Response to referees: 'How to Construct a Minimal Theory of Mind'

We are extremely grateful for a helpful referee's report. We accept the suggestions and have addressed each criticism in the revised version of our paper; the changes are explained in detail below. We were also asked to reply to another referee's comment, which we do below.

Referee's Comment

"I wonder if this manuscript is different enough from other papers published by the same authors on the same topic. Although the manuscript does not identify the authors, its content makes their identity quite clear. Although a bit of repetition is not a bad idea because each journal reaches a different audience, the editor needs to decide whether the amount of overlap between this manuscript and previously published contributions is significant enough to prevent the publication in Mind and Language."

> The referee is right that our identity is not hard to guess from this paper. Our paper has been through two rounds of reviews with four referees at another journal (Cognition). As three of the four reviews were positive, and as our paper has been cited once or twice, we imagine that this referee believes, incorrectly, that an earlier version of this paper has been accepted for publication elsewhere. Certainly we have only written one other paper together, and this is very different from the submitted paper.

Report 1
(We have added numbering in order to refer to the referee's comments)

[1] Let me begin by saying I am sympathetic to the author(s)' skepticism regarding the preponderance of claims as to infants' and other animals' theory of mind abilities as of late. However, I think it is a mistake to attempt to find fault with such research simply because one holds an opposite viewpoint, just as it is, paradoxically, a mistake to champion such findings, because it is one's preferred hypothesis to believe or desire such commonalities or mental continuities. In some sense, I worry that the authors are just as guilty of lack of objectivity as the researchers whose work they critique for a number of reasons, some of which are elaborated further in what follows. Having said that, I think that such a review is important, especially at such a time when the field is dominated by the opposing viewpoint and one-sided research, and such a critical evaluation of this research has been sorely missing.

> We are grateful for these observations. To address the worry that we are finding fault with some theory of mind research simply because we hold an opposite viewpoint (and to address concerns later in the Report, particularly [6] below) we have re-written what was Section 1 and divided it into three shorter sections (Sections 1, 2 and 3 in the new paper). We now start by motivating our question with research that does not involve controversy about infants or nonhumans (see Section 2). We now also explain more clearly that the motivation for our

project does not depend on showing that other views are incorrect, only that there are plausibly at least some cases where no hypothesis yet has decisive support. We argue that this is sufficient to motivate consideration of alternative hypotheses (see Section 3). We now also stress that we do not claim to have established the correctness of the hypotheses we discuss, nor the incorrectness of the main alternatives. In this paper our concern is primarily with theoretical issues.

[2] I don't think the distinction between theory of mind abilities and theory of mind cognitions is made clear on page 2 (first full paragraph). It seems to me from that reading, that if one is exploiting the knowledge or mental states of another, one must be cognizant of those states. If what one means is to manipulate one's behavior without being aware of the underlying mental states giving rise to those behaviors, that seems to me something different. If that is what the author(s) is trying to describe, that needs to be made more explicit here.

> This is a helpful suggestion. We have provided a more detailed description of the distinction between theory of mind abilities and theory of mind cognition (pp. 2-3), and now introduce the distinction with an additional example (p. 2).

[3] I think the logic in the last full paragraph on page 2 may be flawed. Just because the development of theory of mind in humans may depend on language, explicit training, certain cultural experiences etc. does not require that the same is true of other species. Similar abilities may arise under different circumstances in different species, with different necessary conditions through convergent evolution. Furthermore, to assume that the only manner in which some components of theory of mind must be present is if all components are present would be a mistake. Think mosaic cognition. It is possible that, for other species, fully developed aspects of the theory of mind system may be present, without some of the dimensions on which a fully developed theory of mind system depends in humans. It is dangerous to become locked into a way of thinking that is anthropocentric, such that we are looking for only exactly what exists in humans, and that develops in exactly the same manner.

> We agree and have therefore removed this argument, replacing it with alternative arguments which treat the cases of infants and nonhumans separately (pp. 6-7 and p. 8 respectively). While we still discuss the possible relevance of minimal theory of mind to research on nonhuman cognition rather than considering only on the human case, we are now more cautious.

[4] At the top of page 3, in parentheses the authors write "for an opposing view, see.." but no reference follows.

> We have added the missing reference.

[5] The author needs to make clear at the outset what they mean by propositional attitude concepts.

> We are grateful to the reviewer for pointing out that this term is not obvious and have re-written to avoid using the term at all. (One thing the Report made clear to us was that the paper introduces too many concepts and arguments; we have reduced both.)

[6] It takes the authors until the fourth page to indicate that the goal of the paper is to come up with a description of a minimal system for theory of mind. For the first three pages I believed I was going to be reading a critical review of recent papers purportedly evidencing theory of mind abilities in very young children and non-humans. The goal of the paper needs to be made more clear in the opening paragraph and in the abstract.

> The reviewer is right (thank you!). We have re-written the abstract and the opening paragraph as suggested, and made corresponding changes to the introductory sections of the paper and the conclusion. In the light of this point and points [10] and [11] below we also came to the realisation that the earlier version of our paper contained too much material which should really be in a dedicated review paper. We have therefore removed all but the most essential review material from this paper.

[7] The paper is well written and well referenced but it is a little information-dense, which may make it less readable for the average reader. Attempts to break up the density of the text with more transparent and accessible examples may help.

> We have added new examples (e.g. p. 2) and reduced density by cutting some material from earlier version.

[8] I don't think the section on Pure Behavior Reading (pages 5-7) is adequate or very readable. It did not make explicit how this type of reading can explain behavior in a way that allows an organism to predict the behavior of another, without ever going beyond simply reading external behavior.

> In response to this concern as well as to some other points (including [11] and [14] below) we have removed this section and no longer discuss pure behaviour reading at all. This required simplifying the presentation of minimal theory of mind, which may also help to address some of the referee's concerns about 'information density' (see [7]).

[9] At the bottom of pg. 8 the author discusses the distinction between object and goal directed actions in a footnote. I think this distinction should be moved to the body of the text and further elaborated on.

> In order to simplify things (and to avoid increasing the paper's length) we no longer make use of this distinction.

[10] In order to understand the Povinelli and Vonk behavioral abstraction example on pg. 9, the author needs to explain the paradigm that this example is in reference to in more detail. A naive reader will not understand the reference to dominant and

subordinate animals and which food they are orienting to without being very knowledgeable of the original study by Hare et al, on which this example is based. The discussion of the caching studies by Clayton and Emery on page 16 suffers from a similar lack of transparency. Furthermore, the author should cite Povinelli & Vonk (2004) as it contains more full-blown examples than given in the 2003 paper by the same authors.

> We agree with the referee. We now cite Povinelli & Vonk (2004). For reasons explained under [11] below and [6] above, and in order to limit the length of this paper, we have moved discussion of Povinelli and Vonk's notion of behavioral abstraction out of this paper.

[11] The author(s) may be oversimplifying the Povinelli, Penn and Vonk argument regarding behavioral abstractions to some degree on pg. 9. It is not the case that another organism, such as a chimpanzee, could NEVER generalize across contexts or circumstances for which it has never experienced that precise event. The whole point of the "abstraction" is that chimpanzees may be possible of forming abstractions (forming general classes or categories of events/objects). It is just that these general classes or categories never involve unobservables, such as mental states (Vonk & Povinelli, 2006). So, a chimpanzee that reasons that another chimpanzee can never respond to him if HE himself cannot see his conspecific's face can generalize that if the other chimpanzee is turned away, has his face covered by a blanket, is behind an opaque barrier, is in darkness etc., he will know not to beg with a visual hand gesture to that conspecific, but none of this requires that he has generalized that these are all cases where the conspecific cannot SEE him, because seeing is an internal mental state. But he is still capable of seeing the relationship between different contexts and scenarios – some of which he may not have previously experienced. I believe the point being made by Penn, Holyoak and Povinelli (2008) was that the only commonality in their examples was that in all cases there was a similar mental state that would have to be understood, such as seeing, or knowing, so that responding similarly in all cases would demonstrate having a theory of mind.

> The referee is right that we were giving a simplified version of the view under discussion. On reflection (and after re-considering Vonk & Povinelli, 2006) we believe that proper discussion of this view, which has been developed with in a series of papers over more than a decade and often seems to be misunderstood, requires much more detail. (Something similar is true of the main competing views, which we also do not discuss in detail.) We therefore decided it would be best to avoid discussing the details of this view (or the main opposing view) in this paper (see also points [6], [8] and [19]). While this has the disadvantage that relations between our view and the views of Povinelli, Penn and Vonk and their various supporters and opponents are not made explicit, it seems best to discuss this on another occasion in order first to present our own view.

[12] I think the authors need to be clearer about the difference between "encountering" and "seeing" on pg 14 when discussing the Moll & Tomasello (2006) experiment. If the experimenters controlled for contact and line of sight, etc., then it seems difficult to explain what encountering means other than visual access.

> This is a good suggestion. We did this, but then found that one consequence of removing discussion of pure behaviour reading (for reasons given under [8]) meant that we could not explain the relevance of minimal theory of mind to Moll & Tomasello's (2006) experiment. We have therefore cut this experiment and related material.

[13] Why can't you act in a goal directed fashion towards objects you have not yet encountered (pg. 16 – third condition of goal directed actions)? For instance preparing to dress up to look nice for a blind date to go to a restaurant you've never eaten at before (even if you don't yet know where you'll be going). Even animals set out to forage with the goal of finding food sometimes without having yet encountered a particular food source that they might then exploit subsequently. Upon seeing a novel food source they would still then recognize it, through generalization, as a food to be captured and consumed.

> This is a misunderstanding, but one we are entirely to blame for. The principle was not supposed to be true but only a useful heuristic. We have now explained this more carefully (see p. 11)

[14] In some sense I feel that the author(s) is just replacing mental state terminology with non mental state terminology without explaining how the new terms really work. For instance using registration instead of representation and encountering instead of seeing. The distinction between these sets of terms, in terms of what it means cognitively to the organism isn't always clear. The explanation on page 24 does nothing to help here. I especially don't understand how registration and encountering are not representational (pg. 33). Perhaps I am completely dense, but still, I suspect something is missing in the level of explanation that might pose problems for other readers as well.

> To address these concerns we have also added an earlier discussion of how encountering and seeing differ on pp. 12-13. We hope that the simplified, partly re-written and shorter presentation of the minimal theory of mind construction (in Section 4) helps to clarify this as well. The notion of *encountering* is defined on pp. 10-11 and *registration* is defined in two steps on pp. 13-14 and p. 16.

[15] At the bottom of page 19, the sentence should read 'they had a chance" or "the chance"

> Thank you, done.

[16] On page 26 in the puppet experiment, why would subjects be reaching behind the screen – what would their goal be? Why would they be expecting to find a puppet or wanting to grab one?

> We were initially puzzled by this comment. We only ever describe the protagonist as reaching, not the subjects. Suspecting that our description of the experiment might not have been sufficiently clear we have re-written this part of it (p. 20).

[17] On pg. 29 in discussing Scott and Baillargeon's paradigm, the author states that "along with reasoning about the protagonist's goal.."reasoning about objects present is necessary for success. Wouldn't reasoning about goals constitute reasoning about mental states, and therefore be evidence for some level of theory of mind?

> We have added an extended discussion of goal-directed action on pp. 9-10 in order to be clearer about the distinction between goals (or outcomes) and intentions. (Incidentally, even if reasoning about goals did constitute reasoning about mental states, this would not be an objection to the view we are defending; what matters for our purposes is not that the subjects are not reasoning about mental states but only that they are not reasoning about false beliefs concerning identity.)

[18] On pg. 31, the authors, in describing a study by Kovacs et al (2010) note that infants' expectations about an object's presence or absence is modulated by others' expectations, but it isn't clear from their description whether these expectations could be formed strictly on the basis of the behavior of the other individuals, without necessarily referring to their internal thoughts and expectations –which would be critical to the author's argument.

> The expectations in question are the subject's (i.e. the infant's) expectation that the ball is present (say) and the other's (i.e. the smurf's) expectation that the ball is absent (say). We are puzzled by the 'these expecatations could be formed ...' because we assume that the subject cannot both expect the ball to be present and, simultaneously, expect the ball to be absent. So the subject only forms one expectation. We initially thought the referee might be suggesting that the subject's knowledge of the other's expectation could be be formed on the basis of observing the other's behaviour (and since the other is static in this experiment, this means observing the other's posture and the other's presence or absence). We agree. But this can't be what the referee intends to suggest because it is not an objection to our interpretation (or the authors') of this study. On the interpretation we discuss, the subject has reason to expect the ball to be present (say) but also knows that the other expects the ball to be absent; and knowledge of the other's expectation interferes with the subject's own expectation. We infer that the referee is suggesting that this is not the only interpretation. The reviewer may be suggesting that when the protagonist is absent for the crucial transition, the bare fact of the protagonist's absence during the transition (not any knowledge of the protagonist's expectations) makes the subject less likely to form an expectation concerning the presence of absence of the ball. Whatever exactly the suggestion is, we agree with the referee that other interpretations of this experiment are possible and do not wish to place too much weight on a single study using a relatively new paradigm,

however groundbreaking. Accordingly we have revised the text and added a new footnote (footnote 20 on page 25) to reflect this concern.

[19] Overall, it is not clear enough what the author(s)' proposition really has to add to the existing controversy between mind-reading and behavior-reading accounts of experiments aiming to show that young children and non-humans evidence human-like theory of mind. [b] In fact, I am not convinced that the authors correctly represent the behavior-reading or behavioral-abstraction accounts of such findings. As such, I cannot recommend publication of the MS in its current form.

> We have made revisions throughout the paper in response to these comments. On [a] the new abstract and introduction and sections 2 and 3, together with the Conclusion, make clearer both what our proposition is and what it contributes. We take [a] to be linked to points [1], [6] and [12] and have already described some changes in more detail under those points. It may also be important that we now stress that we are not trying to resolve the controversy mentioned by the referee but rather to add identify a further, no less controversial hypothesis. In response to [b] we have also made several revisions as explained under [8], [10] and [11] (which we take [b] to be summarizing).