Actual Knowledge

Abstract: This response argues that when you represent others as knowing something, you represent their mind as being related to the actual world. This feature of knowledge explains the limits of knowledge attribution, how knowledge differs from belief, and why knowledge underwrites learning from others. We hope this vision for how knowledge works spurs a new era in theory of mind research.

1. Introduction

Since the publication of Premack and Woodruff's (1978) article "Does the chimpanzee have a theory of mind?" in this journal, researchers have taken the point of theory of mind to be determining the content of others' thoughts—what exactly it is they think or want. And it has become an accepted truism that this capacity is essentially for predicting and explaining others' behavior (the literature, *passim*). The central and basically only point we are going to make in this response is that this way of understanding theory of mind has gotten it all backwards.

We think that the capacity for *theory of mind*, in its most basic form, is not primarily concerned with the content of others' thoughts which sometimes happens to reflect the actual world; it's primarily concerned with the content of the actual world, which other minds happen to reflect. The signature features of this basic capacity—that it is factive, that it requires more than justified true belief, that it allows you to represent others as knowing *more* than you, and that it is not modality specific—suggest that the capacity did not evolve specifically for predicting and explaining others' behavior. After all, the ability to represent someone as knowing where they hid the cookies from you, for example, isn't particularly useful for predicting where they'll go to get the cookies. It's true that there are cases in which this basic capacity will be useful for determining the content of others' thoughts or predicting and explaining their behavior, but our point is that the signature features of knowledge make it clear that this basic form of theory of mind did not evolve for this purpose in particular. Instead, we've argued, the signature features of this basic capacity suggest it evolved to help you interact with and learn from others, precisely because it allows you to keep track of what they understand about the actual world.

So, despite what a few philosophers said in their commentaries to Premack and Woodruff's article now more than forty years ago, the core cases in theory of mind research should never have been ones involving false beliefs. Much more revealing are ones in which someone else is better informed than you, rather than worse. So if you're in the business of reading *Behavioral and Brain Sciences* in search of paradigms that separate one's own representation of the world from others' understanding of it, these seem like good ones that might reveal something about our basic capacity for theory of mind. Perhaps it's finally time to let go of false beliefs and focus instead on the way things actually are.

A number of the commentaries to our target article objected to this way of understanding theory of mind. Some objected to the idea that theory of mind representations concern the actual world rather than the contents of others' minds (Section 2). Others argued that the signature limitations on this basic form of theory of mind simply boil down to limits on what kind of *content* one can or does attribute to others' minds (Section 3-4). Still others objected that theory of mind really is for predicting and explaining others' behavior, as opposed to coordinating on the actual world or learning about it from others (Section 5).

By and large, we think these commentaries are onto something. In fact, we think that if you can just accept that knowledge is concerned with the actual world, you'll get to have your cake and eat it too. There are limitations on what kind of content you represent others as knowing, but those boil down to limitations on the kind of content you think the actual world involves. And sometimes this basic capacity for theory of mind does let you figure out the content of others' thoughts, but those are just the cases in which you also happen to know whatever it is someone else knows. If you *do* happen to know the answers to a test question (say you wrote the exam), you'll know a great deal about what is going on in the minds of the students who also know the answers. In such cases, you'll even be pretty decent at predicting what answers they'll give and explaining why they gave those answers. But in all the other cases, where you don't know what it is that someone else knows, the exact same capacity will still be incredibly useful—this time for learning about the actual world itself, which is one of the things that the capacity for knowledge representation is particularly well-designed for.

Other commentaries saw the merit in our basic vision for theory of mind and came up with a surprising number of elegant suggestions for improvement. We did our best to take these to heart, and we'll point out the many places where they've made the central argument clearer or more convincing, or have expanded this basic vision toward new horizons (Sections 6-7). So much for the introduction, and now on to the details of how to actually understand knowledge.

2. Knowledge and the actual world, or, the truth

As we were at pains to point out in the target article, knowledge, unlike belief, is factive. While there are different ways to spell out the factivity condition on knowledge (see the commentary by **Nagel**), what is not controversial is that it ensures that you cannot represent others as knowing anything you take the actual world to preclude. Thus, there is some uncontroversial correspondence between your own understanding of the actual world and your attribution of knowledge to others. But what does this correspondence amount to?

A number of the commentaries propose that this basic form of theory of mind essentially involves your own understanding of the world (among others, see the commentaries by **Durdevic & Krupenye, Tomasello**, and **Westra**). On this view, when you represent others as knowing something, you represent their mind as being related to the actual world as you understand it. You yourself take the actual world to be some way, and representing someone as knowing something involves representing them as having the right kind of relation to that part of the world. The things that they know, then, are things involved in your own understanding of the world. As emphasized in the commentary by **Tomasello**, a form of theory of mind with this structure might be as simple as tracking whether other agents are acquainted with physical parts of the world, such that they know where the ball is, or know the woman who knitted mittens for your niece.

One alternative proposal takes a step back from the actual world and proposes that this basic form of theory of mind essentially involves monitoring whether others' understanding conflicts with your own (see, e.g., the commentary by **Deschrijver**). The actual world might happen to inform one's own understanding, but does not play a direct role, since knowledge attribution takes place at the level of tracking the relationship between two minds: your own mind and another's.

A third approach, which we might think of as the farthest from actuality, argues that this basic form of theory of mind, if it is a genuine form of theory of mind, must be meta-representational in the same way that belief is (Leslie, 1987). That is, it must essentially involve you representing someone else's independent representation of the world, and accordingly is not essentially concerned with the actual world as you understand it. With varying degrees of commitment, versions of this approach are discussed by **Dudley and Kovács; Gordon; Kampis & Csibra,** and **Binmore**.

So, why do we think knowledge involves one's own understanding of the actual world? One reason is that if you don't think this, you don't have a natural way of accounting for why knowledge but not belief is factive. If attributing knowledge (but not belief) involves understanding others' minds in relation to the actual world, then the factivity of knowledge comes for free. Obviously, you cannot understand someone's mind as being related to some part of the actual world when you think the actual world contains no such part. If knowledge attribution is instead metarepresentational in the same way belief is, then some extra explanation must be given for why this representation just happens to be limited by precisely the bounds of your own understanding of the actual world. It's not that you couldn't give such an account, but we can't see why you would want to. There's a much simpler explanation on offer.

This way of explaining the factivity of knowledge is importantly different from the version proposed by **Nagel**. Nagel proposes that the factivity constraint on knowledge is *modal*: knowledge *necessarily* only binds agents to truths. The natural way of understanding Nagel's suggestion is that when people represent someone as knowing something, they understand the person's mental states as having this property (necessarily being true). From our perspective, the trouble with this approach is that the data suggest that knowledge attribution is unlikely to involve representing any such modal property. Non-human primates, for instance, seem to have a remarkable capacity to attribute knowledge (see §4.1), but we'd be pretty shocked if they have the capacity to represent anything as being necessarily true. And if they don't have the capacity to represent anything as necessarily being true, then a fortiori they don't have the capacity to represent others' mental states as necessarily being true. So, we can be pretty sure that if non-human primates do attribute knowledge, knowledge does not involve reasoning about which truths hold across possible worlds. And there's evidence that human adults aren't all that different (Turri, 2018). In fact, humans seem relatively happy to attribute knowledge in exactly the cases that philosophers designed to illustrate that knowledge cannot be attributed when such modal properties are violated (see, e.g., Colaço et al., 2014, on fake barn intuitions and Turri, 2016a on reliabilism). Such attributions are to be expected if knowledge concerns the actual world, not merely possible ones. So the first, and perhaps most obvious reason to think knowledge involves the actual world is that this gives you a simple explanation of both why and how knowledge is factive.[1]

A second reason to find this approach promising is that it also explains why the capacity for knowledge representation is more basic than the capacity for belief representation (see the commentary by **Westra** for a related line of reasoning). If knowledge attribution involves understanding others' minds in relation to the world, then one can maintain a single

[1] We'd like to note that some of the authors (WB and JT) have recently challenged views on which the factivity of knowledge requires that one can only know things that are strictly speaking or precisely true (Buckwalter & Turri, 2020a; 2020b).

representation of the world, parts of which others also know about. Belief, unlike knowledge, cannot involve one's own understanding of the world in the same way because beliefs can be false. Thus, belief, unlike knowledge, requires an independent representation of the world—the world merely as understood by another—which must be maintained separately from one's own understanding, and thus can be false.

Once again, another feature of knowledge—its comparative basicness—falls naturally out of our way of understanding knowledge attribution, while other approaches leave this feature unexplained. Consider, for example, the suggestion by **Lassiter** that the basic theory of mind capacity we presented may be better explained by representations of *true* belief rather than knowledge, or the suggestion from **Sobel** that the evidence may be better explained with the notion of *prelief* (i.e., representations that are understood to not be real, but also are not understood to be false, as in pretense). What remains perplexing is why attributions of true belief or prelief, which require the same resources as genuine belief representation, would show all of the signature features of a more basic cognitive capacity: emerging early in phylogeny, ontogeny, processing time, and may be processed automatically and persist in the face of other cognitive impairments.

Of course, one could go on to give some further explanation of why some additional difficulty emerges in cases of false beliefs in particular, e.g., **Deschrijver** proposes a difficulty with conflict monitoring. However, such explanations face the challenging task of carving apart the cases that are genuinely difficult from cases that seem similarly complex but are not as difficult. Consider, for example, a recent piece of empirical evidence that demonstrates a nuanced capacity for attributing knowledge in monkeys (Horschler, Santos, & MacLean, 2019). In this experiment, monkeys watched an experimenter who saw a piece of fruit move into one of two containers in a display in front of them. A screen then blocked the view of the experimenter and one of two things occurred. In half of the conditions, the fruit itself briefly moved out of the container and then back inside. In the other half of the conditions, the fruit remained where it was, but the container briefly moved off of the fruit and then back on it. In both cases, all objects had returned to the position where the experimenter had last seen them, and using looking time, researchers investigated whether the monkeys expected the experimenter to reach for the fruit where it was last seen. What Horschler and colleagues found was that when the box moved, monkeys continued to expect the experimenter to reach for the fruit where they had last seen it. However, when the fruit moved instead, monkeys no longer expected the experimenter to reach for the fruit where they had last seen it.

If your account proposes a difficulty with conflict monitoring (**Deschrijver**) or that one can only represent true beliefs (**Lassiter**), these results are worryingly hard to explain. Such accounts focus on representations that are independent from the actual world. Thus, when the experimenter doesn't know about them, things that happen in the actual world shouldn't change what the experimenter believes. Accordingly, the most natural prediction for such accounts is that in *both* conditions the experimenter will be represented as having a belief about the location of the fruit, and this belief will happen to be true—it will match the monkey's own ideas about the location of the fruit. So, in both conditions, monkeys should expect the experimenter to reach for the fruit where it actually is. But, of course, monkeys don't do that. Instead Horschler and colleagues found that monkeys only expect the experimenter to reach for the fruit when it was the container, rather than the fruit, that moved.

The difference between the conditions is easy enough to explain, however, if monkeys represent knowledge rather than belief. Knowledge requires more than having a justified true belief (§2; Gettier, 1963). And so when the fruit moves but the experimenter doesn't see it (but then happens to return to the original location), the experimenter might end up with a true belief about the location of the fruit by coincidence, but they do not share the monkey's understanding of the location of the fruit. By contrast, when only the container moves, this should *not* affect the experimenter's knowledge of the fruit, and monkeys should continue expecting the experimenter to act in accordance with this knowledge. This is exactly what they do.

This is just one of a growing number of studies that demonstrate clear failures to represent others' true beliefs while simultaneously demonstrating clear success in representing their knowledge (see Krachun, et al., 2009 and Horschler, et al., 2021). The key difference is that knowledge tasks can be passed by simply keep track of whether the agent understands the relevant part of the actual world, while the true belief tasks require you to construct a separate representation of the world as the agent understands it, which just happens to align with your own understanding, and thus is true.

So in short, if you can accept that knowledge concerns the actual world, you get a surprisingly simple explanation for why knowledge is basic, why it is factive, and how it differs from belief.

3. But what do we know anyway?

Instead of locating the difference between knowledge and belief in the role of one's own understanding of the world, a number of commentaries argued that the essential difference between them concerns the kind of content they allow you to attribute. After all, as **Tomasello** and **Starmans** point out, human languages typically encode an intriguing difference between knowledge and belief. In English (as in many other languages), one can know ways of doing things and know the smell of summer rain, but one cannot *believe* ways of doing things or *believe* the smell of summer rain. **Tomasello** and **Starmans** argue that the content of belief attributions seems to be propositional, while the content of knowledge attributions can be both actual things in the world and abstract matters of fact.

Following their line of thought further, it wouldn't be surprising if there were different mechanisms for understanding, on the one hand, the kind of acquaintance other agents have to physical parts of the world and, on the other hand, their acquaintance with things like abstract propositional truths. Moreover, it is plausible enough that the mechanisms for figuring out what physical parts of the world another is acquainted with may be simpler than the mechanisms for figuring out which propositions another is acquainted with. And as **Tomasello** and **Starmans** point out, much of the evidence for basic knowledge ascriptions in non-human primates and human infants suggest that these populations represent others as knowing about physical objects or having certain skills. So, perhaps all of this points to a key distinction in kinds of knowledge, with a basic form of knowledge attribution that amounts to little more than knowledge by acquaintance or know-how and differs sharply from belief attribution, and a separate more complex form of propositional knowledge ascription that is not more basic than belief ascription but is rather quite similar to it. On this view, the difference in basicness we illustrated in the target article is a matter of the basicness of the content attributed (propositional vs. non-propositional), and not truly a matter of the basicness of the attitude itself (knowledge vs. belief). This all seems quite convincing.

The trouble is that there actually seems to be a simpler explanation for why there isn't great evidence that non-human primates and human infants represent others as having knowledge of abstract propositions. Namely, there isn't great evidence that non-human primates and human infants represent abstract propositions in general. If knowledge attributions essentially involve your own understanding of the world, then the kind of content one can represent others as knowing will depend on what kind of content your own representation of the actual world involves.

Moreover, the similarity between propositional and non-propositional content shouldn't be hard to see here. If you do not understand the actual world to involve any extraterrestrial aliens, you could not represent anyone as knowing them ("knowledge by acquaintance"). And if you do not think there are ways of turning water into gold bullion, you can't represent anyone as knowing how to do that ("knowledge-how"). And in just the same way, if you do not understand the actual world to involve abstract propositions, like "2+7=10" then you certainly will not be able to represent others as knowing this sort of thing either ("propositional knowledge").

Non-human primates (and perhaps very young human infants) may not have the capacity to represent propositions, and thus their knowledge representations will necessarily be restricted to simpler forms of content, whether knowledge-by-acquaintance (**Tomasello**) or even just visual perspective (**Asaba, Chuey, and Gweon**). And if this is right, then such knowledge representations are also likely to be guided by specific attention to cues such as eye gaze or direct perception (**Call**; **Grossmann & Dela Cruz; Dudley & Kovacs**). However, for human adults who clearly can and do represent the world in something closer to propositional terms, the same capacity may be used to represent others as having knowledge of abstract truths. For example, as emphasized beautifully by **Mikhail**, human adults represent others as having moral and legal knowledge. While unquestionably abstract, these rules make up part of our understanding of the world, and given that, we have no trouble representing others as sharing our understanding of them. Note that in the latter kinds of cases, we agree with **Westra** that there is reason to think the content of knowledge is propositional and with **Farina & Lavazza** who argue that knowledge is content-involving and representational.)

Importantly though, even for unambiguously propositional content, attributing knowledge seems easier than attributing belief. One completely uncontroversial piece of evidence is that young children succeed at attributing propositional knowledge (e.g., "Sally does not know her marble is in the basket.") before they succeed in attributing similarly propositional beliefs (e.g., "Sally believes her marble is in the box."). Similarly, adults are faster to correctly attribute or deny knowledge claims than they are to correctly attribute or deny corresponding belief claims, even when the term used for knowledge is explicitly propositional, e.g., 'savoir' in French, which only takes propositional complements (Phillips, et al., 2018). A third piece of evidence comes from the commentary by **Bricker**, who used EEG to show that propositional knowledge representation elicited a weaker P3b amplitude than belief representation (Bricker, 2020). Thus, even when knowledge attributions unambiguously involve propositional content, they continue to show signs of emerging earlier, being simpler, and requiring less processing than matched belief attributions.

So, it turns out the surprisingly simple solution is that the mechanism for representing knowledge is just the same across all of these different kinds of cases—you are just figuring out what parts of the world someone else understands--and seeming differences in the complexity of knowledge attributions across species or development arise simply from the complexity of representing different parts of the actual world. (See **Rosenbaum, Halilova, and Pathman** for related commentary on the difference in complexity between episodic and semantic content in knowledge attribution.)

The upshot of our view is that knowledge attributions won't be limited to any particular type of content (propositional, knowledge-how, etc.). Knowledge attributions can be as rich as your own understanding of the world. It is for this reason that we suspect that the capacity for knowledge attribution we provided evidence for in the target article will be not be fully captured by approaches that place limits on the content of basic theory of mind attributions, for example, reducing it to representations of visual perspective (**Asaba, Chuey, & Gweon**), uninterrupted perceptual access (**Dudley & Kovács**), skill (**Carpendale & Lewis**), goals (**Schlicht et al.**), episodic experience (**Kampis & Csibra**), or knowledge by acquaintance (**Tomasello**). While each of these commentaries does an excellent job of pointing to specific aspects of knowledge we can attribute to others, it would be quite surprising on each of these views if knowledge attribution just happened to work in much the same way in all the other cases as well. That is, each of these cases shares the signature features of knowledge attribution (§2 of the target article). We don't think this is surprising though. Each of these cases involve various aspects of the actual world, and representing others as knowing that part of the world will work similarly in each case.

If you are wondering at this point whether we are really proposing that knowledge attribution may function in essentially the same way in non-human primates as it does in human infants and adults, let us be clear. We are. In fact, the commentary by **Moss** suggests that it might even extend to philosophers. Moss argues that the history of philosophy suggests that explicit theories of knowledge preceded those of belief in the Presocratics. As she argues, this suggests that explicitly theorizing about knowledge may be easier for creatures like us than explicit theorizing about belief. It may then be no coincidence that this empirical fact aligns with the other ones we reviewed in our target article and may provide yet another indicator that knowledge is more basic than belief for creatures like us—even those of us who are philosophers.

4. Knowing what you don't know

A third objection that was touched on by a number of the commentaries was that theory of mind is for predicting and explaining behavior (see, e.g., **Binmore**, **Gordon; Dudley & Kovács**). This perspective makes sense if one is committed to belief being the most basic theory of mind representation. But if we are right that knowledge is more basic than belief, then the trouble faced by this approach is that the more basic form of theory of mind seems oddly ill-designed for action prediction and explanation in particular (see **Bazhydai & Harris** for a similar line of reasoning). One notable feature of knowledge representation is that it seems to require more than justified true belief. But of course, justified true belief should be more than sufficient if your goal is just to predict someone's actions. Even unjustified false beliefs will do. A second notable feature of knowledge representation is that it allows you to represent others as knowing more than you yourself know. Others know all sorts of things you don't. But just knowing *that* others know more than you doesn't do you much good if your primary goal is predicting what they are going to do or explaining why they did what they did. So it's odd that our theory of mind capacity would have these particular features if it primarily evolved for the purpose of predicting and explaining behavior.

In contrast, if knowledge attributions involve representing others as understanding the actual world, then the ability to represent others as knowing more than you isn't particularly puzzling. In fact, it's precisely what you'd expect. When you represent someone as knowing more than you, you represent them as knowing something about the actual world you do not. You probably do not know how to play the zither, but you do think that there are in fact ways of playing the zither. And if you didn't think there was a fact of the matter, you couldn't represent someone as knowing that fact. For example, those of us who don't think each person's soul weighs a certain amount can't represent others as knowing the amount each soul weighs. Further, in cases where you yourself don't exactly know something, but you have a pretty good idea about it, you have a correspondingly good idea of what it is that the other person knows. And when you yourself have a great idea about the relevant part of the world, you'll have a correspondingly great idea about the content of someone else's mind. If you know why you randomly assign participants to conditions in a controlled experiment, and you represent someone else as knowing why too, then you'll have a great idea of exactly what it is they know. Not only do you know the precise content of their mental states, but you'll be able to predict what they'll do, and explain why they did what they did.

So it's not that we don't think knowledge representations can be used for prediction and explanation or that these representations don't reflect the content of others thoughts, it's just that the traditional view gets it backwards. Knowledge concerns the actual world and which parts of it others understand. In some cases, others' understanding of the actual world will align with yours, and in those cases, you will know the content of others' thoughts, and be able to predict and explain their behavior. But there are also cases in which others' know more about the actual world than you do. Our point is that your own representation of the actual world plays much the same role both when you attribute knowledge to another of some fact you *do* know and when you attribute knowledge of facts you do *not* know. In both cases, you are representing another as understanding some part of the actual world (the way *that* part of the world actually is). What is changing is simply your own understanding of that part of the world (see **Durdevic & Krupenye** for related discussion).

This proposal for how to understand others as knowing more than you (egocentric ignorance) differs in important ways from the suggestions raised in many of the commentaries. For comparison, consider the proposal by **Tomasello** that the basic form of knowledge involves only knowledge by acquaintance. On this view, non-human primates only represent others as having been acquainted (or not) with physical objects in the world. Following **Kampis and Csibra**, suppose that the mechanism here works by simply tagging which physical objects someone is acquainted with. As Kampis and Csibra point out, such a mechanism does not seem to allow for representations of egocentric ignorance. To make this concrete, consider the success apes have in representing conspecifics as knowing whether there is a piece of fruit in a given box even when they themselves do not know (e.g., Kaminski et al. 2008). In such cases, subjects don't actually represent there being a piece of fruit in the box, and thus it's hard to see how they could tag that object as having been acquainted with the relevant conspecific. What this example illustrates is the difficulty in accounting for egocentric ignorance faced by views that reduce knowledge representations to simple representations like acquaintance, tagging, visual perspective, or perceptual access. Even more perplexing is how this kind of knowledge representation could be extended to understanding others as knowing how to crack open a nut (as non-human primates do, Rapaport & Brown, 2008) or knowing how to play the zither (as we humans do). What others know in such cases are not objects that can be tagged and clearly cannot be reduced to some particular visual perspective.

At the same time, our proposal for how to understand what happens when you represent others as knowing *less* than you (altercentric ignorance) also differs from those discussed in the commentaries. For example, **Deschrijver** suggests that altercentric ignorance may amount to simply attributing no representation whatsoever to another agent – much like the representation you attributed to the Prince of Liechtenstein before reading this sentence. But, just as it is possible to represent someone as sharing your knowledge of some particular part of the world (e.g., knowing a person) without representing them as sharing all your knowledge (e.g., knowing all the people you know of), it is possible to represent someone as *not* sharing your knowledge of a particular part of the world, without representing them as not sharing any of your knowledge. That is, the basic capacity for knowledge attribution allows for representations of knowledge and ignorance about specific parts of the world (this point provides a helpful contrast with the suggestion from **Gordon** that we may simply attribute all of our knowledge to others by default). In fact, much of the evidence we reviewed demonstrates precisely this kind of specificity in attributions of knowledge and ignorance. Consider simple studies in which non-human primates will selectively steal the piece of food that a dominant competitor does not know about (Hare, Call, Agnetta, & Tomasello, 2000). Success on these tasks requires that chimpanzees selectively represent the dominant competitor as ignorant of the existence of one piece of food while knowledgeable about the other. If they simply attributed no representation whatsoever to the other chimpanzee, they should be equally likely to take either piece of food (since the other chimpanzee would be equally unaware of both). When one represents others as ignorant, there must be specific parts of the world you do not represent them as knowing.

Importantly, the kind of ignorance we have been discussing does not involve representing someone else as being *aware* of their own ignorance (this would require a separate capacity involving metarepresentation, see **Durdevic & Krupenye**). The difference is that when you represent another agent as being selectively ignorant about some part of the actual world, the predictions you'll make concern only the other parts of the actual world they do know about (you'd predict that they'd be upset about you eating the food they do know about, but not the food that they don't know about). But if you are able to represent other agents as having some awareness of their own ignorance (knowing *that* they don't know), then the predictions you'll make may also concern the ignorance itself, and what other agent's might do to alter their ignorance. As emphasized by **Royka and Jara-Ettinger**, one possibility is that such a meta-representation requires a mind that can employ some kind of symbolic negation operator, allowing you to represent the agent as knowing that they *do not* know. Future work may want to explore this possibility.

We have been arguing for an understanding of knowledge ascription that is both rich and flexible in some ways and notably limited in others. However, both the richness and the limits arise from a single unassuming commitment: when one represents others as knowing or not knowing something, one represents them as knowing or not knowing something about the actual world.

5. Give learning a try

This way of understanding knowledge fits seamlessly with our claim that knowledge is for learning. When you represent your friend as knowing how to ride a bike, even though you don't know how to, you take them to understand something about the actual world: a way in which bikes can be ridden. You are not particularly interested in their ideas about how to ride a bike, independent of whether they actually work; what you take them to know and what you want them to teach you is how to actually ride a bike.

A number of commentaries pushed back on this basic idea, arguing that the capacity for belief representation is a better candidate for underwriting learning from others, especially given the success of cultural evolution in humans in particular (**Gordon**; **Dudley & Kovacs; Richardson**; **Salazar; Sobel**).

One form of this objection was succinctly put by **Richardson**, who argues that knowledge cannot be both what humans share with non-human primates and what explains humans' unique capacity for cultural accumulation. Stated this way, we couldn't agree more. While we do think that the capacity for knowledge attribution is likely shared with non-human primates, we agree that knowledge is not what explains humans' unique capacity for cultural learning. Rather, we suspect that humans' unique cultural accumulation of knowledge is instead explained by our unique *representational capacities*— perhaps the capacity for representing abstract propositions, encoding information linguistically, and so on. We believe these sorts of capacities, not the capacity for knowledge attribution, is what differentiates humans from other species. But of course, none of this means that knowledge attribution doesn't play a central role in the process of accumulating cultural knowledge. If knowledge attribution works the way we've been arguing, then changes in domain-general representational capacities will result in changes in what we can and do represent others as knowing, which in turn will change what we can learn from them. And so, while non-human primates may accumulate knowledge of foraging techniques (e.g., Musgrave, et al., 2020), human infants may accumulate knowledge of the names of novel objects (**Bazhydai & Harris**), and human adults may accumulate knowledge of math, all while using the same basic capacity for representing others as knowing something about the actual world.

While this response may help to address the differences in the content of cultural learning in non-human primates, there are clearly differences not only in content but also in frequency and tendency. We suspect that our proposal has little to contribute in explaining these differences. There are myriad ways in which humans are both more social and more successful in communicating than our primate relatives (see Henrich, 2015 for a discussion).

A second form of the objection that knowledge is for learning, raised by **Sobel** among others, is that the processes for selectively determining who to learn from may be better accounted for by a form of belief representation. In a helpful response, the commentary by **Bazhydai and Harris** provides a beautiful accounting of the empirical evidence that knowledge rather than belief representations support selective learning in infants. The body of work they discuss demonstrates that infants selectively learn from others who are knowledgeable and selectively pass on information to those who are ignorant, all while not yet demonstrating any real capacity for the kind of metarepresentation required by belief. This literature similarly helps to address the point raised by both **Kampis and Csibra** and **Handley-Miner and Young** that if knowledge representations are going to be useful for social learning, they need to be accompanied by mechanisms for determining who actually knows what you want to know. The literature on trust in testimony provides remarkably thorough evidence for how these mechanisms may function, and we hope that this literature will become increasingly integrated into theory of mind research (see **Bazhydai & Harris and Salazar,** Harris et al., 2018). On a related note, one would expect that what we choose to teach others is also guided by knowledge, and here again, there is a growing body of evidence that knowledge plays a key role in guiding the information we provide to others (Turri, 2016b).

Finally, it may be worth being explicit that none of this means that belief attribution cannot also support social learning. However, if we are correct about the essential difference between knowledge and belief, then the cases in which belief attribution plays an essential role will be ones in which what you need to learn is something specifically about how others think, not how the world actually is. When the emperor wears no clothes, successfully predicting and coordinating with others certainly will require belief-based social learning. Still, we suspect such cases make up the periphery rather than the core of learning from others, especially in the course of primate evolution.

6. What to do with belief?

Throughout, we have been arguing for a central way of understanding the differences between knowledge and belief attribution. An important separate question, which was raised in a number of the commentaries, instead asks how these two forms of attribution may be related to one another (**Bender & Gatewood**, **Ninan**, **Nagel**, **Brakel**, **Kano & Call**, **Durdevic & Krupenye**).

As pointed out in the commentary by **Nagel**, the account of knowledge attribution we've given can occur entirely independently from belief attribution. Thus, our view differs in an important way from standard philosophical views of knowledge, according to which knowledge entails belief. It is important here to keep in mind the difference between the philosophical claim about the concept of knowledge and the psychological claim we have made (see the commentary by **Gerken** for a similar concern). We've argued that the representation of knowledge does not entail the representation of belief. This point is directly supported by the empirical evidence. For example, we've argued that monkeys can represent knowledge but not belief. And if that's right, it clearly can't be the case that their representing knowledge entails their representing belief. This same point is also supported by the growing experimental philosophy evidence for cases in which people will attribute knowledge but not belief (Myers-Schulz & Schwitzgebel, 2013; Yuan & Kim, forthcoming). Such cases can naturally be described as ones in which we take the agent to have access to the relevant part of the world even though this access doesn't exhibit the normal impact on the agent's thoughts or behavior (see the commentary by **Brakel** for related ideas).

One way of thinking about this independence between knowledge and belief aligns with the proposal from **Kano and Call**, according to which the capacities for knowledge and belief attribution are entirely separate. Kano and Call agree that human infants seem to have an ability to attribute knowledge but not belief. However, they are moved by the studies providing evidence for false belief representation in apes (Kano et al., 2019, Krupenye et al., 2016; see also the commentary by **Durdevic & Krupenye)**. Thus, as they argue, given that the two capacities do not consistently appear together, perhaps they should simply be understood as arising from separate systems. While we are less convinced that the existing research provides sufficient evidence for a capacity for belief representation in non-human primates (and monkeys in particular, see §4.1), we can set this question aside for now. Notice that if Kano and Call turn out to be correct that some non-human primates have a capacity for belief representation, there would remain a remarkably consistent pattern across species: one never finds a capacity to represent beliefs in the absence of a capacity to represent knowledge.

This consistent pattern suggests an alternative way of understanding the relation between knowledge and belief that aligns instead with the proposal from **Ninan**. As argued by Ninan, the capacity for belief representation may depend on a prior ability to represent knowledge. Following an idea from Williamson (2000), Ninan suggests that instances of representing others as believing something may essentially be instances of representing someone as acting *as if they* *knew something.* If this is right, then belief attribution (even in cases where the belief is true), would require a form of counterfactual conditional reasoning. In other words, it would require representing a merely possible way the actual world could have been, and then taking the agent to be related to that world in much the same way we take others to be related to the actual world when they know things about it. Three features make Ninan's proposal intriguing. The first is that it could explain the general pattern whereby belief attribution appears later in development than knowledge attribution. The second is that it fits well with the empirical correspondence one finds in human development between counterfactual conditional reasoning and belief attribution (Riggs & Peterson, 2000). And the third is that it provides one way of understanding why there are cases in which knowledge does not entail belief, since knowing something does not entail acting as if one knew that thing (Meyers-Schulz & Schwitzgebel, 2013; Radford, 1966).

While we are not yet sure whether belief representation should be understood as depending in some way on knowledge representation, this is clearly an important area for future research.

7. One thousand flowers

There remains a great deal we do not know about the basic theory of mind capacity we have been concerned with. At least partially, this is because knowledge attribution has received comparatively less attention than belief attribution in the history of theory of mind research. So, while we agree with **Richardson**, **Dudley and Kovács**, and **Call and Kano** that our paper should probably not incite a wholesale abandonment of the study of belief attribution, we want to emphasize the range of commentaries that pointed to important new questions and future directions for the study of knowledge. We hope these questions spur a new era in theory of mind research.

7.1. Catching up

In the past forty-plus years— starting from the proposal of the false belief task in the commentaries to Premack and Woodruff's (1978) article in this journal—we have learned a great deal about belief representation. We have largely reached a consensus on the neural substrates involved in representing false beliefs (e.g., Saxe & Kanwisher, 2003). We have developed elegant ways of computationally modeling the process of belief attribution and update (e.g., Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Schulz, & Tenenbaum, 2020). We have an increasingly good idea of when the capacity for belief attribution arose over the course of evolution (e.g., Marticorena et al., 2011). And we have thoroughly studied the extent to which humans automatically represent others' false beliefs (e.g., Apperly, et al., 2006; Kovács et al., 2010; Phillips, et al., 2015). Yet, as pointed out in many of the commentaries, we have corresponding gaps about each of these when it comes to knowledge.

7.1.1. The neuroscience of knowledge attribution

The commentaries by **Bricker** and **Gordon** call for the emergence of the neuroscientific study of knowledge attribution. Bricker's EEG study (Bricker, 2020) is a helpful early step in this direction. He finds that belief representation demands more neural resources than knowledge representation as indicated by differences in P3b amplitude. A clear implication of this finding is that knowledge representation—even propositional knowledge representation—does not depend on belief representation, since representing the agent's knowledge requires less processing than representing the agent's beliefs. Still, many open questions remain. Because theory of mind networks have quite literally been defined by false beliefs (i.e., a false belief vs. false photograph contrast, Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011), we don't yet have much of an understanding of the neural mechanisms involved in knowledge representation. Thus, an important and completely open question is whether knowledge representation recruits the same theory of mind network as belief representation or relies on a distinct set of neural substrates.

7.1.2. The computation of knowledge attribution

The commentaries by **Asaba, Chuey, and Gweon**, **Royka and Jara-Ettinger**, and **Krupenye,** emphasize the importance of understanding the computational processes that underwrite knowledge attribution. The existing work on computational theory of mind relies on inferences over belief states, whether through Bayesian inference (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017), inverse reinforcement learning (Jara-Ettinger, Schulz, & Tenenbaum, 2020), or another mechanism (Koster-Hale and Saxe, 2013). What the current proposal suggests is that there may be simpler forms of theory of mind computation that do not require representing and reasoning over the potentially huge number of wrong beliefs an agent may have. Moreover, if the current proposal is correct, then the computations that underwrite knowledge attribution may instead directly recruit one's own understanding of the world, which would serve to drastically reduce the space of possible knowledge states necessary to reason over. We hope future work takes up this challenge.

7.1.3. The evolution of knowledge attribution

While we've argued that knowledge arose before belief, this does not settle the question of *when* the capacity for knowledge representation actually evolved. The commentary by **Veit** suggests this capacity arose during the Cambrian explosion. Maybe so, but either way, this is an empirical and testable claim we hope is taken up in future work by studying knowledge representations in species less related to us than non-human primates, such as corvids, canines, or even octopuses.

7.1.4. The automaticity of knowledge attribution

The commentary by **Surtees and Todd** points out that a great deal remains to be done in studying implicit, spontaneous, or automatic knowledge representations. As we've argued, and was expanded on by Surtees and Todd, most of the current evidence is based on visual perspective taking tasks, which at best can only provide suggestive evidence for the automatic calculation of genuine knowledge representations. (The evidence is less equivocal about belief representations, which clearly do not seem to be automatic.) While there have been a few studies that have looked at abstract knowledge rather than visual perspective taking (e.g., Dungan and Saxe, 2012), the question of whether we automatically calculate what others know, and what the limits of these calculations are, remain important questions for future work.

7.2. Looking forward

In addition to commentaries proposing that we need to understand knowledge in the same ways we've come to understand belief, other commentaries emphasized that there are aspects of knowledge that merit studying on their own grounds.

In this vein, the commentary by **Gerken** points toward the importance of studying the biases and limits of knowledge representation. As Gerken argues, some interesting features of knowledge representations may provide further clues to how this capacity functions. We agree that studying the signature limits of knowledge ascription is an important and productive avenue for future work. We suspect that this approach will also help uncover ways in which knowledge and true belief attribution come apart, the importance of which was emphasized by **Lassiter** and **Durdevic & Krupenye**.

Similarly, the commentary by **Machery, Barrett, and Stich** argues for the importance of studying cross-cultural and cross-linguistic variation in knowledge ascription (also emphasized by **Bender & Gatewood**). While it would be surprising if there was genuinely no cross-cultural or cross-linguistic variation in knowledge ascription, the extant evidence indicates that many of the notable features of knowledge attribution exhibit remarkable cross-cultural stability. For example, a well-known developmental finding is that there is remarkable stability in the order in which children pass a battery of theory of mind tasks (Wellman & Liu, 2004), and variations across languages and cultures are relatively minor (e.g., Shahaeian et al., 2011). Moreover, Machery, Barrett, and Stich (and their colleagues) have found robust evidence that knowledge is denied across cultures in Gettier cases (Machery et al., 2017) and that knowledge ascriptions are equally insensitive to stakes across cultures (Rose et al., 2019). And there is even new evidence for cross-cultural stability in the tendency to attribute knowledge in cases where belief is denied (Yuan & Kim, forthcoming). In their commentary, Machery, Barrett, and Stich hint at some preliminary evidence that there may be cases in which this last feature of knowledge is not exhibited. If those results hold up, it would certainly be interesting and important. Still, we do not think that such a finding by itself would be problematic for our general proposal. If the capacity for knowledge representation is indeed basic in the way we've argued, one should expect a lot of generality across languages and cultures, but probably not strict universality (see Strickland, 2017). More importantly though, the only way to know whether this generality claim holds up is to do the difficult and important cross-cultural work being done by Machery, Barrett, and Stich. Thus, we echo their call to continue investigating cross-cultural and cross-linguistic variation in knowledge attribution.

A final group of commentaries emphasized the importance of better understanding knowledge representations in our social lives, especially in cases in which we interact with and learn from others.

The commentary by **Bazhydai** and **Harris** calls for studying the relationship between knowledge representation and active solicitation of teaching from and to others. As they emphasize, an important but as of yet unanswered question is whether young children exhibit higher rates of soliciting information from others when they represent them as knowing something rather than (merely) truly believing it. In a similar vein, the commentary by **Erdemli, Audrin, and Sander** suggests that the process of learning from others may partially be driven by "social epistemic emotions" and "affective social learning". We hope that researchers working on active solicitation begin to research these important questions.

Relatedly, **Handley-Miner & Young** and **Asaba, Chuey, and Gweon** emphasize the importance of studying knowledge representation in cases of real-world complexity, where people may only have partial knowledge and you may even be uncertain about who has knowledge or how much knowledge they have. Following much of the empirical work, we have emphasized cases where knowledge is relatively clear-cut. However, many real-world cases involve precisely the kind of uncertainty Handley-Miner and Young point out. The literature on trust in testimony provides a rich resource to draw on (see Harris et al., 2018 for a recent review), but better understanding knowledge attribution in the face of such uncertainty clearly remains an important avenue for future work.

Schlict et al., and **Woo, Tan, and Hamlin** and **Asaba, Chuey, and Gweon** all raise important questions concerning theory of mind about others' goals or preferences. An ability to represent others' goals and preferences, much like the ability to represent knowledge, appears early in development and before belief representation (see the commentaries by **Schlicht** and **Woo, Tan, & Hamlin**). Notice that when you represent others as having goals or preferences, these seem to involve the actual world. Others may have a goal of getting to a particular part of the actual world (say, the top of a hill), or a preference for eating some part of the world (say, cookies). An intriguing possibility then is that this form of theory of mind, much like knowledge, essentially involves one's own understanding of the actual world. And if this is correct, then we would *not* expect an early facility in attributing goals or desires, when the object of those goals or desires is precluded by the actual world (e.g., wanting to eat a cookie now that was already eaten yesterday). We hope future work investigates this possibility.

And with that, let us turn to a new chapter in theory of mind research.

References

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic?. P*sychological Science*, 17(10), 841-844.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 1-10.

Bricker, A. M. (2020). The neural and cognitive mechanisms of knowledge attribution: An EEG study. *Cognition*, 203, 104412.

Buckwalter, W., & Turri, J. (2020a). Knowledge and truth: A skeptical challenge. *Pacific Philosophical Quarterly*, 101(1), 93-101.

Buckwalter, W., & Turri, J. (2020b). Knowledge, adequacy, and approximate truth. *Consciousness and Cognition*, 83, 102950.

Colaço, D., Buckwalter, W., Stich, S. & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme*, 11(2), 199–212.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *Neuroimage*, 55(2), 705-712.

Dungan, J. and Saxe, R. (2012), Matched false-belief performance during verbal and nonverbal interference. *Cognitive Science*, 36: 1148–1156. doi: 10.1111/j.1551-6709.2012.01248.x

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121-123.

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

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5), 836-848.

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

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59(4), 771-785.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, 69, 251-273.

Henrich, J. (2015). *The secret of our success*. Princeton University Press.

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. *Cognition*, *190*, 72-80.

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2021). How do non-human primates represent others' awareness of where objects are hidden? *Cognition*, 212, 104658.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123, 101334.

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, 201910095. https://doi.org/10.1073/pnas.1910095116

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830-1834.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308), 110–114. https://doi.org/10.1126/science.aaf8110

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

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., ... & Hashimoto, T. (2017). Gettier across cultures 1. *Noûs*, 51(3), 645-664.

Marticorena, D. C., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. Developmental science, 14(6), 1406-1416.

Musgrave, S., Lonsdorf, E., Morgan, D., Prestipino, M., Bernstein-Kurtycz, L., Mundry, R., & Sanz, C. (2020). Teaching varies with task complexity in wild chimpanzees. *Proceedings of the National Academy of Sciences*, 117(2), 969-976.

Myers-Schulz, B., & Schwitzgebel, E. (2013). Knowing that P without believing that P. *Noûs*, 47(2), 371-384.

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, 26(9), 1353-1367.

Phillips, J., Knobe, J., Strickland, B., Armary, P., & Cushman, F. (2018). Evidence for evaluations of knowledge prior to belief. In Proceedings of the *Cognitive Science Society*.

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

Radford, C. (1966). Knowledge: by examples. *Analysis*, 27(1), 1-11.

Rapaport, L. G., & Brown, G. R. (2008). Social influences on foraging behavior in young nonhuman primates: learning what, where, and how to eat. *Evolutionary Anthropology: Issues, News, and Reviews*, 17(4), 189-201.

Riggs, K. J., & Peterson, D. M. (2000). Counterfactual thinking in pre-school children: Mental state and causal inferences. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (p. 87–99). Psychology Press/Taylor & Francis (UK).

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., ... & Zhu, J. (2019). Nothing at stake in knowledge. *Noûs*, 53(1), 224-247.

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

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, 47(5), 1239.

Strickland, B. (2017). Language reflects "core" cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive Science*, 41(1), 70-101.

Turri, J. (2016a). A new paradigm for epistemology: from reliabilism to abilism. *Ergo*, 3(8), 189–231.

Turri, J. (2016b). *Knowledge and the norm of assertion: An essay in philosophical science*. Open Book Publishers.

Turri, J. (2018). Primate social cognition and the core human knowledge concept. In M. Mizumoto, S. Stich, & E. McCready (Eds.), *Epistemology for the rest of the world: linguistic and cultural diversity and epistemology* (pp. 279–290). Oxford University Press.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75(2), 523-541.

Williamson, T. (2002). *Knowledge and its Limits*. Oxford University Press on Demand.

Yuan, Y. & Kim, M. (forthcoming). Cross-Cultural Universality of Knowledge Attributions. *Review of Philosophy and Psychology*