Referential and Cooperative Bias: in Defence of an Implicit Theory of Mind.

Josef Perner

University of Salzburg

Commentary in symposium on Katharina Helming, Brent Strickland, and Pierre Jacob's "Solving the puzzle about early belief-ascription," published on Brains Blog:

http://philosophyofbrains.com/2016/10/17/symposium-on-helming-strickland-and-jacob-solving-the-puzzle-about-early-belief-ascription.aspx.

Acknowledgements.

I thank Brian Leahy and Michael Huemer for valuable corrections and improvements to an earlier draft.

Helming, Strickland, and Jacob (2016: HSJ) propose a very interesting explanation for why children show sensitivity to an agent's false belief long before they pass the traditional false belief task around 4 years. Their core argument is that children understand false belief from early on but the experimenter's test question makes the traditional test so much more difficult. For, the experimenter's question requires children to shift from a third-person stance of watching the agent to a second person stance interacting with the experimenter. This shift creates two biases that favour the wrong answer (p 24):

> "This tension between the second- and the third-person perspectives generates two *biases*, one of which highlights the *epistemic* perspective on the object's actual location that is shared by the children and the experimenter, at the expense of the mistaken agent's different epistemic perspective on the object's location. The other bias *motivates* young children to help the mistaken agent achieve the goal of her instrumental action."

These biases ring a bell in my memory: The cooperative bias has played a role from the early days of ToM research, has been explicitly tested, and found to be ineffective. The referential bias I have used in connection with discourse referents (a variety of mental files). I will use mental files theory to sharpen the distinction used in our original theory between implicit and explicit understanding of belief. I will show how referential bias can be nicely integrated in this theory.

1. The cooperative bias

My earliest memory is of the paper by Perner, Leekam, and Wimmer (1987, p 127):

> "The *third hypothesis* concerned pragmatic considerations about the form of Wimmer & Perner's test question: 'Where will he look for the chocolate?' …. The experimenters intended this question as a test of children's ability to predict the protagonist's behaviour …. Subjects, however, may have taken this question as a request to help the protagonist to find the object ….

> They may have glossed the actual question as: 'Where *should* he look?' In the present study we tried to minimize this danger by asking 'Where does he *think* the chocolate is?'"

There was no difference between test questions, as confirmed in the metaanalysis by Wellman et al (2001). Yet, the think-question may not be the most convincing way to suppress the cooperative bias. Fortunately, Garnham & Perner (2001) provided a more pertinent test. They had the agent approach the object by different slides. Each slide had a precipitous drop. To soften the agent's arrival children could move a mat to the relevant slide. The mat was not used for the first condition in which two measurements were taken. Children's eye gaze to the top of the slide that led to the place where the agent thought his object was (believed location) served as an indirect measure. They were asked an explicit question about where the agent would go as a direct measure. In a second condition children were asked—before they were allowed to move the mat—where they will move the mat. The correct answer is to where the agent will come down. Moreover, should children interpret the question normatively as 'Where *should* you move the mat?' their answer would be the same, since there is no point in moving the mat to where the agent *should* but will not come down. So, cooperative bias should produce correct answers to this mat-move question. Yet, children's answers were no better than to the traditional question.

In a third condition children simply had to move the mat. Here, children had a stronger tendency to move the mat to the false-belief location than in their answers to questions. It was noted that some children moved the mat spontaneously, others hesitated and needed prompting. In a second experiment children's spontaneity was systematically recorded. The majority of hesitant children moved the mat to the object's actual location, while most of the spontaneous responders moved the mat correctly to the believed location—about the same proportion as showed correct anticipatory looking. In short, too much reflection leads to incorrect responses.

These results are compatible with research on implicit knowledge: It is not available for verbal or symbolic responses or delayed action, nevertheless has measurable influence on spontaneous action and eye gaze. This has been shown for blindsight patients as well as subliminal perception and susceptibility to visual illusions in healthy persons. Perner and Clements (2000) outline these parallels with early belief competence. It was for these parallels that we chose to speak of an *implicit understanding* of false belief. And this choice was underlined by Ruffman et al (2001) who found that children, who fail the FB-question but show the expectant looking to the believed location deny any possibility that the agent might appear at that very location to which they, nevertheless, look in expectation of his return.

From this brief reminiscence into ancient history we learn that there is evidence against the purported cooperative bias and that there is supportive evidence for the original proposal that the early sensitivity to belief is based on implicit knowledge. The arguments put forward by HSJ against "cultural constructivism" do not apply, since they are based on the so-called constructivists' denial of an early sensitivity to belief, while an implicit understanding implies that children have information about agents' perspectives. However, implicitness, in contrast to the nativist position defended by HSJ, also implies that something is missing from full explicit understanding. The implicit-explicit distinction, unfortunately, lacks a clear view of what is supposedly missing. With the help of mental files theory I can give a more precise specification of the missing element and provide a place for referential bias.

2. Mental files theory of mind

We used mental files theory (Perner & Brandl 2005; Perner, Rendl, & Garnham 2007) to explain why children start to appreciate alternative names for an object at the same time as they start to understand false belief (as tested with the traditional task). This led to the discovery that children have comparable problems with identity statements (Perner, Mauer & Hildenbrand 2011). In our current version (Perner & Leahy 2016) we use Recanati's framework. For present purposes the critical elements are:

(1) To think of an object is to have a mental file for it.

(2) One can think of an object in different ways (modes of presentation), e.g., of a rubber die that is also an eraser, as *the die* or as *the eraser*. This is captured by having two coreferential files for the same object (referent).

(3) One may think of the die and the eraser as different objects or one may know of their identity. This is captured by linking the two files (Perry 2002, ch. 10).

(4) To capture another agent's perspective about an object one entertains a coreferential file indexed to that agent. The indexed file is called a vicarious file (one uses it as if it were the agent's regular file for the object; Recanati 2012, chapters 14, 15).

(5) Our developmental assumption: children can have coreferential files before they can link them (around 4 years). This explains why mastery of the standard false belief task, the alternative naming game, and identity statements emerge at this age.

To explain these data we assume that calling the object a *die* or an *eraser* (any two sortals will do) creates different files for the same referent object. Young children familiar with these words will do this but start linking files, i.e., become fully aware that the die is the eraser, around age 4. Without a link children can talk about the die or the eraser, but when talking about it *as the eraser*, they fail to see that it is a *die* as well, which the alternative naming game (Doherty & Perner 1998) requires. With the ability to link the files children succeed. On mental files theory (Recanati 2012), representing another person's perspective also requires linking files. Another person's perspective is captured by coreferential object files that are used as if they were the other person's regular files (vicarious files). For instance, to capture Mistaken Max's false belief about the location of his chocolate in the original false belief task one needs a regular file to represent the chocolate from one's own view, where it is in its new location, and a vicarious chocolate file to capture Max's view that the chocolate is still in its old location. In Recanati's framework vicarious files must be linked to their corresponding regular files. Hence, without linking before age 4 children can have no vicarious files and fail the false belief test. Once they can link files they pass the FB test, succeed on the alternative naming game, and make sense of identity statements.

The most natural way to account for early sensitivity to belief is to allow for *unlinked vicarious files*. Recanati is quite right, that such files could not fulfill their proper function, but they come in handy for our developmental concerns. If in the alternative naming game the younger children can have two coreferential files for the same object without being able to link them, it seems natural to assume that

this might also happen for vicarious files. In the case of alternative identities children record in their die-file what happens when the object is used from the perspective of being a die, and in their eraser files what they learn about it under the perspective of being an eraser. Similarly children record in their regular chocolate-file what happens to the chocolate from their own perspective, and in their vicarious chocolate-file for Max only the events that Max witnesses from his perspective. The vicarious file records events that happen within Max's perceptual field (what Apperly & Butterfill, 2009, have called "registration" and Perner & Roessler, 2010, "experiential record"). Thus children can have *unlinked vicarious files* that contain information about other agents' perspective before they can link the files to their regular files. Without linking children have no voluntary access to their vicarious files when thinking about events from their own point of view. Hence, which view they take at a particular moment depends on extraneous influences that pull the child's attention to one or the other file: "referential pull".

3. Referential bias

Perner et al (2007, Table 1) identified three sources of referential bias:

(1) *Referential ambiguity*: In the typical condition, there is a clearly known location competing with the believed location of the object, hence potential referential ambiguity as to which location might be meant. This contrasts with the "disappear" condition (Wimmer & Perner 1983), where the object disappears from the scene and therefore leaves only the believed location as a dominant possibility. That this matters can be seen in the fact that Clements & Perner (1994) found children looking to the believed location only at the end of the third year (around 2 years 11 months); using a disappear condition Southgate et al (2007) found it at 2 years. Recently Wang and Leslie (2016) manipulated this factor and found a change in looking from a preference for the believed location in the disappear, to no preference in the non-disappear condition, to the actual location in a knowledge condition.

(2) *Desire targeting reality*: In some sense Mistaken Max wants to go where his object actually is, not where he thinks it is. Hence emphasizing Max's goal before asking the test question as in the traditional test should create a strong reality bias, which has been reported by Rubio-Fernandez and Geurtz (2015, Experiment 2).

(3) *Assertive commitment*: Why does asking a question about where Mistaken Max will go make the false belief task so difficult? One is not accountable for where one happens to look as a story unfolds. One's lookings can follow what might happen or what somebody thinks might happen. In contrast, answering a question commits one to giving a correct answer. Awareness that one's answer should reflect reality provides a reality bias.

As all three of these factors work in favour of the chocolate's real location in the traditional false belief task, children who cannot yet link files will give the wrong answer. Once able to link files children can voluntarily switch from their own view of the chocolate (regular file) to Max's view of the chocolate, and do not depend on extraneous factors to make them switch to Max's perspective. For that reason they pass the standard false belief task as they become able to link files.

The explicit question in the traditional task seems the most important factor that makes the task so difficult. This factor is supposed to be covered by source number 3 above. Admittedly, our criterion there is rather loose intuition and not theoretically well founded. Time to look to HSJ's analysis for help.

4. HSJ's analysis

> "the referential bias itself rests on two components: first of all, participants must identify the object (the marble) in accordance with the experimenter's referential action. If and when they have done so, their attention is primed to the object's actual location." (p 28).

To fathom their first component we have to look at the *experimenter's referential action* within the where-prediction question, e.g., 'Where will Sally look for her marble?' (p 27-28):

> "In order to correctly answer the experimenter's where-prediction question, participants must form a thought involving a reference to the location of the mistaken agent's (i.e. Sally's) marble. While the constituent 'the location of Sally's marble' is not linguistically encoded in the experimenter's question, participants must in effect form a thought including a mental reference to the location of Sally's marble …. Now, the reference of 'the location of Sally's marble' can be fixed in participants' minds with respect to one of two competing epistemic perspectives: Sally's mistaken perspective (not shared by the children) or the epistemic perspective shared by participants and the experimenter with whose communicative action they are currently engaged from a second person perspective. While the latter determines the marble's actual location, the former determines the empty location. We surmise that the latter trumps the former."

Unfortunately, the authors do not elaborate on their surmise. And it is far from obvious that they surmise correctly. To me—but I might be missing something terribly obvious—the location reference inherent in the question "Where will Sally look for her marble" is to the *location where Sally will look for her marble*. For someone with a solid theory of mind (HSJ's assumption) this would be unambiguously the location where Sally thinks her marble is. Hence, the referential bias—thus construed—should be in favour of the believed location and should yield a result opposite to what we need to explain. In contrast, on my assumption that children have an incomplete (unlinked vicarious files) understanding of the mind, the inherent reference in the question would not create any particular bias. In sum: no help from this part of HSJ's analysis. Fortunately, their second component looks much more promising (p 28):

> "Secondly, the mistaken agent's incorrect epistemic perspective on the object's location is further downgraded (or overshadowed) in the participants' minds by the rival epistemic perspective on the object's location which they share with the experimenter who is addressing them."

Clearly, a question ought to shift the child's attention to the perspective shared by child and question asker. This will highlight the shared information of the object's real location and create a referential bias towards the object's actual location. So this analysis constitutes a much better reason for why the question creates a reality bias than our intuitive factor "assertive commitment." Moreover, it can be nicely incorporated into my framework of an implicit understanding based on unlinked vicarious files.

This component contains, however, a problematic feature in conjunction with HSJ's assumption that children are endowed with a theory of mind. For children and experimenter also share the perspective that *Sally thinks the marble is in its old place*. This should create an opposing referential bias for the believed location and neutralise the referential bias for the real location. Notice that this argument does not apply to implicit understanding based on unlinked vicarious files, since such files need not represent the experimenter's view of another agent's view. Hence, only the reality bias created by their own perspective about the object's real location, which they share with the experimenter, will be effective.

5. Conclusion

I took a look back at our original explanation for the discrepancy between children looking in expectation of an agent to the believed location and claiming that he will appear at the real location. We had argued that their looking was governed by an implicit understanding and their answer to the question by an explicit understanding (theory). I now used mental files theory to fill an old theoretical gap. I distinguished implicit from explicit understanding with the difference between unlinked and linked vicarious files. Unlinked vicarious files contain information about the agent's perspective (the implicit understanding) but cannot be accessed at will from the child's own perspective contained in their regular files. But children who can't link files still show sensitivity to the agent's perspective if extraneous factors shift them to the vicarious files. Since the verbal false belief test strongly biases children towards their own perspective, they fail the test until they can link coreferential files and can access the vicarious file even when entrenched in their own perspective. The need of linking files for the verbal false belief test also explains why passing this test coincides with mastery of other tasks that require linking of coreferential files (e.g., alternative naming, identity statements).

HSJ's analysis of the pragmatics of asking a test question provides an improvement over our earlier intuitions about why the question creates a referential bias for the object's real location. Their analysis dovetails well with the shift from implicit to explicit understanding and, in fact, better than with their background assumption that children's use of a well-formed theory of mind is blocked by the reality bias created by the test question. They also lack a clear explanation of why this blockage disappears at around 4 years.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?. Psychological review, 116(4), 953.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. Cognitive development, 9(4), 377-395.

Doherty, M., & Perner, J. (1998). Metalinguistic awareness and theory of mind: Just two words for the same thing?. Cognitive Development, 13(3), 279-305.

Garnham, W. A., & Perner, J. (2001). Actions really do speak louder than words—But only implicitly: Young children's understanding of false belief in action. British Journal of Developmental Psychology, 19(3), 413-432.

Helming, K. A., Strickland, B., & Jacob, P. (2016). Solving the puzzle about early belief-ascription. Mind & Language.

Perner, J., & Brandl, J. L. (2005). File change semantics for preschoolers: Alternative naming and belief understanding. Interaction Studies, 6(3), 483-501.

Perner, J., & Clements, W. A. (2000). From an implicit to an explicit theory of mind. Beyond dissociation: Interaction between dissociated implicit and explicit processing, 273-294.

Perner, J., & Leahy, B. (2015). Mental files in development: dual naming, false belief, identity and intensionality. Review of Philosophy and Psychology, 1-18.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. British Journal of Developmental Psychology, 5(2), 125-137.

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. Science, 333(6041), 474-477.

Perner, J., Rendl, B., & Garnham, A. (2007). Objects of desire, thought, and reality: Problems of anchoring discourse referents in development. Mind & Language, 22(5), 475-513.

Perner, J. & Roessler, J. (2010). Teleology and Causal Reasoning in Children's Theory of Mind. In J. Aguilar & A.A. Buckareff (Eds.), Causing Human Action: New Perspectives on the Causal Theory of Action (chapter 14, 199-228). Cambridge, MA: Bradford Book, The MIT Press.Perry, J. (2002). Identity, personal identity and the self. Hackett Publishing.

Recanati, F. (2012). Mental files. Oxford University Press.

Rubio-Fernández, P., & Geurts, B. (2015). Don't Mention the Marble! The Role of Attentional Processes in False-Belief Tasks. Review of Philosophy and Psychology, 1-16.

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. Journal of experimental child psychology, 80(3), 201-224.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. Psychological Science, 18(7), 587-592.

Wang, L., & Leslie, A. M. (2016). Is Implicit Theory of Mind the 'Real Deal'? The Own-Belief/True-Belief Default in Adults and Young Preschoolers. Mind & Language, 31(2), 147-176.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. Child development, 72(3), 655-684.