Tracking and Representing Others’ Mental States

Stephen A. Butterfill
<s.butterfill@warwick.ac.uk>

This is a slightly longer version of a chapter published in Andrews and Beck (eds), Routledge Handbook on Animal Minds

1. Introduction

Few things matter more than the mental states of those nearby. Their ignorance defines limits on cooperation and presents opportunities to exploit in competition. (If she’s seen where you stashed those mealworms she’ll pilfer them when you’re gone, leaving you without breakfast. And you won’t get that grape if he hears you sneaking past.) What others feel, see and know can also provide information about events otherwise beyond your ken. It’s no surprise, then, that abilities to track others’ mental states are widespread. Many animals including scrub jays (Clayton, Dally and Emery 2007), ravens (Bugnyar, Reber and Buckner 2016), goats (Kaminski, Call and Tomasello 2006), dogs (Kaminski et al. 2009), ringtailed lemurs (Sandel, MacLean and Hare 2011), monkeys (Burkart and Heschl 2007; Hattori, Kuroshima and Fujita 2009) and chimpanzees (Melis, Call and Tomasello 2006; Karg et al. 2015) reliably vary their actions in ways that are appropriate given facts about another’s mental states. What underpins such abilities to track others’ mental states?

There is a simple and quite widely accepted answer. As in humans, so in other animals: abilities to track others’ mental states are underpinned by representations of those mental states. Some people seem less confident about lemurs or monkeys than chimpanzees, perhaps in part because these animals’ abilities to track others’ mental states appear less flexible (Hare et al. 2003; Burkart and Heschl 2007). Others caution that there is currently insufficient
evidence to accept that any nonhuman animals ever represent others’ mental states (e.g. Whiten 2013; Heyes 2015). But overall the simple answer—abilities to track others’ mental states are underpinned by representations of those mental states—is endorsed by many of those cited above for at least some nonhuman animals.

The simple answer will appear inescapable if we assume that tracking others’ mental states must, as a matter of logic, involve representing others’ mental states. But this assumption is incorrect. Contrast representing a mental state with tracking one. For you to track someone’s mental state (such as a belief that there is food behind that rock) is for there to be a process in you which nonaccidentally depends in some way on whether she has that mental state. Representing mental states is one way, but not the only way, of tracking them. In principle it is possible to track mental states without representing them. For example, it is possible, within limits, to track what another visually represents by representing her line of sight only. More sophisticated illustrations of how you could in principle track mental states without representing them abound (e.g. Buckner 2014, pp. 571ff). What many experiments actually measure is whether certain subjects can track mental states: the question is whether changes in what another sees, believes or desires are reflected in subjects’ choices of route, caching behaviours, or anticipatory looking (say). It is surely possible to infer what is represented by observing what is tracked. But such inferences are never merely logical.

To learn what underpins abilities to track others’ mental states, we would therefore need to evaluate competing hypotheses. In recognising this, we immediately face two requirements. The first requirement is a theoretically coherent, empirically motivated and readily testable hypothesis on which tracking mental states does not involve representing mental states. This requirement is currently unmet (Halina 2015, p. 486; Heyes 2015, p. 322) and, as the next section argues, surprisingly difficult to meet.

2. Pure Behaviour Reading: Cast the Demon Out

Pure behaviour reading is the process of tracking others’ behaviours, including their future behaviours, independently of any knowledge, or beliefs about, their mental states. Can research on pure behaviour reading supply hypotheses on which tracking mental states does not involve representing mental states?

Contrast two approaches to theorising about behaviour reading. One focusses on the behaviourist counterpart of Laplace’s demon. The behaviour reading demon has unlimited cognitive capacities, perfect knowledge of
history and can conceptualise behaviours in any way imaginable. Although blind to mental states, it can predict others’ future behaviours at least as well as any mindreader (Andrews 2005, p. 528; Halina 2015, p. 483ff). Invoking the behaviour reading demon makes vivid the point that the existence of abilities to track others’ mental states does not logically entail representations of mental states. But the behaviour reading demon is little use when it comes to generating testable hypotheses. Not even the most exacting rigour requires excluding the possibility that an animal is a behaviour reading demon before accepting that it can represent mental states.

The other approach to theorising about behaviour reading concerns actual animals rather than imaginary demons. Byrne (2003) studied a particularly sophisticated case of behaviour reading in Rwandan mountain gorillas. The procedure for preparing a nettle to eat while avoiding contact with its stings is shown in figure 1. It involves multiple steps. Some steps may be repeated varying numbers of times, and not all steps occur in every case. The fact
that gorillas can learn this and other procedures for acquiring and preparing food by observing others’ behaviour suggests that they have sophisticated behaviour reading abilities (Byrne 2003, p. 513). If we understood the nature of these behaviour reading abilities and their limits, we might be better able to understand their abilities to track mental states too.

We seek an account of pure behaviour reading to generate testable hypotheses about tracking mental states without representing them. This requires us to focus on the actual behaviour reading capacities of particular animals. Unfortunately relatively little is known about pure behaviour reading, even in humans.

Behaviour reading involves at least three components: segmentation, categorisation and structure extraction. First, it is necessary to segment continuous streams of bodily movements into units of interest. Adult humans can readily impose boundaries on continuous sequences of behaviour (Newtson and Engquist 1976; Zacks et al. 2001; Kurby and Zacks 2008), as can infants (Baldwin and Baird 2001). Segmentation is probably complicated by the fact that units are not performed one after another but can overlap, as when you bat away a fly while squeezing a lemon and talking on the phone. A further complication is that identifying unit boundaries probably involves taking into account the context in which a behaviour occurs (Newtson, Engquist and Bois 1977, p. 847). How could segmentation be achieved? Commencement and completion of a goal or subgoal typically coincide with dramatic changes in physical features of the movements such as velocity (Zacks, Tversky and Iyer 2001). Baldwin and Baird express this idea graphically with the notion of a ‘ballistic trajectory’ that provides an ‘envelope’ for a unit of action (Baldwin and Baird 2001, p. 174). Research using schematic animations has shown that adults can use a variety of movement features to group behavioural chunks into units (Zacks 2004; Hard, Tversky and Lang 2006).

A second component of behaviour reading is categorisation. Adult humans spontaneously label units of behaviour as ‘running’, ‘grasping’, or ‘searching’ (say). This is categorisation: two units which may involve quite different bodily configurations and joint displacements and which may occur in quite different contexts are nevertheless treated as equivalent. How are categories identified in pure behaviour reading? One possibility is that some categorisation processes involve mirroring motor cognition. When a monkey or a human observes another’s action, there are often motor representations in her that would normally occur if it were her, the observer, who was performing the action (see Rizzolatti and Sinigaglia 2008, 2010 for reviews). Further, in preparing, performing and monitoring actions, units of action are represented motorically in ways that abstract from particular
patterns of joint displacements and bodily configurations (e.g. Rizzolatti et al. 1988; Rizzolatti, Fogassi and Gallese 2001; Umiltà et al. 2001; Hamilton and Grafton 2008; Cattaneo, Sandrini and Schwarzbach 2010; Koch et al. 2010; Villiger, Chandrasekharan and Welsh 2010). And, in observation as performance, what is represented motorically depends on contextual factors such as the presence of particular objects on which certain actions might be performed (Gallese and Sinigaglia 2011). Taken together, these findings indicate that one process by which units of action are categorised is the process by which, in other contexts, your own actions are prepared.

A third component of behaviour reading is structure extraction. Many actions can be analysed as a structure of goals hierarchically ordered by the means-ends relation (see figure 2 for an illustration.) A behaviour reader should be able to extract some or all of this structure. But how? Units of behaviour that are all involved in bringing about a single outcome are more likely to occur in succession than chunks not so related. This suggests that transitional probabilities in the sequence of units could in principle be used to identify larger structures of units, much as phonemes can be grouped into words by means of tracking transitional probabilities (Saffran, Newport and Aslin 1996; Gómez and Gerken 2000). We know that human adults can learn to group small chunks of behaviour into larger word-like units on the basis of statistical features alone (Baldwin et al. 2008). A statistical learning mechanism required for discerning such units is automatic (Fiser and Aslin 2001), domain-general (Kirkham, Slemmer and Johnson 2002) and probably present in human infants (Saffran et al. 2007) as well as other species including monkeys (Hauser, Newport and Aslin 2001), pigeons (Herbranson and Shimp

Figure 2: A routine action with a complex, hierarchical structure.
2016), songbirds (Abe and Watanabe 2011) and rats (Toro and Trobalón 2005; Murphy, Mondragón and Murphy 2008). It is therefore plausible that at least some animals’ pure behaviour reading abilities enable them to extract some of the hierarchical structure of actions.

Our primary concern here with behaviour reading is as a potential basis for abilities to track others’ mental states without representing them. But behaviour reading is plausibly important in other ways. In mindreaders, behaviour reading is thought to be useful or even necessary for identifying intentions and other mental states (Newtson, Engquist and Bois 1977, p. 861; Baldwin et al. 2001, p. 708). Behaviour reading may also matter for efficiently representing events (Kurby and Zacks 2008), identifying the likely effects of actions (Byrne 1999), predicting when an event likely to be of interest will occur (Swallow and Zacks 2008, p. 121), and learning through observation how to do things (Byrne 2003). And of course a special case of pure behaviour reading, ‘speech perception’, underpins communication by language in humans.

What are the limits of pure behaviour reading? While little is currently known about pure behaviour reading outside of communication by language, it is perhaps reasonable to assume that structure extraction depends on domain-general learning mechanisms. After all, such mechanisms appear sufficient and there is currently little evidence for domain-specific mechanisms. This assumption allows us to make conjectures about the limits of pure behaviour reading. One limit concerns the kinds of structure which animals of a given species might extract. How complex are the patterns of behaviour which can be extracted? One hypothesis is that, humans at their most reflective aside, animals can usually learn and identify only those patterns which can be captured by a regular grammar (Fitch and Friederici 2012). Another, perhaps more readily exploitable hypothesis concerns non-adjacent dependencies. There is a non-adjacent dependency in my behaviour when, for example, my now having a line of sight to an object that is currently unobtainable because of a competitor’s presence results in me retrieving the object at some arbitrary later time when the competitor is absent. In this case, my retrieving the object depends on my having had it in my line of sight, but there is an arbitrary interval between these events. The hypothesis is that structures involving non-adjacent dependencies are relatively difficult to learn and identify, and that difficulty increases as the number of non-adjacent dependencies increases.¹

¹ Compare de Vries et al. (2012). Of course, whether non-adjacent dependencies are intrinsically difficult depends on the cognitive architecture (Uddén et al. 2012). There is evidence that monkeys (Ravignani et al. 2013) and chimpanzees (Sonnweber, Ravignani and Fitch
To generalise, suppose we assume that extracting structure in pure behaviour reading involves domain-general learning mechanisms only. Then whenever an animal’s abilities to track others’ mental states are underpinned by pure behaviour reading only, her abilities to learn and identify patterns should be subject to domain-general limits. Since birdsongs are discriminable and involve diverse behavioural structures (Berwick et al. 2011), we might take the *Birdsong Limit* as a rough working hypothesis: structures not found in birdsong cannot be extracted in pure behaviour reading.

Although not designed to test such limits, some existing experimental designs involve features which plausibly exclude explaining subjects’ performance in terms of pure behaviour reading only. To illustrate, consider the sequence of events in the ‘misinformed’ condition of Hare, Call and Tomasello (2001, Experiment 1). A competitor observes food being placed \([A]\), the competitor’s access is blocked \([B]\), stuff happens \([X^N]\), food is moved \([C]\), more stuff happens \([Y^N]\), and the competitor’s access is restored \([D]\). Finding evidence that chimpanzees can learn to identify patterns of this form \([ABX^NCY^ND]\) and use them to predict the conspecifics’ behaviours would represent a major discovery.

While it is probably impossible and certainly unnecessary to exclude the possibility that an animal is a behaviour reading demon, it turns out to be quite straightforward (in theory, at least) to exclude the possibility that its actual behaviour reading abilities are what underpin its abilities to track others’ mental states. Even in advance of knowing much about the processes and representations involved in pure behaviour reading, the assumption that structure extraction depends on domain-general learning mechanisms makes it unlikely that the relatively sophisticated abilities of corvids and great apes (say) to track others’ mental states could be underpinned by pure behaviour reading only. We may need to look elsewhere for readily testable and plausible hypotheses on which great apes’ and corvids’ relatively flexible abilities to track others’ mental states do not involve representing mental states.

3. **End False Belief about False Belief**

The overall question for this chapter is, What underpins abilities to track others’ mental states? I started by arguing that answering this question would require evaluating competing hypotheses. To do this requires, first, identifying a theoretically coherent, empirically motivated and readily testable hypothesis on which tracking mental states does not involve representing mental states.

---

2015) can learn patterns involving one non-adjacent dependency.
mental states. But, as we have just seen, it is surprisingly difficult to do this, at least for the most flexible animals. Pure behaviour reading seemed to be the obvious source of such hypotheses; but although the partial account of pure behaviour reading just considered yields readily testable hypotheses, these are already plausibly excluded by existing data. In the absence of an alternative, should we accept, provisionally, that in at least some nonhumans, tracking mental states does after all involve representing them?

There are at least two obstacles to accepting this. The first is a false belief about false belief. The false belief task (Wimmer and Perner 1983) is sometimes regarded as an acid test of mental state representations (see Bennett’s, Dennett’s and Harman’s influential responses to Premack and Woodruff 1978). Awkwardly, chimpanzees and other nonhuman animals have so far mostly thwarted efforts to show that they can track others’ false beliefs (e.g. Call and Tomasello 1999; Kaminski, Call and Tomasello 2008; Krachun et al. 2009, 2010; Marticorena et al. 2011). False belief tasks continue to yield many important discoveries concerning humans (e.g. Milligan, Astington and Dack 2007; Devine and Hughes 2014). But there are reasons to doubt that the false belief task, despite its enormous value, is an acid test of mindreading. First, it is possible to track others’ false beliefs without actually representing them (Butterfill and Apperly 2013). Second, there is evidence that typically developing humans can represent incompatible desires before they can represent false beliefs (Rakoczy, Warneken and Tomasello 2007). Having an ability to track false beliefs is therefore not sufficient for being able to represent beliefs and probably not necessary for being able to represent mental states. So whether we accept that any nonhumans can represent others’ mental states should not hinge on whether they can track false beliefs in particular. As Premack and Woodruff (1978, p. 622) suggested, a false belief task is ‘another arrow worth having in one’s quiver rather than the assured bullseye that the philosophers suggest it is.’

There is a second, more challenging obstacle to accepting that some nonhumans can represent mental states. After claiming that ‘chimpanzees understand … intentions … perception and knowledge,’ Call and Tomasello (2008) qualify their claim by adding that ‘chimpanzees probably do not understand others in terms of a fully human-like belief–desire psychology’ (p. 191). This is true. The emergence in human development of the most sophisticated abilities to represent mental states probably depends on rich social interactions involving conversation about the mental (Slaughter and Gopnik 1996; Peterson and Slaughter 2003; Moeller and Schick 2006), on linguistic abilities (Milligan, Astington and Dack 2007; Kovács 2009), and on capacities to attend to, hold in mind and inhibit things (Benson et al. 2013;
Devine and Hughes 2014). These are all scarce or absent in chimpanzees and other nonhumans. So it seems unlikely that the ways humans at their most reflective represent mental states will match the ways nonhumans represent mental states. Reflecting on how adult humans talk about mental states is no way to understand how others represent them. But then what could enable us to understand how nonhuman animals represent mental states?

The view that tracking mental states involves representing them leaves too many options open, as Call and Tomasello’s nuanced discussion shows. It is not a hypothesis that generates readily testable predictions. We need a theoretically coherent, empirically motivated and readily testable hypothesis on which tracking mental states does involve representing mental states (compare Heyes 2015, p. 321). Identifying such a hypothesis is the second requirement we would have to meet in order to evaluate competing hypotheses about what underpins abilities to track others’ mental states. And to meet this second requirement we must first reject a dogma.

4. Reject the Dogma of Mindreading

Representing physical states, such as the masses or temperatures of things, requires having some model of the physical. Little follows directly from the fact that an individual can represent weight or other physical properties: everything depends on which model of the physical underlies her capacities. And if we ask, ‘What model of the physical characterises her thinking?’, we find that there are multiple, experimentally distinguishable candidate answers (e.g. McCloskey, Washburn and Felch 1983; Kozhevnikov and Hegarty 2001; Oberle et al. 2005). Her physical cognition might be characterised by a Newtonian model of the physical, or perhaps on an Aristotelian model. And it might involve one or another measurement scheme. Perhaps, for example, she distinguishes the weights of things relative to her abilities to move them. Or maybe she relies on a system of comparisons. Different models of the physical and different systems of measurement generate different predictions about the limits of her abilities to track physical events.

Likewise for mental properties. The conjecture that someone can represent mental states—that she is a mindreader, or that she has a ‘theory of mind’—does not by itself generate readily testable predictions. Everything depends on which model of the mental characterises her mindreading.

In asking which model of the mental—or of the physical—characterises a capacity, we are seeking to understand not how the mental or physical in fact are but how they appear from the point of view of an individual or system. Specifying a model is a key part of providing what Marr calls a ‘computational
description’ of a system (Marr 1982). The model need not be thought of as something used by the system: it is a tool the theorist uses in describing what the system is for and broadly how it works.

When an animal is suspected of mindreading we must ask, How does she model the mental? But it will make no sense to ask this question as long as we are in the grip of a dogma. The dogma is that there is just one model of the mental and representing mental states involves the use of that model. Or, more carefully (to accommodate insights about development such as Wellman, Fang and Peterson (2011) offer), the dogma is that all models of the mental comprise a family in which one of the models, the best and most sophisticated model, contains everything contained in any of the models.

The dogma implies that only animals whose capacities approximate those humans exhibit in talking about mental states can be mindreaders. In rejecting the dogma we also remove any reason to make this assumption. Different mindreaders may rely on different, incommensurable models of the mental and different schemes for distinguishing mental states. Mindreading in other animals need not be an approximate version of mindreading in adult humans any more than medieval physical thought approximates contemporary physical theories.

To see how strange endorsing the dogma would be, contrast the mental with the physical. The briefest encounter with the history of science reveals that there are several models of physical phenomena like movement, mass and temperature. Some models are more accurate but also relatively costly to apply, while others are easier to apply but less accurate. And there appear to be different kinds of physical cognition which involve different—and incommensurable—models of the physical.² It would be astonishing to discover that there is one privileged model such that all physical cognition can be understood by reference to that particular model. The dogma of mindreading tacitly guides discussion only because, by contrast with the rich array of flawed theories of the physical, there is a scarcity of scientific theories of desire, intention and the other mental phenomena that animals can track. But this scarcity can be alleviated.

5. Construct Models of the Mental

To understand what underpins abilities to track others’ mental states, we need to construct models of the mental that we can use to formulate and test

hypotheses about different mindreaders’ capacities. And, as in the physical case, we don’t want models of the mental that are maximally accurate. Instead we seek models good enough to underpin successful belief tracking in a useful range of situations and simple enough that their realisation need demand little in the way of conceptual sophistication and cognitive resources.

But what is a model of the mental? On a widely accepted view, mental states involve subjects having attitudes toward contents (see figure 3). Possible attitudes include believing, wanting, intending and knowing. (Not every model of the mental need include these attitudes.) The content is what distinguishes one belief from all others, or one desire from all others. The content is also what determines whether a belief is true or false, and whether a desire is satisfied or unsatisfied.

There are two main tasks in constructing a model of mental states. The first task is to characterise some attitudes. This typically involves specifying their distinctive functional and normative roles by developing a theory of the mental. The second task is to find a scheme for specifying the contents of mental states. This typically involves one or another kind of proposition, although some have suggested alternatives (e.g. Braddon-Mitchell and Jackson 1996, p. 163).

One model of the mental is specified by minimal theory of mind (Butterfill and Apperly 2013, 2016), which repurposes a theory offered by Bennett (1976) in building on insights offered by Gómez (2009) and Whiten (1994) among others. This theory—or, rather, series of nested theories—specifies states with stripped down functional roles whose contents are distinguished by tuples of objects and properties rather than by propositions. These features ensure that, although minimal theory of mind is capable of underpinning abilities to track mental states including false beliefs in a limited but useful range of situations, realising minimal theory of mind need involve little conceptual sophistication and cognitive resources.

---

3 For examples, see Bratman (1987) on intention or Velleman (2000, chapter 11) on belief.
sophistication or cognitive resources.

The construction of minimal theory of mind enables us to specify some simple models of the mental, and to generate testable hypotheses about how mindreaders model minds. One such hypothesis concerns infant humans. The hypothesis is that a minimal theory of mind describes the model of the mental which underpins mindreading processes in these subjects. A key prediction of this hypothesis has so far mostly been confirmed (Fizke, Butterfill and Rakoczy 2013; Low and Watts 2013; Low et al. 2014; Wang, Hadi and Low 2015; see Low et al. 2016 for an overview), although there are some apparently contrary findings too (Scott, Richman and Baillargeon 2015; Kampis and Kovács, May 26, 2016) and some theoretical objections (e.g. Carruthers 2013, 2015a, 2015b; Christensen and Michael, forthcoming). A minimal model of the mental might capture how minds appear from the point of view of some mindreading processes in some humans.

Consider a related hypothesis about nonhuman animals: abilities to track mental states in some nonhumans are underpinned by capacities to represent mental states which involve a minimal model of the mental. (This hypothesis was suggested by Apperly and Butterfill 2009.) This hypothesis avoids objections arising from views on which nonhumans have representational powers whose emergence in human development involves linguistic abilities and communicative exchanges. It also generates testable predictions about the limits of mindreading in nonhumans, including predictions which distinguish hypotheses about minimal theory of mind from hypotheses about pure behaviour reading. And there is already a hint that one such prediction is correct (see Karg et al. 2016; they don’t mention this, but a signature limit of minimal mindreading is inability generally to do level-2 perspective taking).

6. Distinguish Kinds of Mindreading Processes

Testing the hypothesis about minimal theory of mind in nonhumans is especially challenging, however. In humans, minimal theory of mind generates otherwise unexpected predictions because humans typically develop abilities to track mental states beyond anything minimal models of the mental could underpin (e.g. Rakoczy et al. 2015). But nonhumans’ abilities to track others’ mental states are currently thought to be so limited that fewer of the predictions generated by minimal theory of mind are otherwise unexpected. It may appear that invoking minimal models of the mental will amount merely to telling another ‘just so’ story.

Note, however, that a model of the mental is supposed to be just one ingredient in a theory of mindreading. A model of the mental is the basis
for a computational description (in Marr’s sense) of a mindreading system. And although essential, no computational description by itself provides a full theory. We will also need hypotheses about the kinds of representations and processes mindreading involves.

Such hypotheses may be motivated by research with human subjects. There is evidence for two (or more) mindreading processes in adult humans (e.g. Wel, Sebanz and Knoblich 2014; Furlanetto et al. 2015). One mindreading process is more automatic than the other in the sense that whether or not it occurs is, to a significant degree, independent of your tasks and intentions. By distinguishing the contributions of different kinds of mindreading processes to performance, we may succeed in identifying processing characteristics distinctive of automatic mindreading (e.g. Qureshi, Apperly and Samson 2010; Todd et al. *under review*). It may turn out, for example, that automatic and nonautomatic processes in humans can be distinguished by their sensitivities to time constraints, to interruptions in perceptual processing or to cognitive load, or by their effects on motor control versus other kinds of planning. Discovering signature processing characteristics of automatic mindreading in humans could allow us to test the (currently unsupported) conjecture that some nonhumans’ capacities to track others’ mental states involve processes similar to those that underpin automatic mindreading in humans.

This conjecture about processes can be linked to the earlier hypothesis about models (see section 5 on page 10). There are theoretical and empirical reasons for hypothesising that minimal models of the mental characterise automatic mindreading only (Apperly and Butterfill 2009; Low et al. 2016). If we can be confident that this hypothesis holds for adult humans, confirming the conjecture linking automatic processes in humans and nonhumans would support the view that a minimal model of the mental characterises nonhumans’ capacities to track mental states.

The suggestions in this section are wildly speculative, of course. But they illustrate a point that is not speculative at all. Constructing models of the mental does enable us to identify theoretically coherent, empirically motivated and readily testable hypotheses on which representations of mental states underpin abilities to track them. But this is just a first step towards understanding varieties of animal mindreading, one that opens the way for further theorising about the kinds of processes that underpin mindreading.

7. Conclusion

What underpins abilities to track others’ mental states? To answer this question we would need at least two competing hypotheses to evaluate. First,
we would need a theoretically coherent, empirically motivated and readily testable hypothesis on which tracking mental states does not involve representing mental states. Although no such hypothesis currently exists (Halina 2015, p. 486; Heyes 2015, p. 322), there is a body of research on behaviour reading from which a theory capable of generating readily testable predictions might be derived (see section 2). Second, we would need a readily testable hypothesis on which representations of mental states underpin abilities to track them. The construction of minimal theory of mind enables us to generate one such hypothesis (see section 5).

How plausible are these hypotheses? Even in advance of having a theory of behaviour reading, we might assume that extracting structure in behaviour reading depends on domain-general learning mechanisms only. Given this assumption, it seems unlikely that the relatively flexible mental-state-tracking abilities of corvids and nonhuman great apes are underpinned by behaviour reading only (section 2). This may motivate the search for alternative theories on which tracking others’ mental states does not involve representing them. It may even justify accepting, provisionally at least, that at least some animals other than humans represent mental states.

To accept this is not yet to have a theory about mindreading capable of generating readily testable predictions, however (see section 3). Understanding abilities to track others’ mental states is not simply a matter of categorising them as involving or not involving representations of mental states. Instead we need to understand how different mindreaders model the mental.

It may be tempting to rely on a pretheoretical, commonsense understanding of what mental states are in formulating and testing hypotheses about what underpins abilities to track others’ mental states. But a model implicit in adult humans’ talk about mental states is unlikely to be the model of the mental that characterises chimpanzees’ or corvids’ mindreading (section 4). Because different mindreaders may rely on different, incommensurable models of the mental and different schemes for distinguishing mental states, we need to identify models of the mental that we can use to generate readily testable hypotheses about different mindreaders’ capacities. The construction of minimal theory of mind illustrates how to do this.

The hypothesis that abilities to track mental states in some nonhumans including great apes and corvids are underpinned by capacities to represent mental states which involve a minimal model of the mental has three things going for it. It makes precise what researchers should care about in asserting that animals other than humans can represent others’ mental states. It isn’t already known to be false, and there is even a hint that its predictions are
correct (section 5). And it has no theoretically coherent, empirically motivated and readily testable competitors—at least not yet. So if a minimal model of the mental doesn’t characterise any nonhuman animals’ abilities to track other mental states, what does?
References


Burkart, Judith Maria, and Adolf Heschl. 2007. ‘Understanding visual access in common marmosets, Callithrix jacchus: perspective taking or behaviour reading?’ *Animal Behaviour* 73, no. 3 (March): 457–469. doi:10.1016/j.anbehav.2006.05.019.


Fizke, Ella, Stephen A. Butterfill and Hannes Rakoczy. 2013. ‘Toddler’s understanding of false beliefs about an object’s identity’. In *Poster presented at the Biennial Meeting of the SRCD, Seattle*.


Kampis, Dora, and Ágnes Melinda Kovács. May 26, 2016. ‘Fourteen-month old infants track multiple objects and object identity when they represent others’ beliefs’. In *Talk at ICIS, New Orleans, LA*.


Melis, Alicia P., Josep Call and Michael Tomasello. 2006. ‘Chimpanzees (Pan troglodytes) conceal visual and auditory information from others’. Journal of Comparative Psychology 120 (2): 154–162. doi:10.1037/0735-7036.120.2.154.


