Commentary on Butterfill and Apperly's "How to Construct a Minimal Theory of Mind?"

Tadeusz Zawidzki
Associate Professor, Philosophy
George Washington University

1. Introduction

Conceptual innovation in the sciences of social cognition has not kept up with the recent dramatic growth in new and surprising empirical evidence. In particular, most interpreters of this evidence seem limited to categorizing sociocognitive capacities as either exercises of a metarepresentational theory of mind or mere behavioral generalization (Baillargeon et al. 2010; Tomasello & Call 2008; Povinelli & Vonk 2004). Nonhuman, infant human, and adult human interpreters, it seems, can be only either Skinnerians or Cognitivists. But this is surely a false alternative. A number of theorists have recently defended the possibility of alternatives (Sterelny 2003; Maibom 2007; Andrews 2007; 2008; 2012; Hutto 2008; Apperly & Butterfill 2009; Apperly 2011; Zawidzki 2011; 2012; 2013). According to these theorists, nonhuman and infant human interpreters, as well as adult human interpreters under cognitive load, employ socio-cognitive competences more sophisticated than mere behavioral generalization, yet not as sophisticated as full-blown theory of mind, i.e., the metarepresentational "folk psychology" that enables normal adult humans to attribute propositional attitudes like belief and desire. However, none of these proposals approaches the precision and clarity of Butterfill and Apperly's (henceforth B & A) systematic treatment. Not only do they provide formally precise specifications of the components of such an intermediate socio-cognitive competence, they propose detailed empirical tests capable of establishing its "signature limits" (ms, p. 18), and thereby distinguishing it from other forms of social cognition. In my view, B & A's proposal is a tour de force, and ought to have a lasting impact on the study of social cognition.

In what follows, I offer what I consider friendly requests for clarification, meant to help hone a project that I think is substantially on the right track. The requests concern four topics. First, I address the distinction between tracking a mental state and representing it as such. Second, I discuss what I think is a related issue: the constraints that B & A's theory of socio-cognitive *competence* places on the *implementation* of socio-cognitive capacities. Third, I evaluate the experiments they propose for distinguishing between minimal and full-blown theory of mind. Fourth, and finally, I evaluate their reasons for classifying minimal theory of mind as a theory of *mind*.

2. Tracking vs. Representing as Such

The distinction between tracking a mental state and representing it as such plays an important role in B & A's proposal. The reason is that it enables them to distinguish between theory of mind abilities and theory of mind cognition. Theory of mind abilities require tracking states of mind; however, the cognition underlying these abilities needn't involve representing those tracked states of mind as such. For example, it may be possible to track some states of mind, e.g., perceptions, by representing only certain non-intentional behaviors, e.g., gaze direction, as such. This would constitute using theory of behavior cognition to track a mental state. B & A's

arguments suggest that theory of behavior cognition is not powerful enough to robustly track most mental states, at least not in the sorts of experimental settings that routinely test for socio-cognitive competence. But the distinction between what is tracked and what is represented by cognitive processes employed in the tracking opens up space for B & A's positive proposal: mental states like perceptions, desires, and beliefs can be tracked using cognitive mechanisms that do not represent these states. Instead, they represent relations between organisms, goals, objects, and locations. In other words, they argue that it is possible to pass tests of metarepresentational competence, in virtue of tracking representational mental states, without actually instantiating metarepresentational competence, i.e., without representing mental states. This makes sense only if there is a distinction between tracking something and representing it as such.

B & A motivate this distinction by appeal to other cases. For example, an animal might track toxicity without representing it as such, in virtue of representing some proxy for toxicity, e.g., odor. But this distinction makes sense only relative to some theories of representational content. For example, on Millikan's teleofunctional theory of representational content (1984), there doesn't seem to be anything more to it than what B & A call "tracking". According to Millikan, a mental state represents P just in case it enables organisms that token it to track P in such a way that their biological needs involving P are promoted. On this view, even if an organism is using odor as a proxy for toxicity, since the evolutionary explanation of why the organism tokens states that co-vary with odors must appeal to toxicity, these states represent toxicity; in Millikan's terms, toxicity is a "normal condition" on the odor detectors' performing their proper functions.

Of course, there are many alternatives to Millikan's theory. Nevertheless, I think B & A need to make very explicit what they mean by representational content. Millikan's theory is partly motivated by an important point about content: mental states typically do not "directly" represent their contents. Access to content is typically mediated by intermediate surrogates. We represent toxicity via certain effects it has on our sensory apparatus. And similarly for other distal or unobservable properties, events, and states, including mental states. So, just because our representations of such states are mediated by the detection of perceptually available concomitants, does not mean that the representations do not represent the states as such. For this reason, if tracking as B & A understand it is not sufficient for representing as such, they need to say what more is needed.

I suspect that, for B & A, the distinction between tracking and representing as such is one of degree. From what they say about minimal theory of mind, it seems that it is minimal precisely because it can track mental states in some contexts and not others. This naturally suggests that, as one's capacity to track mental states becomes more robust, i.e., one comes to appropriate expectations about behavior caused by mental states across a wider variety of tasks, tracking bleeds into representing as such. But if this is their view, then it raises another problem: it is no longer clear how to experimentally distinguish the implementational implications of their view from those of a competitor.

3. Constraints on Implementation

At one point in their argument, B & A make clear that their thesis is intended entirely at the level of a competence specification, in Marr's (1982) sense. They claim that their competence specification has no implications for implementation. However, this is hard to square with the conjectures in their concluding section. It seems clear from their discussion there that their thesis implies that human social cognition is mediated by two separate systems that remain distinct across lifespan. This is clearly a claim about architecture, about implementation. And it is this claim that many will find hardest to accept.

The reason is that many of the foremost researchers in the field defend a radically different architecture for human social cognition. For example, Baillargeon et al. (2010) seem to endorse a picture consistent with Fodor (1992) and Leslie's (1987) claims that concepts of mental states are part of an innate human competence that is exercised equally in infants and adults whenever behavior is interpreted. The differences in performance are all accounted for by limitations on domain-general capacities required for the application of these concepts in new contexts. So, for example, 15-month-olds pass non-verbal versions of the false belief task, while three-year-olds fail verbal versions *not* because these two tasks engage distinct *socio-cognitive* capacities – minimal theory of mind and full-blown theory of mind respectively – but because verbal versions of the false belief task require certain domain-general, attentional and executive capacities that aren't fully mature until after three years of age. So, on this view, and contrary to B & A, the same socio-cognitive competence – full-blown, metarepresentational theory of mind – is employed on all of these tasks, and performance differences are due to factors extrinsic to social cognition.[1]

This debate matters on a number of levels. Most obviously, it shows that B & A cannot ignore implementation. Whether or not the competence underlying some theory of mind abilities is distinct from the competence underlying others is surely a matter of implementation. But, more importantly, this debate is deeply related to the question raised above, regarding the difference between tracking and representing as such. As I suggested above, in the absence of an explicit theory of representational content, B & A seem restricted to understanding the distinction between tracking and representing as merely one of degree. As an interpreter develops more robust tracking abilities, i.e., can come to accurate expectations about behavior caused by mental states in a greater variety of tasks, tracking bleeds into representing as such. But if this is the picture, then it is hard to see how to experimentally distinguish B & A's two-system view from its nativist competitor. Both views predict the same developmental arc: children gradually develop more robust tracking abilities.[2] On one view, they represent mental states as such from very early on, in virtue of their innate conceptual endowment, and differences in performance are traced to the maturation of domain-general capacities. On the other view, children do not represent mental states as such until quite late in development, and this explains differences in performance.

---

[1] I draw this characterization from conversations I have had with J. Robert Thompson, though he shouldn't be held responsible for the way I've expressed it here.

[2] Differences in adult performance depending on cognitive load don't seem to distinguish these hypotheses either, since such differences can also be explained in terms of overtaxed domain-general capacities.

It's not clear how to distinguish these competing hypotheses about implementation, without first getting clear on what B & A mean by representing a mental state as such. If this notion had more meat to it than just tracking the mental state very robustly, it might be easier to experimentally distinguish their hypothesis from the nativist one.

4. Experimental Tests of Signature Limits

B & A propose two experimental tests for full-blown theory of mind. That is, they argue that an interpreter employing only minimal theory of mind could not pass these tests, while an interpreter employing full-blown theory of mind could. These are supposed to be nonverbal tests for attributing beliefs about identity, about whether or not two apparent objects are really the same. They motivate their claim that such tests can be passed only by full-blown mindreaders by appeal to a Frege case. Suppose an interpretive target believes that Charly is in Baltimore. Furthermore, suppose that Charly is Samantha. From this, a full-blown mindreader would *not* conclude that the target believes Samantha is in Baltimore, since the target might not know that Charly and Samantha are the *same* person. However, a minimal mindreader, in B & A's sense, *would* come to the analogous conclusion about their non-mentarepresentational surrogates for belief: if the target *registers* that Charly is in Baltimore, and Charly is Samantha, then the target also registers that Samantha is in Baltimore, since registration is a relationship between the target and an object. No representation is involved (ms, p. 19).

The thinking here appears to be the following. There is no way to even entertain the question whether Charly is Samantha without a metarepresentational capacity. This is a question about whether or not two representations apply to the same object. I think B & A are right about this. There may be other ways to test for metarepresentational capacity, but surely demonstrating the ability to track beliefs about identity is a sufficient test for it. However, I do not think B & A's nonverbal tests for tracking beliefs about identity succeed. Deflationary interpretations of passing these tests are possible, and once this is appreciated, it becomes hard to imagine a nonverbal way of testing for tracking beliefs about identity.

Consider the first test. A doll with two aspects is presented to a subject and her interpretive target. The target sees only one aspect at a time, in a way that suggests there are two dolls behind an occluder, while the subject sees that there is just one. The doll then departs from behind the occluder, with only one aspect visible to the interpretive target. If the target next reaches behind the occluder, a *full-blown* mindreader should *not* be surprised, having attributed to the target the belief that the two aspects belong to two different dolls, rather than to one and the same doll. A *minimal* mindreader, on the other hand, *should* be surprised since her target should have *registered* that one and the same object has now departed from behind the occluder.

But the subject would also show lack of surprise under the following non-metarepresentational interpretation. The target has failed to register that two "objects" are connected to each other in such a way that wherever one goes the other one goes. On this interpretation, the puppet's two faces are not two aspects of one and the same object. They are two separate "objects", "glued" to each other in some sense. Here, there is no attribution of a false belief about identity. The subject thinks her interpretive target sees two "objects", like her. These "objects" are just what *we* call

"different aspects of the same puppet"; and the interpretive target *fails to register the fact* that they are connected in such a way that they always move together. So the subject will expect her target to reach around the occluder because she has not registered that wherever one object goes the other must go; not because she mistakes one object for two.

Now consider the second test. Two perceptually indistinguishable balls are shown to the subject yet, at the same time, hidden from the interpretive target behind an occluder. One of the balls is shown to the target on one side of the occluder, and then returned behind it. Then the second ball is shown to depart from behind the occluder on the other side. B & A argue that if the interpretive target reaches behind the occluder, and the subject is employing a metarepresentational theory of mind, then she *should* be surprised. The reason is that a full-blown mindreader would attribute to the target a false belief about identity: that the second ball departing from behind the occluder is the *same* as the first ball shown and then returned behind the occluder. Such a mindreader should not expect her target to reach for the ball, which she thinks she has just seen depart from behind the occluder. In contrast, a minimal mindreader would not be surprised by the reach; in her view, the interpretive target has registered that one ball returned behind the occluder and that a different ball departed; so she should expect the target to reach for the first ball.

However, as before, a non-metarepresentational interpretation is possible. Suppose the subject merely attributes the following to her interpretive target: the target has failed to register that there is a pair of balls behind the occluder. In fact, she has incorrectly registered the numerosity of the set of balls behind the occluder as one. Therefore, when one ball departs, she shouldn't expect there to be any left. Thus, if she reaches around the occluder to retrieve the remaining ball, the subject should be surprised. One can attribute (mistaken) registrations of numerosities of sets of objects, without attributing beliefs about the identities of these objects.

The reason it is relatively straightforward to reinterpret such nonverbal tests of tracking beliefs about identity is related to a point B & A make about registration. Although they illustrate the concept in terms of objects at locations, they also claim that the concept is highly flexible, and that, in principle, interpreters can attribute registrations of a variety of different properties (ms, p. 17). For example, they suggest that scrub jays likely attribute registration of food types, rather than particular items of food. But if it's possible to attribute registrations of even very abstract properties like food types, then it should be possible to attribute (mistaken) registrations of properties like numerosities of groups, e.g., there being one object behind an occluder rather than a pair. And it should be possible to attribute failures to register relations between objects, e.g., the fact that two objects are connected or glued together such that where one goes the other always goes. Once we allow for registrations of such abstract facts, it becomes very difficult to test for the ability to track beliefs about identity nonverbally. The reason is that it requires knowing how nonverbal subjects carve up their worlds into objects, and it's hard to see how to determine this behaviorally.

5. Why is Minimal Theory of Mind a Theory of *Mind*?

On one level, B & A's labels for the abilities and competences they describe are unproblematic. "Theory of mind" has become a term of art in social cognition research, and, in many contexts, it

should not be taken too literally. Most of the time it means whatever competence is measured using standard experimental paradigms, like false belief tasks. There needn't be any implication that subjects literally have theories of mind.

However, on another level, this terminology is potentially pernicious. Language matters in science, and some ways of describing phenomena may suggest general assumptions about cognition, as well as avenues for future research, that are unwarranted. If there is no literal sense in which nonhuman animals, or infant humans, or adult humans under cognitive load employ a theory of mind to pass tests of socio-cognitive competence, then it is potentially misleading to label it as such. Moreover, B & A seem sensitive to this issue. They devote a whole section of the paper to defending their terminology: minimal theory of mind really is a theory of *mind*, they argue (ms, p. 18).

If so, it is worth asking in virtue of what minimal theory of mind is a theory of mind. B & A answer this question carefully. They acknowledge that, on one standard understanding of what makes something a theory of mind, their minimal theory of mind is not really a theory of mind. If one thinks that the use of metarepresentations is essential to theory of mind, then B & A acknowledge that the competence they describe is not a theory of mind: none of the four principles that constitute it is metarepresentational. However, they argue that on a different understanding of theory of mind, according to which it essentially involves the attribution of variables intervening between stimuli and responses, their minimal theory of mind really is a theory of mind. The reason is that their fourth principle, that incorrect registrations can be causes of behavior, involves the attribution of precisely such intervening variables. More specifically, on the fourth principle, registrations are intermediate variables that play a subset of the causal roles characteristic of beliefs, i.e., they give rise to principles generalizing across all goal directed actions, can be assigned correctness conditions, and causally influence actions. Registrations are not quite beliefs or representations, because they are not attitudes to propositions, cannot refer to never-existent objects, and are not subject to doxastic norms. But, it seems, they're enough like beliefs to count as mental; hence their importance to minimal theory of *mind*.

I do not think this reasoning is sufficient. First of all, it is not clear why minimal theory of mind should count as a theory of mind rather than, say, a theory of the brain. Presumably some brain states play a subset of the causal roles of beliefs, give rise to principles generalizing across all goal directed actions, and can be assigned correctness conditions. So why not call the competence they describe "minimal theory of brain"? In general, just because a competence represents some subset of the causal properties of some object or substance, this does not mean that it is perspicuously characterized as a theory of that object or substance under any description. For example, presumably many nonhuman animals employ a competence that represents some of the important causal properties of water. Does this mean that they deploy theories of H_2O? This is a systematically misleading way of characterizing such competences. Calling minimal theory of mind a theory of *mind*, just because it represents some features of minds is, I think, similarly misleading.

But I think there is a more serious problem with B & A's argumentation here. They give no reason for thinking of registrations as *intervening* variables. This language encourages the assumption that minimal mindreaders attribute internal, unobservable states, conceived as

residing within the targets of their attributions. I'm not sure, however, why registrations, even as understood on B & A's fourth principle, need be conceived of as internal, unobservable states of interpretive targets. As B & A acknowledge, relations to external facts can also have a causal influence on behavior. For example, correct registration is a relation between the interpretive target and some state of the world, and it causes successful behavior (ms, p. 14). B & A may respond that this is not the case for incorrect registrations. These can bear no relations to actual states of the world, since the situations of which they are registrations do not obtain. However, why conclude that incorrect registrations must therefore be internal, intervening causal factors? Why not simply say that minimal mindreaders assume that *relations to non-actual situations* can influence the behavior of their interpretive targets? This seems strange on first reading, but consider goals, which B & A do not treat as intervening causal factors (ms, pp. 9-10). An interpretive target's relation to her goal is a relation to a non-actual situation, which influences the target's behavior. If goals needn't be intervening states, then why must incorrect registrations be intervening states?

If registrations needn't be treated as intervening variables, then I do not think there is any sense in which minimal theory of mind is really a theory of *mind*. And I'm not sure why this even matters. If the only alternatives were theories of mind and theories of behavior, then it might matter that minimal theory of mind is a theory of mind; after all, it is far more abstract and sophisticated than a theory of behavior. However, I think what is so groundbreaking about B & A's project is precisely that it introduces conceptual resources that enable us to eschew this false alternative. Why can't the competence they describe constitute an entirely separate kind of socio-cognitive capacity? For example, why not call it a theory of "rational action," to use Gergely & Csibra's (2003) terminology; or perhaps a version of Daniel Dennett's "intentional stance" (1987; Zawidzki 2012; 2013)?

6. Conclusion

B & A's proposal is highly original, worked out in impressive detail, and highly relevant to current research into social cognition. I believe it has the potential to be a game changer. I suspect the questions I've raised have relatively straightforward answers. Minimal theory of mind promises to become a crucial component of the conceptual tool-kit employed in the sciences of social cognition.