COMMENTARY

Young children flexibly attribute mental states to others

Peter Carruthers[a,1]

Developmental social psychology has been in turmoil since 2005, when the first study using implicit, nonverbal measures with preverbal infants appeared to show that they could attribute false beliefs to other agents, forming expectations about an agent's behavior on that basis (1). This was an exciting finding, suggesting that core components of human "mindreading" abilities (or theory of mind) might be innate or innately channeled early in development. Before that time, since the first published investigation of children's mindreading in 1983 (2), it had been a near-consensus in the field that children are incapable of attributing false beliefs to other people until around the age of 4 y, suggesting a protracted period of learning (3). However, all these earlier studies had used elicited responses (normally verbal ones) with children, whereas the new paradigm relied on spontaneous behavior (in this case, looking behavior, which is longer when the infant's prior expectations are violated). Since 2005, there have been dozens of other positive findings using implicit measures, including not just expectancy-violation looking but also anticipatory looking (4), spontaneous helping behavior (5), and more. Most of these studies have tested barely verbal infants in the first half of the second year of life, but some have shown that infants can track the false belief of another agent from as young as 6 mo of age (ref. 6; see refs. 7 and 8 for review and discussion of such findings). However, the field remains mired in controversy, and a number of deflationary explanations of the implicit mindreading results have been proposed. In PNAS, Király et al. (9) make an important new contribution to these debates, showing that young children's use of false-belief information is much more flexible than any of these deflationary accounts would predict.

Deflationary Accounts

While some have claimed that the implicit false-belief findings can be explained in terms of low-level stimulus factors or associations, which bias the infants' attention or subsequent behavior (10), the consensus among most researchers in the field is that there are now too many infant studies, using too many variations in materials and methods, for this account to be plausible. This is because many different explanatory factors will need to be postulated across the various studies. It is much more parsimonious to postulate a common underlying explanation (namely, mindreading ability). Moreover, this deflationary approach, which has not been scientifically fruitful, is confined to seeking post hoc explanations of the findings of others.

However, a second and more fruitful deflationary account likewise postulates a single factor underlying the infant results (although it postulates two systems overall when older children and adults are included). It claims that there two separate systems for tracking the mental states of other people. One is fast and automatic but inflexible, and present from early infancy; the other is slow and working-memory dependent but flexible, and acquired slowly over the first few years of life through social and communicative exchanges with other people (11). While this two-systems view has been criticized elsewhere on theoretical (12) and empirical (13) grounds, Király et al. (9) provide important new evidence against it. Also of significance is that, in the course of doing so, (*i*) they replicate a previous false-belief study with toddlers (14), given that there have been a number of claims in the field of failure to replicate (15); and (*ii*) they provide additional evidence that children as young as 3 y are capable of accessing and drawing inferences from an episodic memory of a prior event.

The Study Method

Király et al. (9) modified a method previously used to show false-belief understanding in 17-mo-olds (14) but now targeted at 36-mo-olds, which seems to be the age at which young children first become capable of using episodic memories of previous events (but still a full year earlier than they begin to pass elicited

[a]Department of Philosophy, University of Maryland, College Park, MD 20742

Author contributions: P.C. wrote the paper.

The author declares no conflict of interest.

See companion article 10.1073/pnas.1803505115.

[1]Email: pcarruth@umd.edu.

versions of the false-belief test). The basic idea is that after being familiarized with two unfamiliar (and unlabeled) objects, the child watches the experimenter place the objects in two different boxes before leaving the room. In her absence, another experimenter sneakily switches the locations of the objects. The first experimenter then returns, opens the boxes in such a way as to enable the child (but not the experimenter) to see the contents, points to one of the boxes, and says, "Remember, I put a sefo in here; can you pass me the sefo?" ("sefo" is a nonce word). To interpret which object the first experimenter means to refer to, children have to realize that she has a false belief about the respective locations of the two objects; so, although she is pointing to one of them, she really means to refer to the other. The outcome measure is the object that the child selects or first reaches toward. Using this method, 17-mo-olds have been shown to interpret the speaker's request in a belief-dependent manner (14), a finding that is replicated as a control experiment with 18-mo-olds in Király et al.'s Experiment 2 (9).

Experiment 1

In Király et al.'s Experiment 1 (9), the method described above was adapted to include a pair of sunglasses, which the first experimenter put on before the second experimenter switched the locations of the objects in front of her, after which she left the room. In her absence, the children were encouraged to play with the sunglasses. Half of the children discovered, surprisingly, that the sunglasses were completely opaque, whereas the other group found that they could be seen through in the way one might expect. The first experimenter then returned, opened the boxes, and asked to be passed the sefo, as in ref. 14. Although the experimenter had been present when the objects were switched, the group of children who had subsequently learned that the sunglasses were opaque should thereafter realize that she was ignorant of the switch and interpret her request accordingly, whereas the other, normal-sunglasses group should interpret her as pointing toward the object she intends. This is exactly the result obtained with 3-y-olds (but not with 18-mo-olds, who, it is thought, as yet lack episodic-memory capacities).

Note that to succeed in the false-belief condition in this task, the children have to have access to and draw inferences from their episodic memory of the earlier switching event. Either at the time when they discover that the sunglasses are opaque, or later when asked to pass the sefo to the experimenter, they have to recall that the experimenter was wearing the sunglasses when the contents of the two boxes were switched. Also, in light of that memory, they have to update the belief previously attributed to the experimenter from a true one (assuming she had observed the switch through her sunglasses) to a false one (since the opaque glasses actually prevented her from seeing the switch happen).

This is a remarkable finding. It shows that social cognition by this age is not just an automatic online process, as two-systems theorists maintain. On the contrary: Although spontaneous, it is under a form of executive control. Even if the episodic memory of the event in question were evoked automatically (rather than searched for) by the child's own experiences when exploring the opaque sunglasses, the child still has to perceive the relevance of that memory to the experimenter's current mental state, and this needs to lead her to update her representation of the experimenter's belief, either at that moment or later when interpreting the experimenter's request. Indeed, if the studies cited in ref. 9 are correct that children first become capable of episodic remembering at around the age of 3 y, this means that children immediately start to integrate those capacities into their mindreading of other people's mental states.

Far from being inflexible and automatic, it seems that young children's mindreading abilities can make use of executively controlled searches of episodic memory and/or working-memory–demanding inferences drawn on the basis of such episodic memories.

Experiment 3

Király et al.'s Experiment 3 (9) then provides a conceptual replication of this result with another group of children. However, this time, the children did not need to rely on episodic memory to update a true belief to a false one, but rather to update a false belief to a true one. In this experiment, the first experimenter left the room before the objects were switched, as in ref. 14, but the children were then encouraged, after the switch, to leave through the same exit to bring the experimenter back. Half of the children found her peering into the experimental room through a one-way mirror (in which case she would have observed the switch take place, and would thus have true beliefs about the objects' locations, not false ones), whereas the other half found her in the room without seeing the mirror (which was covered with a curtain). Here, too, to interpret the experimenter's later request to be handed the sefo, the children in the first group would need to access an episodic memory of the switching event, which took place in the absence of the experimenter but is now updated to include her watching through the one-way mirror, to reason that she knows the current locations of the objects. Children in the second group, in contrast, should interpret her request as grounded in a false belief about those locations and should therefore hand her the object she is not pointing toward. This is just what was found.

These new findings are deeply problematic for two-systems theories of the difference between infant/toddler and 4-y-old mindreading competence. Far from being inflexible and automatic, it seems that young children's mindreading abilities can make use of executively controlled searches of episodic memory and/or working-memory–demanding inferences drawn on the basis of such episodic memories. But that leaves in place a puzzle, of course: Why is it that children fail elicited/verbal versions of these tasks before the age of 4 y if they nevertheless possess the underlying competence to flexibly attribute mental states to others and draw appropriate inferences from those attributions? There are, however, a number of proposals in the literature for resolving this puzzle, either in terms of the greater executive-function demands of elicited tasks (8) or the increased mindreading load imposed by such tasks, given that speech interpretation and communication themselves implicate mindreading (7), or by appealing to younger children's poor pragmatic-interpretation skills (16, 17). It remains for future work to adjudicate these accounts or (since they are mutually consistent) to uncover their respective contributions to young children's failures in verbal tasks.

1 Onishi KH, Baillargeon R (2005) Do 15-month-old infants understand false beliefs? *Science* 308:255–258.
2 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.
3 Wellman HM, Cross D, Watson J (2001) Meta-analysis of theory-of-mind development: The truth about false belief. *Child Dev* 72:655–684.
4 Southgate V, Senju A, Csibra G (2007) Action anticipation through attribution of false belief by 2-year-olds. *Psychol Sci* 18:587–592.
5 Buttelmann D, Carpenter M, Tomasello M (2009) Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition* 112:337–342.
6 Southgate V, Vernetti A (2014) Belief-based action prediction in preverbal infants. *Cognition* 130:1–10.
7 Carruthers P (2013) Mindreading in infancy. *Mind Lang* 28:141–172.
8 Scott RM, Baillargeon R (2017) Early false-belief understanding. *Trends Cogn Sci* 21:237–249.
9 Király I, Oláh K, Csibra G, Kovács ÁM (2018) Retrospective attribution of false beliefs in 3-year-old children. *Proc Natl Acad Sci USA*, 10.1073/pnas.1803505115.
10 Heyes C (2014) Submentalizing: I am not really reading your mind. *Perspect Psychol Sci* 9:131–143.
11 Butterfill S, Apperly I (2013) How to construct a minimal theory of mind. *Mind Lang* 28:606–637.
12 Carruthers P (2016) Two systems for mindreading? *Rev Phil Psychol* 7:141–162.
13 Buttelmann F, Suhrke J, Buttelmann D (2015) What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *J Exp Child Psychol* 131:94–103.
14 Southgate V, Chevallier C, Csibra G (2010) Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Dev Sci* 13:907–912.
15 Baillargeon R, Buttelmann D, Southgate V (2018) Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cogn Dev* 46:112–124.
16 Westra E (2017) Pragmatic development explains the false-belief task. *Rev Phil Psychol* 8:235–257.
17 Westra E, Carruthers P (2017) Pragmatic development explains the Theory-of-Mind Scale. *Cognition* 158:165–176.