There must be more to development of mindreading and metacognition than passing false belief tasks

doi:10.1017/S0140525X0900065X

Mikolaj Hernik,[a] Pasco Fearon,[b] and Peter Fonagy[c]

[a]*Baby Lab, Anna Freud Centre, London, NW3 5SD, United Kingdom;* [b]*School of Psychology and Clinical Language Sciences, University of Reading, Reading, RG6 6AL, United Kingdom;* [c]*Research Department of Clinical, Educational and Health Psychology, University College London, London WC1E 6BT, United Kingdom.*

mikolaj.hernik@annafreud.org
http://www.annafreudcentre.org/infantlab/mhernik
r.m.p.fearon@reading.ac.uk
http://www.reading.ac.uk/psychology/about/staff/r-m-p-fearon.asp
p.fonagy@ucl.ac.uk
http://www.ucl.ac.uk/psychoanalysis/unit-staff/peter.htm

Abstract: We argue that while it is a valuable contribution, Carruthers' model may be too restrictive to elaborate our understanding of the development of mindreading and metacognition, or to enrich our knowledge of individual differences and psychopathology. To illustrate, we describe pertinent examples where there may be a critical interplay between primitive social-cognitive processes and emerging self-attributions.

Carruthers makes a good case that self-awareness of propositional attitudes is an interpretational process, and does not involve direct introspective access. He also argues that mindreading and metacognition rely on one cognitive mechanism; however, in this case we are less persuaded by the evidence which hinges on Carruthers' reading of well-rehearsed data from autism and schizophrenia. We think that these two predictions have distinct bases and it is at least conceivable that there are two dissociable interpretative meta-representational systems capable of confabulation: one self-directed, one other-directed. Thus, the argument in favour of model 4, over, say, a version of model 1 without a strong commitment to non-interpretative access to self-states, is based purely on parsimony. Our intention is not to defend such a two-system model, but rather to point out that even if one accepts that metacognition involves interpretation, mindreading and metacognition may still be dissociable. Furthermore, Carruthers pays little attention to the differences between input channels associated with first- and third-person mindreading and the surely distinct mechanisms (arguably within the mindreading system) that translate them into attitude-interpretations. As a result, we worry that Carruthers may end up with a rather impoverished model that struggles to do justice to the broader phenotype of first- and third-person mindreading, its development, and the ways in which it may go awry in psychopathology.

Carruthers' reading of developmental evidence is restricted to the standard strategy of comparing children's performance across false-belief tasks. These are inherently conservative tests of mindreading ability, as false-belief-attribution is neither a common nor a particularly reliable function of the mindreading system (Birch & Bloom 2007; Keysar et al. 2003). Clearly, there are earlier and more common abilities central to development of third-person propositional-attitude mindreading – for example, referential understanding of gazes (Brooks & Meltzoff 2002; Senju et al. 2008) or pretense. However Carruthers does not discuss development of the mechanism that is central to his model. He also overlooks evidence that the tendency to engage in pretence has no primacy over the ability to understand pretence in others (Leslie 1987; Onishi et al. 2007).

There are other developmental areas potentially useful to Carruthers' argument. Several socio-constructivist accounts (e.g., Fonagy et al. 2002; 2007) attempt to describe the developmental mechanisms by which early social-cognitive competences, expressed especially in early interactions with the attachment figure (Sharp & Fonagy 2008), give rise to metacognitive awareness. Arguably, the most advanced of these theories is the social-biofeedback model proposed by Gergely and Watson (1996; 1999; Fonagy et al. 2002; Gergely & Unoka 2008). Currently, this model assumes that in repetitive episodes of (mostly) nonverbal communication (Csibra & Gergely 2006) mothers provide marked emotional "mirroring" displays which are highly (but inevitably imperfectly) contingent on the emotional displays of the infant. By doing so, mothers provide specific forms of biofeedback, allowing infants to parse their affective experience, form separate categories of their affective states, and form associations between these categories and their developing knowledge of the causal roles of emotions in other people's behaviour.

It is important to note that socio-constructivist theory is an essential complement to Carruthers' model 4, bridging a potentially fatal gap in his argument. People do *attribute* propositional emotional states to the self, and it seems reasonable to assume that their *actual* emotional states (propositional or not) play a role in generating such attributions. Carruthers' current proposal under-specifies how the mindreading system, which evolved for the purpose of interpreting others' behaviour, comes to be capable of interpreting primary somatic data specific to categories of affective states and of attributing them to the self. Furthermore, according to Carruthers, when the mindreading system does its standard job of third-person mental-state attribution, this sort of data "play little or no role" (target article, sect. 2, para. 8). Presumably, they can contribute, for example, by biasing the outcome of the mindreading processes (like when negative affect leads one to attribute malicious rather than friendly intentions). However, in first-person attributions, their function is quite different. They are the main source of input, providing the mindreading system with cues on the basis of which it can recognize current emotional attitude-states. The social-biofeedback model assumes that the mindreading system is *not readily* capable of doing this job and spells out the mechanism facilitating *development* of this ability. Putting it in terms of Carruthers' model 4: it explains how primary intra- and proprioceptive stimulation gains attentional focus to become globally accessible and how the mindreading system becomes able to win competition for these data.

Research on borderline personality disorder further illuminates the value of the socio-constructivist model (Fonagy & Bateman 2008). The primary deficit in borderline personality disorder (BPD) is often assumed to be a deficit in affect self-regulation (e.g., Linehan 1993; Schmideberg 1947; Siever et al. 2002). We have evidence of structural and functional deficits in brain areas of patients with BPD normally considered central in affect regulation (Putnam & Silk 2005). Accumulating empirical evidence suggests that patients with BPD have characteristic limitations in their self-reflective (metacognitive) capacities (Diamond et al. 2003; Fonagy et al. 1996; Levy et al. 2006) that compromise their ability to represent their own subjective experience (Fonagy & Bateman 2007). There is less evidence for a primary deficit of mindreading (Choi-Kain & Gunderson 2008). Evidence from longitudinal investigations suggests that neglect of a child's emotional responses (the absence of mirroring interactions) may be critical in the aetiology of BPD (Lyons-Ruth et al. 2005), more so even than frank maltreatment (Johnson et al. 2006). We think that the BPD model may become an important source of new data that could illuminate relationships between mindreading and self-awareness and their developmental antecedents. We suggest that children who experience adverse rearing conditions may be at risk of developing compromised second-order representations of self-states because they are not afforded the opportunity to create the necessary mappings between the emerging causal representations of emotional states in others and emerging distinct emotional self-states.

ACKNOWLEDGMENTS
The work of the authors was supported by a Marie Curie Research Training Network grant 35975 (DISCOS). We are grateful for the help and suggestions made by Liz Allison and Tarik Bel-Bahar.

Banishing "I" and "we" from accounts of metacognition

doi:10.1017/S0140525X09000661