Replies to Three Commentaries on Minimal Theory of Mind

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

November 8, 2013

Contents

1 Introduction **1**

2 Predictions: How are we doing so far? (HR) **2**
2.1 Signature limits 2
2.2 Multiple steps in acquiring non-minimal theories of mind . . 3
2.3 The rabbit/carrot and the rescue/sting 4

3 Mindreading vs behaviour reading (SS) **5**
3.1 Behaviour reading: two approaches 5
3.2 How minimal theory of mind differs 7

4 Tracking vs Representing (TZ) **10**

5 Intervening variables (TZ) **11**

6 Computational Descriptions vs Mechanisms (TZ, SS) **13**

7 Minimal theory of what? (TZ) **14**

8 Infants vs Scrub-jays (SS) **15**

9 Conclusion **16**

1. Introduction

We are grateful to Hannes Rakoczy, Shannon Spaulding and Tadeusz Zawidzki for three illuminating and very helpful critical commentaries. Here we report some of what we have learned from them and answer the objections.

2. Predictions: How are we doing so far? (HR)

There are multiple theories of mind, including a minimal one and a more sophisticated one associated with adult humans at their most reflective. In our paper we aimed to describe how to construct a minimal theory of mind, and we offered some conjectures. One of these conjectures was that mindreading in infancy involves minimal theory of mind. Hannes Rakoczy's commentary introduces new evidence in support of that conjecture.

This is very welcome. Rakoczy also identifies a gap in our reasoning about how the conjecture about infants mindreading could be tested. We agree with Rakoczy that we made a mistake, and we're delighted that a study by him and his colleagues enables us to fix our mistake. In what follows we describe exactly where we went wrong and how the fix enables us to correct this mistake.

2.1. Signature limits

How can we tell, for a particular type of subject and task, which theory of mind is being used? And how can we tell whether the same theory of mind is being used in two different cases? By means of signature limits.

But what are signature limits? To answer this question, first consider a case where the use of signature limits is well established: physical cognition. Suppose you are interested in a particular cognitive phenomenon, representational momentum (say), and you want to know what sort of theory of the physical underpins it. Does this phenomenon reflect Newtonian principles or a theory on which objects have impetus? Here's how you decide. First, think about the Newtonian and impetus theories. In what situations do the theories make different predictions? Take one of these situations, and let it be one in which only the impetus theory makes an incorrect prediction. This prediction is a signature limit of the impetus theory. Now consider the conjecture that the cognitive phenomenon under study (representational momentum, or whatever) reflects impetus theory. If that conjecture is correct, the signature limit of the impetus theory should be revealed in the cognitive phenomenon. By contrast, the conjecture that the cognitive phenomenon under study reflects Newtonian principles makes no such prediction. Accordingly we can test the conjecture that the cognitive phenomenon reflects an impetus-based theory of the physical by testing predictions derived from signature limits of the impetus theory (Kozhevnikov & Hegarty 2001).

In general, a signature limit of a theory is a set of predictions of the theory which are incorrect, and which are not predictions of other theories under consideration. As the study just mentioned (Kozhevnikov & Hegarty 2001) beautifully illustrates, identifying signature limits can make it possible to test conjectures about which theory underpins a given cognitive phenomenon.

We applied this general idea to the case of mindreading. Just as there are multiple theories of the physical, so also there are multiple theories of mind. Some of these theories are more accurate but relatively complex, others are simpler but less accurate. It would be a mistake to assume that there is just one theory of the physical in terms of which we must understand every cognitive process. Similarly, there is no good reason to suppose that there is just one theory of mind in terms of which all mindreading must be understood. Where there is mindreading we can ask which theory of mind is in play. And signature limits provide a way to answer this question.

2.2. Multiple steps in acquiring non-minimal theories of mind

Hannes Rakoczy in his commentary identifies a gap in our reasoning about signature limits. We claimed that one signature limit of minimal theory of mind is a set of predictions involving false beliefs about the identities of objects (as illustrated by Lois Lane's mistakes about Superman and Clark Kent). Put roughly, minimal theory of mind makes incorrect predictions about what Lois Lane, whose beliefs entail that there are two individuals where in fact there is just one, will do. Now for these predictions to be a signature limit, it is not enough for them to be incorrect. They must also not be predictions of any other theory under consideration. We failed to properly evaluate this additional requirement in our paper because we were making a simplifying assumption. The simplifying assumption was that the only other theory of mind under consideration was the one true, perfectly complete theory of mind. Of course things are not this simple, and not only because there is no such theory of mind. Things are not this simple because acquiring a non-minimal theory of mind is a protracted process involving multiple steps.

We know that acquiring a non-minimal theory of mind involves multiple steps in part because Rakoczy himself, in earlier work with Warneken and Tomasello (2007), showed that three year olds who fail a standard false belief task nevertheless understand the possibility of incompatible desire. So we cannot think of the acquisition of a non-minimal theory of mind as the sudden appearance of an ability to apply the one true, perfectly complete theory of mind. Rather, as several others have suggested, we should think of the acquisition of a non-minimal theory of mind as involving a succession of theories of mind.

Rakoczy's commentary has made us realise that this affects our claim about the signature limits of minimal theory of mind. In a developmental context, for a prediction to be a signature limit of minimal theory of mind it must not be a prediction of any of the theories of mind which appear in the course of acquiring a non-minimal theory of mind. (This is Rakoczy' claim (b) on page 2 of his commentary.)

So far Rakoczy has provided a very welcome theoretical clarification. But every silver lining has a cloud. Consider three- and four-year-olds' progress in acquiring a non-minimal theory of mind. Is there a stage at which they have a theory of mind which enables them to track false beliefs about location (and perhaps other properties) but not about identity? As Rakoczy notes, there is some evidence which might appear to indicate there is such a stage and—prior to the efforts of him and his collaborators—no evidence to the contrary. So there was a huge gap in our justification for the claim that false beliefs about identity are a signature limit of minimal theory of mind in this developmental context.[1]

2.3. The rabbit/carrot and the rescue/sting

Fortunately for us, Rakoczy et al. (2013) provide some exciting new findings which, as a side-effect, fill the gap in our reasoning. These findings indicate that, when three- to six-year-olds' are tested using some explicit measures, understanding false beliefs about location (say) is tied to understanding false beliefs about identity. So in the case of infant mindreading, the other relevant theories of mind do not make the incorrect predictions involving false beliefs involving identity which follow from minimal theory of mind. Thus Rakoczy rescues our favourite signature limit.

We were just about to offer him our hand in marriage when one of us felt a sting in the rescue. In saving the conjecture about infant mindreading involving minimal theory of mind, Rakoczy appears to have shown that one of us was wrong about children's developing understanding of intensionality. We are still thinking about this.

Rakoczy's new findings also allow us to address a problem raised by Tadeusz Zawidzki. Rakoczy's experiments involve what is in fact a single object with two aspects. Everything depends on the subjects knowing this. Zawidzki notes that subjects might see things differently: they might, for example, think of the single object as two objects. In this case, what we as theorists interpret as a test concerning false beliefs about identity would collapse into a test concerning false beliefs about location. We agree with Zawidzki that this is a potential problem. Indeed, it belongs to a family of problems we have encountered in designing experiments. Two things help in overcoming these problems. First, in some cases we can draw on evidence about infants' understanding of identity (e.g. Cacchione et al. 2013), avoiding the need to rely on mere guesses about their perspective. Second, Zawidzki's problem does not arise when the evidence indicates that subjects fail to understand false beliefs involving identity. It only arises when the

[1] Note that this gap in our reasoning should not apply in the case of adult participants because we know that adults can track false beliefs involving mistakes about identity. Failures to track false beliefs involving mistakes identity in adults' automatic mindreading reflect a signature limit of minimal theory of mind (compare Low & Watts 2013).

evidence indicates that subjects do understand false beliefs about identity. So the problem arises in Rakoczy and colleagues' 'explicit' studies, which provide evidence that, when asked to make verbal predictions, subjects find it no harder to take into account another's false belief about identity than a false belief about location. Fortunately in studies like these it is relatively straightforward to check subjects' understanding of the objects. So far, then, the problem can be overcome.

3. Mindreading vs behaviour reading (SS)

Where does behaviour reading end and mindreading begin? We conjecture that infants and others sometimes track others' beliefs by virtue of representing their registrations. (Registration is a state that we constructed in such away that, in a limited but useful range of situations, registration and belief are correlated.) Shannon Spaulding challenges us to say how our conjecture differs from 'the behavioural account' of infants' competence.

3.1. Behaviour reading: two approaches

On confronting this challenge we immediately hit an obstacle. The obstacle is that there are different approaches to characterising behaviour reading. One approach is based on investigations of how humans and others are able to extract patterns of behaviour from continuous bodily movements independently of any knowledge of agents' goals and mental states. This can be understood as a matter of solving three problems. First, there is the problem of identifying basic chunks of behaviour. This is thought to involve sensitivity to a variety of movement features (Baldwin et al. 2001; Zacks 2004; Hard et al. 2006). Second, there is the problem of working out which sequences of chunks form larger units of interest. There is some evidence that transitional probabilities in the sequence of chunks could in principle be used to identify significant units, much as phonemes can be grouped into words by means of tracking transitional probabilities (Saffran et al. 1996; Gómez & Gerken 2000). Further, Baldwin et al. (2008) demonstrated that adults can learn to group small chunks of behaviour into larger word-like units on the basis of statistical features alone. Then, third, there is the problem of identifying hierarchical patterns in behaviour, that is, patterns connecting non-consecutive units. To illustrate, consider observing someone tasked with making a burger. Their task involves several steps whose order is only loosely constrained, where some of these steps can be omitted or replaced (cheese burger, chilli burger) and where steps can be interspersed with irrelevant actions (answering the phone or avoiding a projectile). This is why grouping together all and only the behavioural units involved in making a burger involves discerning hierarchical structure. In general, discerning hierarchical structure appears to be cognitively demanding (Cohen et al. 1990). There appear to be limits on the types of structure that even human adults can learn (Newport & Aslin 2004), and there is no convincing evidence that nonhumans can learn arbitrary hierarchical patterns (Corballis 2007). Insofar as it is possible at all, learning hierarchical patterns may depend on comparing what happens on different occasions and relying on changes in motion features (Byrne 2003).

This first approach to characterising behaviour reading has been much ignored in research on social cognition. This is unfortunate because mindreading of any kind surely depends on behaviour reading. Also, better understanding the mechanisms that make behaviour reading possible (and their limits) may well have consequences for understanding how mindreading emerges in evolution and development.

If we approach behaviour reading in this first way, answering Spaulding's challenge to distinguish minimal theory of mind from behaviour reading is straightforward. The fact that minimal theory of mind involves identifying to which goals actions are directed is already sufficient to distinguish it from behaviour reading (which starts from continuous bodily movements and makes no use of the notion of goal-directed action). In addition, minimal theory of mind is a causal theory (see Section 5 on page 11) whereas the aim of behaviour reading is to extract patterns. Finally, behaviour reading does not involve states which, like registration, are defined by their functional roles.

That was much too easy. Clearly Spaulding does not have this approach to behaviour reading in mind. What does she envisage?

There is an alternative approach to characterising behaviour reading. This alternative approach is unconstrained by investigations of mechanisms. Instead the idea is to start with a set of propositions concerning the conditional probability that a particular type of goal-directed action will occur given that certain other goal-directed actions have occurred. Behaviour reading is then the application of these propositions to predicting (and maybe influencing) which goal-directed actions will occur in the future. In facing Spaulding's challenge we have to think of behaviour reading in this second way.

As an aside, note that we should be cautious in considering hypotheses about behaviour reading as characterised in this second, less constrained way. There is quite good evidence that at least some animals (adult humans) do represent facts about others' mental states and are thereby able to do things that they might not otherwise be capable of. By contrast, there is remarkably little evidence that any animals represent facts about conditional probabilities linking goal-directed actions and are thereby able to do things. If we take the second approach to characterising it, there may turn out to be no such thing as behaviour reading.

This is not to say, of course, that Spaulding's challenge is not a good one. Even if there turns out to be no such thing as behaviour reading, it might still be informative theoretically to contrast it with minimal theory of mind.

3.2. How minimal theory of mind differs

How can we answer Shannon Spaulding's challenge to distinguish minimal theory of mind if we think of behaviour reading as the use of behavioural rules—that is, propositions concerning the conditional probabilities linking goal-directed actions—to make predictions?

An initial response to this challenge would be to note that registration is a state, not a behaviour. If behaviour reading involves applying rules concerning the conditional probability that one type of behaviour will occur given that another has occurred, then our approach is clearly different.

But this response is mistaken. After all, behavioural rules can surely apply not just to behaviours, but also to bodily configurations and their relations to the environment. Thus Povinelli and Vonk in formulating behavioural rules invoke a relation that obtains between an agent and an object when the agent is 'oriented to' that object (Povinelli & Vonk 2003). So our conjecture about minimal theory of mind does not differ from a behavioural rules account just in virtue of invoking states.

How else could we answer Spaulding's challenge to distinguish our conjecture from one involving behavioural rules only? Tadeusz Zawidzki implies that we construe registrations as 'internal, unobservable states'. If this were our view, we might try to distinguish our conjecture by saying that registrations are off-limits to proponents of behavioural rules because, unlike bodily configurations, they are internal, unobservable states. Certainly, some researchers have attempted to contrast behavioural states with mental states along these lines. But this is not our view. It is not our view because we are not yet fully convinced that mental states are unobservable (compare Smith forthcoming), nor that all behaviours are observable. And, more simply, it is not our view because we don't have the first idea how to provide an empirically motivated distinction between the observable and the unobservable, nor between the internal and external. Tadeusz Zawidzki suggests we should place no weight on such contrasts, and we agree.

At this point we despair of finding a quick and easy answer to Spaulding's challenge. But we do have a long and complex one. To answer the challenge we need to ask, What is a mental state? To start with the least controversial, let us assume that a mental state involves three things: a subject (Ayesha or Henry, for example), an attitude and a content (see figure 1 on the next page). Familiar attitudes include believing, wanting, intending and knowing. The content is what distinguishes one belief from all others,

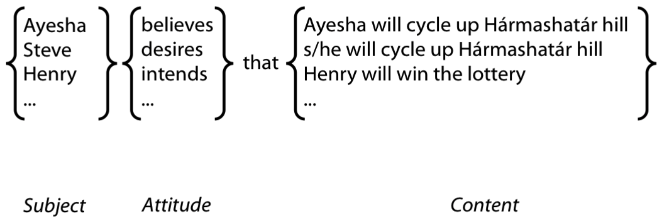
Ayesha
Steve
Henry
...
believes
desires
intends
...
that
Ayesha will cycle up Hármashatár hill
s/he will cycle up Hármashatár hill
Henry will win the lottery
...
Subject
Attitude
Content

Figure 1: Mental states involve subjects having attitudes toward contents.

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. Someone who wanted to construct a model of the mental would need to do at least two things. She would need to characterise some attitudes, which typically involves specifying their distinctive causal and normative roles. And she would need to identify a scheme for distinguishing among contents; this typically involves one or another kind of proposition, although some have suggested other abstract entities including map-like representations.

Theorists occasionally appear to take for granted that there is a unique, well-supported, fully accurate, normative model of the nature of mental states, and that what research on mindreading tests is how well or badly different groups—infants, chimps, scrub jays and the rest—can apply this model in tracking others' mental states. But we doubt that anyone would, on reflection, endorse this view. After all, there is longstanding, substantial debate about the nature of mental states. This debate concerns both how the attitudes should be characterised (e.g. whether belief is characterised by appeal to norms or not) and how the contents of beliefs should be distinguished (e.g. by appeal to one or another kind of proposition). The idea that we theorists have a single well-supported and fully accurate characterisation of the mental states is, as things stand, pure fantasy.

This observation is the inspiration for our approach. Like the Wombles who are famous for making good use of things people leave behind, we are recycling a flawed, much too simple attempt to characterise belief. On this attempt, the attitude is specified by the five principles we give in our paper, and the scheme for distinguishing among contents involves relations between objects and locations (or other properties) rather than propositions. (But note, incidentally, that some varieties of propositions are relations, just more complex relations than those we consider.) As an attempt to characterise belief, the attempt is almost but not entirely unsuccessful. We can therefore turn things on their head. Let us use 'registration' as a term for that state, whatever it is, that the account does characterise. This allows us to make good use of what was formerly a failed attempt to characterise belief: we can use it as a conjecture about the belief-like states that might be represented as proxies for belief.

In characterising registration, we give an account of its functional role by specifying principles relating it to goal-directed action, and we give an account of how the contents of registrations can be distinguished by appeal to relations between objects and locations.[2] Of course registration is simple in that its functional role is not elaborate and in that it is only possible to distinguish between the contents of registrations in a relatively crude way. But registration is not thereby any less of a mental state than full-blown belief or desire (whatever exactly those turn out to be).

But is registration a mental state? On some philosophers' views, belief differs from registration only in that its functional role is more complex and in that specifying its content requires something more sophisticated than a two-place relation. These philosophers' views provide no reason to deny that registration is a mental state. By contrast, other philosophers' views entail that an account of belief requires a special ingredient not needed for characterising registration, perhaps consciousness or normativity. These philosophers' views may provide reasons to deny that registration is a mental state.

What can we conclude? First, that there are views about the nature of belief which, if true, would support the claim that registration is a mental state. And, second, that whether registration is actually a mental state probably hinges on some quite deep controversies.

Our simple-minded impulse is to duck these controversies. We can salvage two things that matter. First, we might have done something towards showing how to construct a minimal theory of mind even if it turns out that the principles we offered aren't quite sufficient and an additional ingredient is needed to introduce genuinely mental states (whatever exactly these are). Second, standard false belief tasks (e.g. Wimmer & Perner 1983) do not test for understanding things like the normative aspects belief. This suggests that on the notion of a mental state implicit in much theory of mind research, registrations are mental states.

So how does minimal theory of mind differ from the use of behavioural rules? It involves ascribing registration, which is a mental state or something very like one.

[2] Incidentally, Spauding writes that 'representing registration consists in representing an agent's recent encounter of an object.' In fact our principles do not provide such a tight link between encountering and registration. The fourth principle allows for the possibility that someone registers an object at a location despite never having encountered it. As we mention in the paper, this is necessary given findings such as those of Träuble et al. (2010).

4. Tracking vs Representing (TZ)

We argue that representing others' beliefs is just one way of tracking them, and that there are other belief-like mental states representing which would enable you, within limits, to track beliefs. This was the point of characterising the notion of registration: in a limited but useful range of situations, someone could track beliefs by virtue of representing registrations.

In making this argument we are depending on a distinction between representing something and tracking it. Consider mass, for instance. Some individuals can represent the masses of things, and thereby track their masses. But, plausibly, there are individuals who cannot represent the masses but only the weights of things. Fortunately, mass and weight are closely related around here. This means that, around here at least, representing objects' weights can be a way of tracking their masses.

Can we be more precise about the notion of tracking? We offered this working definition: To say that someone can *track* others' beliefs is to say that she has an ability which exists in part because exercising it brings benefits obtaining which depends on exploiting or influencing facts about others' beliefs.

Tadeusz Zawidzki notes that there are theories of representation on which the distinction between representing and tracking cannot be drawn in the way we want to draw it. A feature of these theories is that the following inference is valid:

1. S can track beliefs by virtue of representing registrations.

Therefore:

2. S's representations of registrations are representations of beliefs.

Zawidzki challenges us to say what blocks the inference. Let us put the challenge this way. Suppose an individual is able to track others' beliefs by virtue of representing registrations. What more would be needed to make it the case that she is representing beliefs?

This is a good question but we are reluctant to answer it because the same question arises for many things other than belief. It also arises for mass and toxicity, for instance. Given that even quite good philosophers' attempts to identify general truths about representation have met with hardly any success, there may not be much more that we can usefully say to contrast representing with tracking. Even so, perhaps one observation about representation in general will help. Which content a particular representation has is linked both to the conditions under which the representation does or might exist (the input conditions) and also to the conditions under which the ways in which this representation does or might influence action would be beneficial to its subject (the output conditions). Zawidzki may be right that theories of representation which focus exclusively on output conditions are unlikely to succeed in distinguishing representing from tracking. (And a converse problem results from focussing exclusively on input conditions: what ought to be representations of distal phenomena end up as representations of their proximal correlates.) This tells us that an adequate theory of representation will need, somehow, to balance input and output conditions in specifying what is represented. Doing so will yield the required distinction between tracking and representing. But we confess that we have no idea how this might be done.

5. Intervening variables (TZ)

Tadeusz Zawidzki notes that we didn't properly explain the claim that anyone using all the principles we invoked in constructing a minimal theory of mind is thereby treating the belief-like state we characterise, registration, as an intervening variable.

We should start by explaining how we understand the notion of an intervening variable. Suppose you have a causal model of a situation involving two binary variables: the weather can be cold or not and the street can be safe or dangerous for walking. Your model says that cold weather increases the probability that the street will be dangerous for walking (see figure 2 on the following page). In this first model there are no intervening variables. But suppose you now switch to a second, revised model. On this model there is a new variable: ice can be present or absent. This new variable intervenes between cold weather and dangerous streets. That is, cold weather increases the probability that ice will be present, and the presence of ice in turn increases the probability that the street will be dangerous for walking (see figure 3 on the next page). To our simple-minded, ahistorical way of thinking, the state of the ice (present or absent) is an intervening variable. This is true even if the state of the ice is no less observable than any of the other variables in the model. So in invoking the notion of an intervening variable, we are ignoring some of the historical context in which this notion was introduced to focus on what we take to be the core idea.

Now we can explain why anyone using all the principles we invoked in constructing a minimal theory of mind is thereby treating registration as an intervening variable. The principles have three consequences. First, registration is only defeasibly linked to encountering; someone using minimal theory of mind may infer that an individual registers an object at a location even though that individual has never encountered the object at that location. Second, registration is not reducible to action: in some cases it is possible (according to the principles) to register an object at a location just in virtue

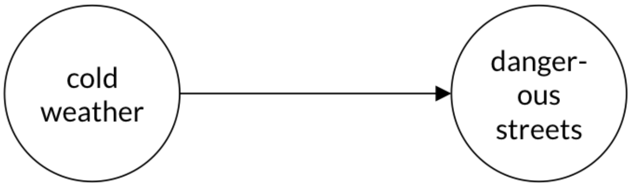

Figure 2: No intervening variable.

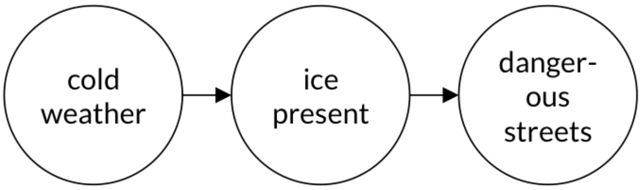

Figure 3: The presence of ice is an intervening variable.

of encountering it there and so without ever performing a goal-directed action whose goal specifies that object. And, third, encountering causally influences registration which in turn causally influences goal-directed action. The first two consequences tell us that registration cannot be identified with either encountering or acting. And the third consequence tells us that registration intervenes between encounters and actions.

Zawidzki might object that the notion of an intervening variable (even as we understand it) only makes sense in the context of a causal model. Why suppose that anyone using minimal theory of mind treats registrations as causal influences on actions? Here we should make a concession. If someone were using minimal theory of mind merely to predict others' actions, there may be no reason to suppose that they are treating registrations as causes. But the principles we offer can support interventions as well as predictions. Accordingly, some individuals may use minimal theory of mind in order to manipulate others' actions, not only to predict them (compare Knudsen & Liszkowski 2012 and Dally et al. 2005). And, inspired by Woodward (2003), we take this to amount to treating the variables as causally related.

6. Computational Descriptions vs Mechanisms (TZ, SS)

Tadeusz Zawidzki is right to call for care and clarity on the relationship between claims at a computational level of description and claims about implementation. He's also right that our wider project involves both. The two-systems view advanced in Apperly & Butterfill (2009) and Apperly (2010) is a view about implementation—that is, about characteristics of the cognitive processes involved in tracking beliefs and other mental states. This view is theoretically motivated by the observation that tracking beliefs and other mental states needs to be both highly flexible and cognitively efficient, which are difficult characteristics to reconcile in a single system. And the two-systems view is empirically motivated by mounting evidence for conclusions that would otherwise be contradictory: that tracking beliefs and other mental states can be both cognitively effortful on the one hand and automatic and efficient on the other.

Our two-systems view makes only a relatively abstract claim about implementation and leaves much open. Importantly, our view is agnostic about whether there are two or more systems. (We did say this at the outset, but it should have been clearer.) As comparison with the case of number indicates, where some researchers propose distinct cognitively efficient systems for representing analogue magnitudes and for individuating small sets (e.g. Carey 2009), it is coherent to suppose that cognitively efficient mindreading involves multiple systems.

Claims about implementation constrain and are constrained by claims at a computational level of description. Anyone who has a two-systems view about abilities to track beliefs and other mental states needs to explain how cognitive efficiency could be achieved, and how flexibility could be achieved. In both cases, these explanations will involve claims at a computational level of description. Facts about what a system represents constrain how cognitively efficient it could be, and how much flexibility it could support.

Minimal theory of mind is a systematic attempt to characterise a form of mindreading simple enough to be achieved by a cognitively efficient system. It is couched at a computational level of description, and could be implemented in a variety of ways. For instance, it could be implemented with map-like representations to capture others' beliefs about the locations and properties of significant objects. So in our overall project, implementation will come in twice: once with the idea that there are two (or more) systems, and again when we ask how minimal theory of mind is implemented. Minimal theory of mind, then, is not an amendment or alternative to our two-systems account, but rather an attempt to deliver on our earlier promissory note that cognitively efficient mindreading is possible.

7. Minimal theory of what? (TZ)

Why is what we call 'minimal theory of mind' a theory of mind? Why isn't it a version of Daniel Dennett's intentional stance or a theory of rational action? Tadeusz Zawidzki challenges us to answer these questions.

Elsewhere (2012), Zawidzki proposes that reflection on Dennett's intentional stance might enable us to better understand what mental states are from the perspective of a mindreader. This is an illuminating idea and we half agree. In particular, we agree that being a mindreader does not necessarily involve much in the way of metaphysical commitments. But we disagree about causation. Dennett emphasises how mental state ascription is bound up with predicting actions but ignores or denies its importance for manipulating others' actions. We guess that theories of mind, whether minimal or not, are typically used not only to encode and predict but also to manipulate others' actions, as in preventing a competitor from stealing food by ensuring she lacks a correct registration of its location. And this, we think, calls for theories of mind that treat mental states as causes of actions (see section 5 on page 11). This is one reason why we tentatively reject the claim that mindreading can be described as taking the intentional stance in Dennett's and Zawidzki's sense.

A further concern applies to both the intentional stance and a theory of rational action. As Zawidzki emphasises (2011; 2012), using either of these involves identifying what it would be most rational to do in different situations. But what it is most rational to do can depend on arbitrary features of a situation in arbitrarily complex ways. So it may be that the way these two, the intentional stance and the theory of rational action, appeal to rationality makes them unsuitable as tools for explaining how abilities to track beliefs could be cognitively efficient. On the other hand, as Zawidzki observes (2011, p. 492), the conjecture that infants use a theory of rational action generates predictions which have been confirmed.

Zawidzki (2011) has also defended the view that infants are not mindreading but rather applying a theory of rational action along the lines of that described by Gergely & Csibra (2003). There are two questions we should answer here. One is whether our minimal theory of mind is really different from Zawidzki's enhanced theory of rational action. This is something we have already indirectly touched on in Section 3.2 on page 7. So let us now take for granted that the differences between the two theories, our minimal theory of mind and Zawidzki's enhanced theory of rational action, are clear enough. The second question is, Why favour the conjecture that infants (or anyone else) uses minimal theory of mind over the conjecture that they use an enhanced theory of rational action?

Let us step back from minimal theory of mind for a moment. Our two-systems view says that humans' abilities to track beliefs and other mental states involve two (or more) systems. Now we suppose that tracking mental states does not necessarily involve representing mental states (see Section 4 on page 10). So we agree with Zawidzki that being committed to multiple systems does not force us to think of cognitively efficient systems as involving representations of mental states. Further, he has argued that a theory of rational action could support abilities to track beliefs in an interesting range of cases (Zawidzki 2011). So we do not think that there is any narrowly theoretical reason to favour a conjecture about minimal theory of mind over an alternative theory of rational action. One good way to decide between these two would involve identifying and testing signature limits of the two theories. In conjecturing that infants' mindreading involves minimal theory of mind, we are betting on its limits. Were are betting that it is more limited than some suppose but less limited that Zawidzki and others think.

8. Infants vs Scrub-jays (SS)

Suppose that some mindreading in infants, adults and non-human animals is cognitively efficient, and that it is cognitively efficient in part because it involves something like minimal theory of mind. More carefully, suppose that it is cognitively efficient in part because it involves a theory of mind that is crude and unsophisticated relative to that employed by many human adults at their most reflective. Shannon Spaulding argues that we cannot simply take for granted that infants, adults and non-human animals all use the same (minimal) theory of mind.

We agree, and did not intended to make this claim. Rather, our observation is that evidence of mindreading in non-human animals, human infants, and adults under cognitive load leads to the same question: How could mindreading be efficient enough for these different groups that have limited cognitive resources? We offer minimal theory of mind as a systematic, computational-level account of how this might be possible. But this does not entail that infants, adults and non-human animals have the same cognitive apparatus for two reasons. First, as Spaulding notes, it is possible to construct variations of the minimal theory of mind we describe. Variations can be characterised by adding and removing principles, and using alternative basic units to those we chose (objects and locations). Second, the very same minimal theory of mind could be implemented in different ways in different groups.

This diversity of possibilities means we are not making the strong assumption that Spaulding warns against, but it also leads to the formidable challenge of deciding whether the theory of mind of an infant is the same as, similar to or different from that of a chimpanzee, scrub jay or a human adult under cognitive load. For this reason we think that signature limits—which have not received much attention in research on mindreading—are particularly valuable. Different theories of mind will yield different signature limits, as will different ways of implementing a single theory of mind. So if it turned out that infants and scrub jays showed the same signature limits on their mindreading abilities we would have a basis for thinking that they were not only both mindreading, but also using roughly the same theory of mind. Of course, since their last common ancestor would have been a lizard-like amniote some 300 million years ago who probably lacked minimal theory of mind, we would bet that this was the product of convergence rather than common heritage.

9. Conclusion

We are much indebted to Hannes, Shannon and Tad for their objections and suggestions. There is quite a bit more to say than we have written here, and their comments will keep us thinking for some time. We hope the discussion will enable us to better understand how to deal with objections and further refine our position.

References

Apperly, I. A. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Hove: Psychology Press.

Apperly, I. A. & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *2009*(116), 4.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–1407.

Baldwin, D., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*(3), 708–717.

Byrne, R. W. (2003). Imitation as behaviour parsing. *Philosophical Transactions: Biological Sciences*, *358*(1431), 529–536.

Cacchione, T., Schaub, S., & Rakoczy, H. (2013). Fourteen-month-old infants infer the continuous identity of objects on the basis of nonvisible causal properties. *Developmental psychology*, *49*(7), 1325–1329. PMID: 22906060.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Cohen, A., Ivry, R. I., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning*, *16*(1), 17–30.

Corballis, M. C. (2007). Recursion, language, and starlings. *Cognitive Science: A Multidisciplinary Journal*, *31*(4), 697–697.

Dally, J. M., Emery, N. J., & Clayton, N. S. (2005). Cache protection strategies by western scrub-jays, aphelocoma californica: implications for social cognition. *Animal Behaviour*, *70*(6), 1251–1263.

Gergely, G. & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gómez, R. & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, *34*, 1221–1235.

Knudsen, B. & Liszkowski, U. (2012). 18-Month-Olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*.

Kozhevnikov, M. & Hegarty, M. (2001). Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review*, *8*(3), 439–453.

Low, J. & Watts, J. (2013). Attributing false-beliefs about object identity is a signature blindspot in humans' efficient mindreading system. *Psychological Science*, *24*(3), 305–311.

Newport, E. L. & Aslin, R. N. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Povinelli, D. J. & Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, *7*(4), 157–160.

Rakoczy, H., Fizke, E., Bergfeld, D., & Schwarz, I. (2013). Theory of mind and understanding intensionality emerge together in development. *submitted*.

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). "This way!", "No! that way!"—3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, *22*(1), 47–68.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–8.

Smith, J. (forthcoming). The phenomenology of face-to-face mindreading. *Philosophy and Phenomenological Research*.

Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy*, *15*(4), 434–444.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*(6), 979–1008.

Zawidzki, T. W. (2011). How to interpret infant socio-cognitive competence. *Review of Philosophy and Psychology*, *2*(3), 483–497.

Zawidzki, T. W. (2012). Unlikely allies: embodied social cognition and the intentional stance. *Phenomenology and the Cognitive Sciences*, *11*(4), 487–506.