations of the essential abilities for a theory of mind is typically 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 subjects 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 backward 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 litmus test 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, forty years later, we find ourselves teaching introductory students that the litmus test for having (or developing) a capacity for theory of mind is the passing a false belief test, and we've collectively written well over 8,000 papers employing or discussing one version or another of this test.[1] To be clear, we're not saying that any of us think that false beliefs are the entirety of theory of mind; we're saying that most of us treat them as the most important way to *test* for theory of mind, and this seems to be just as true today as it was in 1978.

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 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 if they *recognize* the person you're pointing out, or whether they *remember* that one time when...). 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

[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 July, 2018.

[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

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 *seeing* 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 don't exist, 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 take the world to actually be.

We seem stuck. On the one hand, we know that there are a lot of phenomena that we all agree *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 forty years ago, we're wary that these kinds of *factive* theory of mind don't pass muster. They may not be good evidence for genuine theory of mind representations, since these representations are, by definition, tied to the way one actually takes the world to be. Then, 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 an essential 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. Demonstrating a genuine capacity for theory of mind has typically required demonstrating a competence with non-factive representations, and this has been the standard for how we demarcate, for example, which species actually have a genuine capacity for theory of mind

and Kiparsky, 1970).

(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 help to reorient 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 (barring non-mentalistic confounds, e.g., Heyes 2014a) some evidence that they have a capacity for theory of mind. But very little is known about the principles by which infants are able to pass such tests, 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 is 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? And combining the two, like hybrid theories do, obviously won't solve the problem either.

These theories 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).

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). 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. 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; 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

It's going to be helpful to start by setting aside theory of mind and instead focusing on a broader question: How do we discover whether an entity 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 dif-

ferent 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 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 that just happens to covary with number, then we can be reasonably satisfied that, in some way or another, infants are capable of *representing* number. 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, in neuroscience, and so on. The basic idea is that the

ability to respond differentially 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. This typically isn't taken to be the totality of representation, but it is the basis of it.[3]

Returning to research on theory of mind, though, what is often demanded is not so much a sensitivity to others' mental states, but instead a rather sophisticated understanding of those mental states. Consider the influential approach of thinking of the ability for theory of mind as roughly a matter of being able to take the intentional stance (Dennett, 1978). To take the intentional stance toward another system (whether that system is a chimpanzee or a thermostat) is roughly to interpret it as having beliefs and desires: Specifically, the beliefs that it would be rational to have given the evidence available to the system. And then to interpret it as acting in such a way that would rationally satisfy its desires (assuming that its beliefs are true).

On this picture, theory of mind straightforwardly requires a surprisingly broad understanding of beliefs, desires, rationality, and how all of these ideally fit together. Moreover, such an ability will only really come as a suite: you either have the ability to represent all of these mental states at a level of relative complexity or you simply won't be able take this sort of stance. After all, it wouldn't do one much good to take the intentional stance if one could only represent beliefs or only represent desires, or represented either beliefs or desires in a way that they didn't allow them to combined such that they are subject to the demands of rationality.

[3]On standard philosophical theories of mental representation, mere tracking is generally not enough for representation (Dretske 1988; Millikan 1989; Fodor 1990; see also, Gallistel and King (2009). In addition to a detection mechanism, the organism must be able to use the signal carried by the detection mechanism (Dretske, 1988)), which should serve a functionally appropriate role (Ramsey, 2007). But if it is possible to demonstrate that an animal has both of these elements, then we are in a decent position to say that the animal represents that feature.

On reflection though, we're not sure that all of these commitments are things that theory of mind researchers would really want to endorse. Suppose, for example, one applied an analogous standard in research on number. If the appropriate standard for representation is an understanding of the entity represented, then to show that infants are truly representing number and not something lesser, one will additionally need to show that they understand, e.g., that numbers, being numbers, are by definition the kinds of things that can be negative, so if any subject is *really* able to represent number, they should be able to represent negative numbers.

We think it's obvious that such an approach is wrong in the case of number. One does not need to be able to understand number at this level of complexity in order to be able to represent number in the first place. If that were the case, we're not sure whether most adult humans are capable of representing number. And we can't see why much the same thing isn't also true in the case of theory of mind. Requiring that one be able to represent the falseness of beliefs is like requiring that one be able to represent the negativeness of number. Sure, if you could do it, no one would doubt you could represent others' minds, but it'd hardly be a reasonable test of whether you could had the capacity for representing others' minds in the first place.

Our suggestion here is not meant to be radical or really even particularly original. 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 *core* of an ability for the-

ory of mind. With this alternative in hand, we want to start to ask what a picture of theory of mind might look like once we let go of false beliefs.

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 account of the capacity for theory of mind. The proposal has two key requirements. The first is that an organism be able to *track* the contents of another agent's 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—attributing the tracked representation to the other, while using one's own representation for action. 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, around which more sophisticated capacities are built.

The first condition of the theory—that mindreaders be able to track others' representations of the world—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 other agents' understanding of the situation is the thing that needs to be tracked if one is to represent that perspective. 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 agent has theory of mind ability is to find out whether they are sensitive to changes in others' representations of the world.

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 *attributing* 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* understand the world, not on what I think the world really is like. So if I simply attribute my understanding to you (or confuse your understanding 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.

In other words, 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 track others' understanding of the world, but also keep those representations *separate* from their own understanding. Theory of mind, then, is a matter of both tracking and separation. Tracking demonstrates that you have a representation of another agent's understanding. Separation demonstrates that this representation plays the functional role that is appropriate for a theory of mind representation, and is not, e.g., simply 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 for, 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 their understanding of what the world is like. 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. You could 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 agents' 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 contradicts 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 stretch the separate maps analogy, we can imagine that you can construct altercentric maps by 'removing' parts of your own map 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. Supposing that the person is still looking for bananas, you'd have the capacity, for example, to realize

that the person won't immediately look for the banana in the box. Critically though, this sort of ability requires two functionally separate maps – one for predicting what the other person will do, and a second one that you base your own behavior on.

4.2 What hangs on this?

An important consequence of this proposal is that a genuine 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 *beliefs*. If the core capacities of theory of mind are tracking and separation, then the representation and attribution of factive attitudes (another's *knowing* something or *not knowing* something) is sufficient for theory of mind. All that is required is a demonstration that these factive representations both *track* another agent's understanding and are kept *separate* from your 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 for a genuine capacity for theory of mind. One easy way to make this difference clear is to apply it to one of the controversial cases with which we 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 applying some simpler behavioral rule or associative connection that would link silent boxes with being able to retrieve food.[4]

What representations do the monkeys need in order to succeed in 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 the monkey realizes that its approach will not be part of the competitor's understanding of the world. That is, only if the monkey realizes that the competitor's representation of things will not include its taking of the food.

Thus, among other things, the monkey must be able to *track* what things are like from the competitor's perspective. Moreover, the animal's own map of the world must be kept separate from this representation of the competitor's understanding: even though the monkey's picture will include the approach and retrieval attempt, this representation must be separately maintained, because the monkey needs to simultaneously know both that it is approaching the silent box and that the experimenter doesn't realize that this is the case.

In other words, an obvious interpretation of success on this task is that it requires both *tracking* and *separation*. It requires maintaining a representation of what things are like from the competitor's perspective, and using that representation to predict the competitor's behavior while not treat-

[4]See Santos et al. (2006) and Martin and Santos (2016) for arguments against explanations based on simple behavioral rules. We also argue against such alternative explanations in detail in the supplement to this paper, available here: `https://psyarxiv.com/83zhj`.

ing that representation as a complete picture of the world, i.e., keeping it functionally separate from the way the monkey takes the world to actually be.

Importantly, the ability that we've argued these monkeys seem to be exhibiting in this experiment—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 (i.e., attributing to them a map that does not include the banana). 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. Factive representations, like *knowing*, won't help either. Your own map doesn't include the banana being 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 (attributing to them a map that includes the banana being in the new location), 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 map that strictly 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'll look for the banana.

Given how much we think hangs on this, it's going to be worth getting clear on exactly what the difference between factive and non-factive theory of mind boils down to (and how they both differ from not having a capacity for theory of mind at all). In a certain light, factive and non-factive theory of mind can seem quite similar. In the former, you're systematically tracking what the person does *not* represent as being the case; in the latter, 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 important (and missed) distinctions will begin to come into focus.

5 Theory of Mind

Going forward, it's going to be important to be a little more precise about the proposal we're making. We'll generally stick to a high-level description of the proposal, but alongside that, we're also going to provide a slightly more formal way of understanding what we're proposing. We think that being precise enough to get these details right turns out to be important, so we'll keep these formal details nearby for anyone who wants to see how all this

gets worked out. For the most part though, all of these details can pretty safely be kept in the background if you're just interested in understanding our account and its implications for theory of mind at a broad conceptual level.

Start with the basics. Take 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 depend 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 all 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 includes 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.

With this 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 to reflect changes in the way that this agent understands the situation. *Separation* additionally requires that the other agent's map is maintained and updated independently of your own map: one must be able to add or remove propositions from one's own map in a way that does not demand a corresponding change to the altercentric map, and vice versa. With just these few pieces, we 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.

5.1 No theory of mind vs. 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.[5] 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.

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 may 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 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]

[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 $\forall p : p \in M_O \iff p \in M_S$.

[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 $\forall p : p \in M_O$, $\neg p \notin M_S$.

5.1.1 What the difference is not

The distinction we are pointing to will likely become clearer when compared to an important recent suggestion by Martin and Santos (2016). Martin and Santos argued that the best way to make sense of non-human primates' successes on a variety of theory of mind tasks is that they are able to track other agents' *awareness* of the situation. Stated in the terms we've been using, their suggestion is that non-human primates have a single map of the situation and can keep track of whether other agents are aware of *that* situation—only using that map to predict their actions in cases where they are aware.[7] As Martin and Santos suggest, this capacity would not actually allow non-human primates to represent knowledge or ignorance *per se*, but it would still allow them to go some way in making sophisticated predictions about other agents (Martin and Santos, 2016, pp. 380-381). In cases where another agent is aware of the situation, one can treat the other as having one's own map, and in cases where the agent is not aware, one simply has no predictions about what the other agent will do (or just relies on one's priors over potential behaviors in that situation). On our view, this proposal suggests that non-human primates have half of a theory of mind: they can track whether or not an agent is aware of a situation, but they only have a single map of the world, and so they definitely can't keep the other agent's

[7] While Martin and Santos did not formalize their proposal, the most straightforward interpretation would be that non-human primates have some way of tracking whether the agent has awareness of the relevant situation, and if that condition is met, they attribute their entire representation of that situation to the agent; otherwise, they do not attribute any understanding of the situation to the agent. That is:

$$M_O = \begin{cases} M_S & \text{If other is aware} \\ \varnothing & \text{Else} \end{cases}$$

Obviously, in cases where the other agent is aware of the situation, our proposed characterization of not having theory of mind, $\forall p : p \in M_O \iff p \in M_S$, holds; in cases where the agent is not aware, there is no representation, M_O.

map *separate* from their own. This might be a particularly clever way of conditionally predicting others' actions from a single representation of the world, but it's not genuine theory of mind.

An important thing that this comparison should help make clear is that we don't think theory of mind is a matter of whether or not you are able to accurately predict others' behavior based on the situation they are in. And it doesn't change things if you make these predictions contingently on whether you take someone to share your understanding of the situation. You've still only got one representation of the world, and a genuine capacity for theory of mind requires more than this.

5.1.2 What the difference is

But if this isn't enough for theory of mind, then what is? Suppose that what one does is not track just *whether* another agent is aware of the situation, but instead tracks *which* parts of the situation another agent is aware of and which parts the agent is not aware of. Stretching the maps analogy further, suppose that you can do this by starting with your own map and simply removing all of the parts of the map that the agent isn't aware of (or is mistaken about), leaving you with just a bunch of pieces of your original map. Further, suppose that you then use only these pieces to predict what the other agent will do but not the rest of the map. As long as you don't lose your own map of the situation in constructing this map for the other agent, then as far as we're concerned, you'll end up with two *functionally distinct* maps. Here's why: take a case where you remove pieces from your own map and then predict the other agent's behavior based only on the remaining parts. Great—but now, how do you decide what you yourself should do? If, in making your own decisions, you rely on information that was not part

of the other agent's map, then you must have retained the rest of your own map in one way or another. How else could you have a more complete representation of the situation to base your own actions on? Functionally speaking then, what you've done is track another agent's understanding of the situation and keep that representation separate from your own, and thus, we think this is a clear demonstration of the core capacities of theory of mind: tracking and separation. Together, these two capacities give you the ability to make predictions of others based on a different understanding of the situation than the one that you yourself use.

5.1.3 Where and why this difference matters

To see where the distinction we're pointing to matters, let's return to our boxes of bananas. Suppose that I saw you put a banana in one of the boxes (Box 1), but then when I wasn't looking, you also put another banana in the other box (Box 2). If you have only the abilities suggested by Martin and Santos' proposal (2016), then you have two possible approaches to predicting where I would will look for a banana. On the one hand, you might no longer represent me as being aware of the situation, in which case, you should be at chance in predicting which box I would take a banana from, since you shouldn't attribute to me any representation of there being bananas in boxes. Alternatively, you might instead represent me as being aware of the entire situation, i.e., having your own understanding of the situation. But if that's the case, you should also predict that I might go to either box to retrieve a banana, since I'll understand that both boxes have bananas in them.[8] What you can't do, however, is simultaneously know that there is

[8]Let q_1 be the proposition that there is a banana in Box 1 and q_2 be the proposition that there is a banana in Box 2. If you take me to be unaware of the situation, then $M_O = \varnothing$, and trivially, $q_1 \notin M_O$, so you can't predict my actions based on my knowing q_1. If you take me to be aware of the situation, then $M_O = M_S$, and then obviously, if

a banana in both boxes but predict that I will retrieve the banana from the Box 1 but not from Box 2.[9] Being able to do that would require, at a minimum, a capacity for *factive* theory of mind.

5.2 Factive vs. non-factive theory of mind

We've said something about the difference between not having a genuine capacity for theory of mind, and having a capacity for factive theory of mind, but even if you have a capacity for factive theory of mind, there will still be things you cannot do. While *factive* theory of mind would allow you to construct and update an altercentric map that is *not identical* to your own, it will not allow for your own map to be *inconsistent* with another's map, so you still couldn't represent (false) beliefs. After all, as long as your understanding of the situation itself does not involve contradictions, then obviously no subset of the things you take to be true of the situation will be inconsistent with your own 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

$q_2 \in M_S$, then $q_2 \in M_O$ so you also can't predict my behavior based on my *not* knowing q_2.

[9] It may initially be tempting to object by arguing that the proposal of Martin and Santos (2016) need not be so all-or-nothing, and that a more charitable interpretation would be that their proposal allows one to represent others as being aware of some situations, while not being aware of others (i.e., their awareness representation operates over situations rather than entire maps). The trouble with this response is that this version of Martin and Santos's proposal is equivalent to our own, and ours straightforwardly allows for representations of knowledge and ignorance *per se*, which Martin and Santos explicitly deny their proposal allows for (Martin and Santos, 2016, pp. 380-381). It shouldn't be too hard to see why this must be the case. Suppose an agent is not aware of one situation, S_1, but remains aware of other situations, $\{S_2, S_3, ...S_n\}$. To *not* be aware of a given situation, S_1 is just to not be aware of the set of the facts that make up that situation, i.e., $\forall p : p \in S_1, p \notin M_O$. Similarly, to *be aware* of a situation is just to be aware of the facts that make up that situation, i.e., $\forall p : p \in S_2, p \in M_O$. Thus, in cases where the agent is represented as being unaware of S_1, but aware of S_2, then straightforwardly, $S_O \neq \varnothing$ and $\exists p : p \in M_S \wedge p \notin M_O$, which is a more precise way of saying that one has an ability for factive theory of mind—an ability to represent knowledge and ignorance *per se*.

to represent someone as representing that thing *not being the case.*[10]

To be able to do that, one must have the capacity for *non-factive theory* of mind.[11] 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 propositions your map contains). The upshot of having non-factive theory of mind is that you can do something more than understanding that the another person does not represent something: you can now understand that they take that thing to positively *not* be the case.[12]

With this more precise way of thinking about factive and non-factive theory of mind in hand, it should be 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 a matter of whether one can construct and maintain a particular kind of representation: a non-factive representation. That is, one that is *inconsistent* with the way you take the world to actually be.

5.2.1 What the difference is not

When seen for what it is, the difference between factive and non-factive theory of mind is simply not a difference of *whether or not* one represents

[10] 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 $\forall p : p \in M_O$, $\neg p \notin M_S$.

[11] See Nagel (2017) for a complementary discussion of factive mental states in theory of mind that argues that factive theory of mind better explains many aspects of infants' and primates' successes on theory of mind tasks.

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

others' understanding of the world. That is, it is not at heart a difference in theory of mind. Rather, it's a difference that concerns what kind of content one has the ability to represent in a perfectly general way.

One 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-mental 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 separate representation that is inconsistent with your own understanding of the world.[13]

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.[14] 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 (adding p to a hypothetical map) while simply not

[13]See Byrne (2017) for a review of the empirical evidence for this distinction in online processing.

[14]See Kratzer (2012); Peterson and Bowler (2000); Grant et al. (2004) for related discussions.

having any belief as to whether or not p is actually 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.

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. And 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 be so surprising that young children who cannot pass simple verbal counterfactual reasoning tasks also cannot pass the verbal false 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)[15] 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

[15]See Leahy et al. (2014) for a helpful formal way of accounting for the developmental changes in the shift from hypothetical to counterfactual reasoning in Lewis/Stalnaker counterfactual logic.

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, so any test that makes it essential would not be a very good test of theory of mind.

5.3 Altercentric vs. egocentric ignorance

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').[16] 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.[17]

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. Representations of egocentric ignorance are necessarily more complex than rep-

[16] 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 $\exists p : p \in M_S \wedge p \notin M_O$.

[17] In natural language, this is kind of mental state is often ascribed to others using an embedded question under a factive attitude, and is sometimes referred to as 'knowledge-*wh*'.

resentations 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. If that's right, though, then how do you do it? The solution to this problem requires that you instead construct and attribute a more complex kind of representation that involves a set of maps.

To see why, just 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. The reason you would not be surprised by this is that you understood that the agent's map included a banana behind the door. Attributing such a 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 a box 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. 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. 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.[18]

The basic insight is that in cases of 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.[19] 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.

Unlike the distinction between attributing true and false beliefs, the distinction between attributing a single map and a set 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

[18]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 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, $\cap\mathcal{M}_O = M_S$.

[19]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, $\mathcal{M}_O$, rather than some particular single map, M_O.

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 own. 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.

5.4 Moving forward

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 first 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.

6 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 litmus test for theory of mind is passing 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 non-factive content, we want to propose an alternative task, which more precisely tests for the

core abilities of theory of mind without also making the representation of a particular kind of 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 if they aren't already. 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 easy way to implement this is with the kind of situation illustrated in Figure 1. The subject in the experiment is situated such that she can see into two rooms, Room 1 and Room 2, each of which have two empty boxes. The other agent in the experiment is instead placed such that she can see what happens in Room 2 but not Room 1. With this setup in place, a banana is placed in one of the two boxes in Room 1 (in view of only the subject), and another banana is placed in one of those two boxes in Room 2 (in view of both the subject and the agent), and all of the boxes are closed. Two tests are then required. The first (Fig. 2, left) is one that tests whether the subject correctly tracks the agent's understanding of the situation. If she does, then when the agent is in Room 2, she should predict that the agent will look for the bananas in the box where they actually are. At the same time though, when the agent is in Room 1, the subject should be at

chance in her predictions of where the agent will look for the bananas. The second test (Fig. 2, right) instead asks whether the agent has kept her own representation separate. If she has, she herself should be equally willing to retrieve the bananas from the box in which they were placed in either room. Obviously, it's not possible for a single map to give rise to both different responses, and thus succeeding requires both tracking and separation.

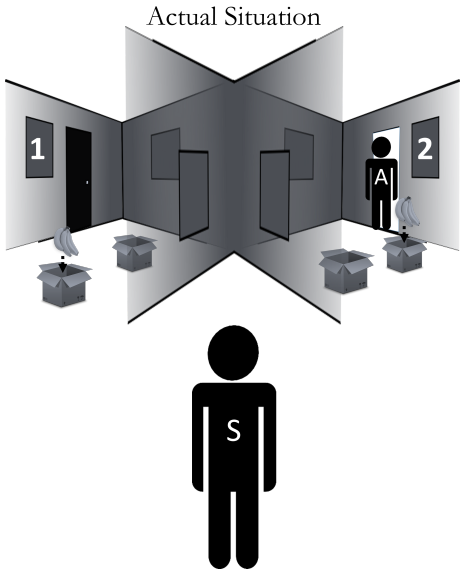

Figure 1: Setup of the diverse-knowledge task

It may be helpful to compare 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 rep-

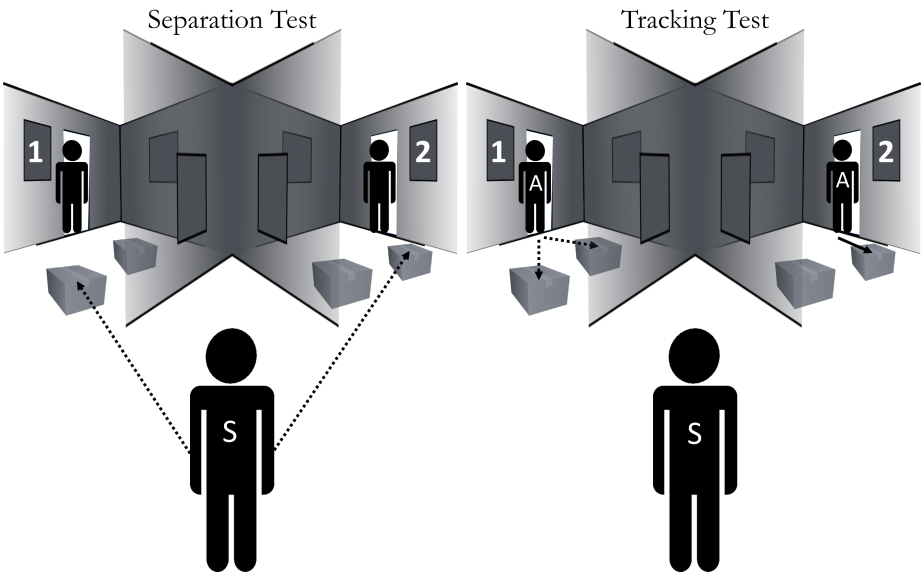

Figure 2: Schematic depiction the Separation and Tracking conditions of the diverse-knowledge task. Dotted arrows indicate the equal preference that the agent or subject should have between multiple locations; solid arrows represent the agent's unequal preference for one location over another.

resentations 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.

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. For example, 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 only be able to see that the agent can see where the object is placed, but not where the object itself is placed. If the subject is able to represent altercentric ignorance of factive content, the subject should not be surprised when the agent retrieves the desired object

from any location in the room the subject couldn't see into. Passing this task would provide prima facie evidence for the capacity to represent egocentric ignorance.

The paradigm can 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. And of course, it's not hard to combine non-factive content with egocentric ignorance either. 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 (different) location where it is placed.

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 another's 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.

6.1 That's it?

The short answer is: Yes, that's it.

The long answer is that throughout we've clearly been making the idealizing assumption that participants' behavior in any particular implementation of these tasks is not better explained by some behavioral rule (or any other nonmentalistic feature) that happens to covary with the agent's understanding of the situation. What's essential for any version of these ex-

periments is to detect whether and to what extent the subject is tracking another agent's understanding of the world and not anything else that happens to be confounded with that understanding (as laid out in §3). As great as we think the diverse knowledge task is, it's not magic, and ruling out these kinds of confounds is just as critical for the diverse knowledge task as it is for any other theory of mind task, including, of course, the false belief task (see, e.g., Heyes, 2014b; Perner and Ruffman, 2005; Povinelli and Vonk, 2004; Ruffman, 2014).[20]

7 Looking backward

From the beginning, our aim has not simply been to develop a new test for theory of mind. Rather, it has been to offer a way for us all to let go of false beliefs in hopes of refocusing on what's essential in theory of mind. While we've sketched one simple way to move forward by testing for only the essential parts of an ability for theory of mind (§6), it's also worth taking a moment to look backward. The past forty years of theory of mind research have, by and large, been conducted with the assumption that false belief understanding is what is critical for demonstrating genuine theory of mind. Yet, if tracking and separation are what is actually essential, then the literature from the past forty years should begin to take on a very different light.

In some cases, research that has been taken to provide clear evidence for theory of mind because it involves false beliefs should start to look more lackluster. To illustrate with one example, consider the various paradigms

[20] In §Section I of the supplement to this article, we expand on why we think the approach we've argued for is a better way of testing for theory of mind from both a theoretical and methodological perspective. The supplement is available here: `https://psyarxiv.com/83zhj`

that have sought to demonstrate theory of mind by showing that the calculation of other agents' perspectives interferes with participants' own responses (see, e.g., Kovács et al., 2010; Apperly, 2010; van der Wel et al., 2014). While these findings are clearly intriguing, they are not well-suited to provide evidence for theory of mind as we've argued it should be understood. Here's the reason: to show an effect, these paradigms require that participants *fail* to keep other agent's understanding of the world separate from their own. And if all of the evidence is evidence of a *lack* of separation, then it is, *ipso facto*, not good evidence for what is essential for having theory of mind.

In other cases though, research that has not been seen as providing good evidence for theory of mind (because it hasn't involved false beliefs) should look more promising. To return to one example, 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). Or, to return to another, rhesus monkeys show intriguing evidence of being able to track others' understanding of the world and keep it separate from their own when stealing food (Santos et al., 2006). Similar patterns have also been found with other non-human primate species using both auditory and visual tasks (see, e.g., Hare et al., 2006; Melis et al., 2006).

It's worth reemphasizing that infants, great apes, and even monkeys seem to be succeeding at these tasks by exercising a capacity for theory of mind that is completely *factive*[21]. What makes these successes particularly interesting is that there is currently no good evidence to date that monkeys have any capacity for non-factive representations (Martin and Santos,

[21] See §Section IV in the supplement to this article where we consider alternative explanations and show how our framework provides straightforward ways of augmenting these paradigms to ensure that they actually do provide clear evidence for factive theory of mind.

2014, 2016), only very limited and mixed evidence in the case of great apes (Kaminski et al., 2008; Krachun et al., 2009; Krupenye et al., 2016), and a growing debate of over the case of non-human infants (Powell et al., 2018; Baillargeon et al., 2018; Dörrenberg et al., 2018). However, completely setting aside the potential for an inability for non-factive theory of mind across all of these cases, the studies we've pointed to provide promising 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 good evidence for a genuine capacity for theory of mind.

Of course, the claim here is not that these infants or non-human primates understand what mental states are in any kind of complex or reflective way, that they are taking some kind of intentional stance, or that they have the same understanding of mental states that adult humans have. Thinking of theory of mind in these terms is exactly what led us astray in the first place. Rather, our claim is that there is good evidence that they can track others' understanding of the world and keep those representations separate from their own understanding. And if that is not theory of mind, we are not sure what is.

Acknowledgements. Among many other people, we would like to thank Lindsey Drayton, Lindsey Powell, Alia Martin, Jessie Munton, Laurie Santos, Brian Leahy, Enoch Lambert, the Yale Cognitive Science Reading Group and the SHAME writing group.

On behalf of the reader, we also thank Matthew Mandelkern for helping us to stay the course in the formalization we provided—invaluable advice for the precision of our proposal.

References

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

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

Baillargeon, R., Buttelmann, D., and Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, 46:112–124.

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.

Byrne, R. M. (2017). Counterfactual thinking: From logic to morality. *Current Directions in Psychological Science*, 26(4):314–322.

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.

Dörrenberg, S., Rakoczy, H., and Liszkowski, U. (2018). How (not) to measure infant theory of mind: testing the replicability and validity of four non-verbal measures. *Cognitive Development*.

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.

Dretske, F. (1988). *Explaining behavior*. MIT Press.

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

Fodor, J. (1990). A theory of content, ii. In *A theory of content and other essays*. Cambridge: MIT Press.

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.

Gallistel, C. and King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Wiley-Blackwell.

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. (2014a). False belief in infancy: a fresh look. *Developmental science*, 17(5):647–659.

Heyes, C. (2014b). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, 9(2):131–143.

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

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.

Leahy, B., Rafetseder, E., and Perner, J. (2014). Basic conditional reasoning: How children mimic counterfactual reasoning. *Studia logica*, 102(4):793–810.

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.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, 86:281–97.

Nagel, J. (2017). Factive and nonfactive mental state attribution. *Mind & Language*, 32(5):525–544.

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

Povinelli, D. J. and Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, 19(1):1–28.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., and Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46:40–50.

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.

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

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.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review*, 34(3):265–293.

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.

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

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.