Is mindreading a gadget?

Pierre Jacob, Thom Scott-Phillips

Is mindreading a gadget?

Pierre Jacob* and Thom Scott-Phillips**

*Institut Jean Nicod,

Département d'études cognitives, ENS, EHESS, CNRS, PSL University, UMR 8129

Ecole Normale Supérieure,

29, rue d'Ulm,

75005 Paris, France

**Department of Cognitive Science, Central European University, Budapest, Hungary.

Department of Anthropology, Durham University, Durham, UK.

Email of corresponding author: jacobpiotr11@gmail.com

The authors are very grateful to two anonymous reviewers for their outstanding comments and wish to gratefully acknowledge support from the European Research Council (ERC) under the European Union Seventh Programme (FP/2007-2013)/ERC Grant 609819.

Abstract

Non-cognitive gadgets are fancy tools shaped to meet specific, local needs. Cecilia Heyes defines cognitive gadgets as dedicated psychological mechanisms (e.g. cooking and sporting expertise) created through social interactions and culturally, not genetically, inherited by humans. She has boldly proposed that many human cognitive mechanisms (including imitation, numeracy, literacy, language and mindreading) are gadgets. If true, these claims would have far-reaching implications for our scientific understanding of human social cognition. Here we assess Heyes's cognitive gadget approach as it applies to mindreading. We do not think that the evidence supports Heyes's thought-provoking thesis that human children are taught to read minds the way they are taught to read words. We highlight a potential circularity lurking behind this analogy, and we explain why we are unpersuaded by Heyes's anti-mentalistic proposal for handling data inconsistent with the gadget view, which others take to be evidence for mindreading in human infancy. We conclude that while human minds may well be filled with gadgets, mindreading is unlikely to be one of them.

1. Introducing cognitive gadgetry

"Cognitive gadget" is a good piece of terminology—it is an innovative terminological gadget. In ordinary language, gadgets are tools created and assembled to meet a particular need. They are products of cultural invention, scaffolding and modification. Some are designed with much precision and finesse, while others are much more haphazard, even make-do-and-mend. Many cognitive processes are like this. The most celebrated examples are literacy and numeracy acquired by human children through explicit schooling (cf. Brem et al., 2006; Butterworth, 2005), but there are many further uncontroversial instances, such as those involved in religious rituals, sports, games and scientific inquiry. In all these cases, the relevant cognitive processes meet particular, local goals, yet they would not exist if not for the impact of specific cultural and social factors.

In her recent book and in several earlier publications, Cecilia Heyes has argued that human social cognition is much more gadget-y than most naturalistically inclined philosophers, cognitive scientists and cognitively-minded anthropologists—who disagree on much else—have tended to assume (e.g. Barrett, 2015; Boyd & Richerson, 1988; Boyer, 2018; Carey, 2009; Carruthers, 2006; Chomsky, 1975; Dretske, 1988; Fodor, 1975; Gallistel, 1990; Millikan, 2004; Pinker, 1997; Sperber, 1996). She contrasts gadgets — domain-specific social cognitive mechanisms that are culturally inherited through social interactions — with instincts, which she characterizes as domain-general cognitive mechanisms that are genetically transmitted from biological parents to offspring.[1] Whereas she takes cognitive gadgets to be mechanisms that respond selectively to social stimuli (e.g. human actions), she takes domain-general mechanisms to respond to both social and non-social stimuli (cf. Heyes,

[1] We assume for the sake of argument that the distinction between social cognitive instincts and social cognitive gadgets is exhaustive, but we do not accept it in actuality. We believe instead that instincts and gadgets are not a dichotomy but instead two end-points of a continuum. See Sperber (2019) and Del Giudice (2019) for elaboration.

2015a). Many human cognitive capacities (e.g. imitation and the language faculty that Pinker, 1994 famously called "the language instinct") have been widely regarded as instincts by cognitive scientists. But they are, Heyes argues, better seen as gadgets. Gadgets rather than instincts are, on her view, what make human social cognition and human social lives so peculiar compared to those of other animals.

What Heyes (2018: 6) patently objects to, in particular, is the idea that the uniqueness of human social cognition could be sealed in the Stone Age mind of Pleistocene hunter-gatherers and genetically inherited by their biological progeny all the way down. Heyes's gadget approach to social cognition is a direct challenge to what she calls "High Church" evolutionary psychology (see e.g. Barkow, Cosmides, and Tooby, 1995; Pinker, 1994) and also to cultural evolutionary theory (see e.g. Boyd and Richerson, 1988; Henrich, 2017). Both these schools of thought tend to assume that what is distinctive of human social cognition is genetically inherited and that only the 'grist' of human social cognition (what is processed), not the 'mills' (the cognitive mechanisms themselves), is culturally inherited.[2]

In Heyes's cultural evolutionary psychological framework, both the mills that make human cultural learning possible and their grist (the outputs of cultural learning) count as cognitive gadgets. Although it is true that the grist is sometimes hard to tell apart from the mill, Heyes's account nevertheless stands in contrast to mainstream thought in several other areas too, in particular cognitive development and experimental cognitive psychology more broadly. As it turns out, one deep challenge facing Heyes's gadget approach to human social cognition is precisely whether "cognitive gadget" is applicable not just to the grist, e.g. mathematical theorems, computer algorithms, scientific theories, narratives or poems, but to the mills themselves, i.e. to cognitive mechanisms such as literacy or numeracy. A sample of

[2] Quite probably not all advocates of either of these schools of thoughts would agree with Heyes's characterization (see also Morin, in press). We will not enter into these debates here.

possible cognitive gadgets includes (but is not limited to) causal understanding, episodic memory, imitation, selective social learning, mindreading, language, numeracy and literacy. Some of these purported examples are more controversial, and have more far-reaching consequences, than others. It is, for example, uncontroversial that literacy is a gadget. Indeed, far from being controversial, "the origins of literacy provide a proof of principle for cognitive gadgets" (Heyes, 2018, p.18). We agree that the question is not whether there are cognitive gadgets, but what they are — which other cognitive processes are gadgets, or gadget-like?

Here we focus on the thesis that mindreading is a gadget. The claim, in short, is that learning to read minds (a cognitive capacity also known as 'mindreading', 'mentalizing', or 'theory of mind') is like learning to read print. Heyes and others have earlier articulated and defended this thesis at some length (Heyes & Frith, 2014), and if it is correct it would reorient, in a fundamental way, a 40-year-old research program in social cognition, focused on developmental and comparative aspects of human mindreading (cf. Baillargeon et al., 2016). It would also, moreover, have far-reaching consequences for fields concerned with the flow of information in human groups and the psychological bases of culture and society, such as cognitive anthropology and cultural evolution (see e.g. Bloch, 2012; Boyer, 2018; Morin, 2015). So, the stakes are high.

We offer a detailed critique, in two steps. First (§2) we summarize and then break down Heyes's arguments for a close analogy between reading minds and reading print. We shall argue that the supposed parallels between the individual cognition of reading minds and reading print are in fact not as deep as they first seem; that children are not taught to read minds as they are taught to read words; and that there is a hidden circularity lurking behind the analogy. We then (§3) respond to Heyes's arguments about the nature of infant mindreading. If mindreading is a cognitive gadget, then it should not emerge early and reliably in ontogeny, but many papers report data consistent with the early emergence of

mindreading in ontogeny. In response, Heyes has offered a reinterpretation of these data, based on the novel theoretical idea of 'submentalizing'. We shall argue that this alternative approach is itself mistaken, thus nullifying the objection. Combining these two classes of argument, we conclude that while human minds may well be filled with gadgets, mindreading is unlikely to be one of them.

We would like to emphasize that, while we reject the thesis that mindreading is a gadget, we do think that human literacy and human numeracy are robust examples of cognitive gadgets in Heyes's sense. Whether human cognitive capacities other than literacy and numeracy are gadgets should be discussed case by case. For example, we think that Heyes's thesis sheds light on some (but not all) features of human imitation. Whether human language, causal understanding and episodic memory too are cognitive gadgets would require a full discussion that we shall not engage in here.

2. The analogy between reading minds and reading words

Ever since Premack and Woodruff's (1978) seminal paper on whether chimpanzees can read minds and the accompanying commentaries by three philosophers (Bennett, 1978; Dennett, 1978; Harman, 1978), most experimental psychologists and cognitive scientists have agreed that one hallmark of mindreading is false-belief understanding (i.e. an individual's capacity to attribute false beliefs to others) because it demonstrates that the individual expects an agent's goal-directed action to depend, not merely on objective features of her environment, but on her mental representation of her environment. (To be clear, false-belief understanding is not coextensive with mindreading; it is just widely seen as a clear hallmark.) In the early stages of this literature, most experimental studies with children used verbal false-belief tasks, in which participants are told a story (with the help of props) about an agent who places her toy in one

of two containers before leaving. In her absence, the toy is moved to the other container.[3] Participants' task is to answer a question asked by the experimenter either about where the mistaken agent is likely to look for her toy or where she thinks her toy is. Most preschoolers, in many different cultures, reliably fail such tasks and point to the toy's actual location; only around 4,5 years of age do most children show reliable success.

Building her argument that mindreading is a cognitive gadget, Heyes agrees with the many social constructivists who take success on verbal false-belief tasks as the hallmark of genuine false-belief understanding and therefore of mindreading, and who dismiss more recent findings based on non-verbal tests (which some psychologists take to be evidence of false-belief understanding in preverbal human infants; see §3). Unlike other social constructivists, however, Heyes rejects both the theory-theory approach (Gopnik, 1996; Gopnik & Wellman, 1992; Gopnik & Wellman, 1994) and the simulation approach (Harris, 1992; Goldman, 2006) to the acquisition of mindreading. She does not think that human children are little scientists who learn mindreading by generating and testing explicit hypotheses on their own (theory-theory), nor does she think that children learn mindreading on their own by combining their introspective knowledge of the contents of their own minds with their capacity to imagine themselves in others' shoes (simulation). Instead, mindreading — mental literacy, as Heyes would have it — is a cognitive gadget that children culturally inherit by social interactions with knowledgeable teachers through a combination of vertical, oblique and even horizontal routes of cultural transmission.

Heyes argues for this provocative thesis in two complementary steps. Her first aim is to dispel what she takes to be a widespread but misplaced suspicion that three distinctive features of mindreading show it to be genetically inherited: neural specialization,

[3] Cf. the famous Sally-Anne task first designed by Wimmer & Perner (1983) and adapted by Baron-Cohen et al. (1985).

developmental disorders and alleged cultural universality. Her second aim is to offer positive evidence that children are taught to read minds by conversations with knowledgeable adults. We address both of these below. In particular, we shall argue that the fact that success on verbal false-belief tasks reflects exposure to adult speech does not demonstrate that false-belief understanding itself is taught through conversation. Nor does the existence of cross-cultural variations in the acquisition of words for psychological states demonstrate cross-cultural variations in mindreading capacities. More generally, much of Heyes's interpretation of this evidence seems to conflate mindreading with folk or commonsense psychology. The former is a 'mill', but the latter is a 'grist', not a 'mill' (cf. Heyes, 2012; 2018).

2.1. Three analogies

As a way to undermine what she views as deceptive evidence for genetic inheritance, Heyes highlights three supposed analogies between literacy and mindreading. We here observe that these parallels are not as deep as they might first seem.

Neural specificity. As Heyes points out, the case of literacy shows that neural specificity is compatible with cultural inheritance: evidence of neural specificity of a cognitive process is not in and of itself evidence of genetic inheritance. Nonetheless there still is an important difference here, which Heyes does not explicitly recognize: namely, while it is now well-established that in the process of acquiring the ability to read, we expropriate a part of the visual cortex that has otherwise been used to recognize patterns, there is no comparable story for reading minds.

There is much evidence based on functional magnetic resonance imaging (FMRI) studies that mindreading is supported by brain activity in the area known as the right temporo-parietal junction (RTPJ) (cf. Saxe & Kanwischer, 2003; Saxe & Wexler, 2005). Some early evidence based on FMRI also suggested that activity in the RTPJ supports attention to unexpected stimuli in general. However, current evidence favors the hypothesis that the main

function of the RTPJ is to sustain mindreading, not that the evolved function of the RTPJ is to sustain general attentional mechanisms, which would be recycled for mindreading during ontogeny, as happens with the visual form area (cf. Young et al., 2010). Moreover, from what is known about the neural bases of mindreading, the relevant brain areas seem also to perform that task from early on in ontogeny. For example, application of non-invasive brain imaging techniques (functional near-infrared spectroscopy) in human infants has recently shown that the very same brain areas, i.e. the temporal-parietal junction in the right hemisphere (RTPJ), which are active when adults watch false-belief scenarios played on videos and are known to be active in adults' false-belief attribution, are also active in 7-month-old infants when they watch the same videos as the adults (cf. Hyde et al. 2018). This is a different story to the case of literacy, where, as we said, it is well-established that acquisition of the cognitive gadget involves the expropriation of one part of the brain that is otherwise used for other purposes (cf. Dehaene & Cohen, 2007, 2011).

Atypical development. According to Heyes, dyslexia is to literacy what autism spectrum disorders are to mindreading: genetically heritable developmental disorders that interfere with the acquisition of a culturally inherited skill (Ramus & Fisher, 2009). We do not dispute that individuals with autism spectrum disorders are impaired in tasks of mindreading, and that autism spectrum disorder exhibits substantial genetic heritability. Nor do we dispute, of course, the fact that literacy is culturally, not genetically inherited. So, we agree that a genetically inherited disorder may interfere with a culturally inherited skill. However, we do not see that this provides any substantive evidence *in favour of* the mindreading-as-cognitive-gadget hypothesis, simply because the fact that autism spectrum disorder is genetically heritable is wholly compatible with both genetic and the cultural inheritance of mindreading.

Cross-cultural diversity. The cultural dimension seems on the surface to be a clear

point of difference between reading minds and reading print. There is clear and manifest cross-cultural diversity in writing and reading systems (ranging from Egyptian hieroglyphs to Arabic, Chinese and Japanese systems), while mindreading seems to exhibit much more cross-cultural unity: for example, so far as is known there are no populations where adults understand the difference between others' knowledge and ignorance but lack false-belief understanding. Addressing this putative difference, Heyes draws attention to the existence of some cross-cultural variability in the acquisition of words for epistemic psychological states. She mentions in particular cross-cultural developmental studies by Shahaeian et al. (2011) showing that children from different cultures go through slightly different stages of cognitive development.

In a seminal study, Wellman & Liu (2004) demonstrated a five-step "theory-of-mind scale" whereby children from a variety of cultures master some verbal mindreading tasks before mastering others. The scale includes a diverse-desires task, a diverse-beliefs task, a knowledge/perceptual-access task, a false-belief task, and a hidden-emotions task, and most children from the US, Canada, Australia and Germany were found to succeed on these verbal tasks in this order (Peterson et al., 2005), and the same pattern of development has been shown to be exhibited by congenitally deaf children born of hearing parents (after being introduced to sign-languages later in childhood than other children) (cf. Peterson & Wellman, 2009; Peterson et al., 2005).

In contrast, Shahaeian et al. (2011) found children from cultures with high collectivist values (e.g. China, Iran) to succeed on "knowledge-access" verbal tasks before they succeed on "diverse-beliefs" verbal tasks, while children from more individualistic cultures (e.g. Australia, the US) tend to show the reverse pattern. In a typical "diverse-beliefs" verbal task, participants see a toy figure of a girl called Linda and a sheet of paper with bushes and a garage drawn on it. They are told that Linda wants to find her cat and that the cat might be

hiding in the bushes or it might be hiding in the garage. They are asked whether they think that the cat is in the bushes or in the garage. This is the own-belief question. Depending on what they answered, they are told that Linda holds the opposite belief. Finally, they are asked to predict where Linda will look for her cat. In a typical knowledge-access verbal task, participants are shown that a box with a drawer contains a toy dog inside. Then they are introduced to a toy figure of a girl called Polly and told that Polly has never seen inside the above drawer. They are asked if Polly knows what is inside the drawer. While children from more collectivist cultures solve the knowledge-access task first, children from more individualist cultures solve the diverse-beliefs task first. Heyes presents this as evidence for cultural variations in the mill of mindreading. But is it really?

On the one hand, it is likely that the frequency of lexical items for distinct epistemic states (e.g. 'know' and 'believe') in adults' speech does reflect some relevant differences between individualist and collectivist cultures. In this light, it is not surprising that children first acquire the lexical item that adults use more frequently in their respective community. If so, it is also not surprising that they succeed on verbal tasks that require them to know the meanings of those lexical items that are more frequently used in their respective community.[4] On the other hand, apart from this relatively minor departure between children from respectively more collectivist and more individualist cultures, there is a robust cross-cultural pattern of development among preschool children (Wellman & Liu, 2004; Peterson et al., 2005; Peterson & Wellman, 2009). Over a wide spectrum of different linguistic and cultural communities, success on diverse-desires verbal tasks (which probe understanding of whether

[4] To be clear, we are only claiming that it is possible (not ruled out by extant data) that cultural variations in the development of mindreading reflect variations in the frequencies of certain mental state terms across cultural-linguistic groups. We are not claiming this has been experimentally demonstrated. Further experimental work would be needed to test this possibility.

different individuals can have different desires) precedes success on diverse-beliefs and knowledge-access verbal tasks (which probe, respectively, understanding of whether different individuals can have different beliefs, and whether different individuals can have different knowledge states), which in turn precede success on verbal false-belief tasks (cf. Westra & Carruthers, 2017).

In sum, there is some cross-cultural variation in the acquisition of words for mindreading, but (1) this variation is not large relative to general, culturally consistent trends, and (2) the existing evidence for variation is evidence only for variation in the words used for mental states and not *ipso facto* for mindreading processes themselves.[5] Given these points, it is at best hugely premature to suggest that there is *bone fide* cross-cultural variation in mindreading abilities.

2.2. Are children taught mindreading through conversation?

As a second challenge to Heyes's arguments in favour of the mindreading-as-cognitive-gadget hypothesis, we now turn to the question of how children might be taught to read minds — as they must be, if reading minds is like reading print. (No child learns to read words without at least some scaffolded learning, after all.) Addressing this topic, Heyes appeals to four kinds of empirical evidence that, taken together, are meant to show that "conversation about the mind" teaches children, not just mental state *labels*, but mental state *concepts* themselves. The idea, in Heyes's (2018: 154) own words, is that children "inherit from their parents and other mindreading experts […] mechanisms that are specialized for the representation of mental states."

The four kinds of empirical evidence that Heyes appeals to are as follows. (i) The correlation in performance on explicit false-belief tasks has been shown to be the same in

[5] As we shall argue shortly, it might also reflect differences in folk psychological theories, which should not be confused with differences in mindreading.

pairs of both identical and non-identical 5-year-old twins (Hughes et al., 2005). (ii) The performances on both mental state language and false-belief tasks of a first cohort of adult deaf users of Nicaraguan Sign Language have been shown to be significantly worse than the performances of a second cohort of adult deaf users of the same language, by which time the language had evolved many more labels for mental states than previously (Pyers & Senghas, 2009). (iii) Children in Samoa, where it is improper to talk about mental states, have been shown to pass verbal false-belief tests much later than in Europe and North America (Mayer & Träuble, 2013). (iv) The performances of healthy young children on various verbal mindreading tests (mainly their ability to productively use terms for mental states) have been shown to be correlated with their exposure to maternal conversation (Taumoepeau & Ruffman, 2006; 2008); and conversely deaf children of hearing parents, who are delayed in their exposure to language, have been shown to be delayed in verbal false-belief performance (Peterson & Wellman, 2009).[6]

There is, however, an alternative — and we believe more parsimonious — interpretation of all these data. All these studies measure mindreading capacities using verbal (or signed) false-belief tasks and knowledge of linguistic terms (including terms in sign

[6] It is an open question what the full effects of deafness are on the social cognition of human infants and children. For example, in a non-verbal false-belief test based on anticipatory looking, Meristo et al. (2012) found a contrast between the performance of typically developing hearing 23-month-old toddlers and deaf 23-month-old toddlers of hearing parents who lacked proficiency with sign language. The hearing toddlers accurately gazed in both the true- and the false-belief conditions, but deaf children gazed at the non-empty location in both the true- and the false-belief conditions. These findings might be taken to vindicate Heyes's claim that children learn to read others' minds from conversation with knowledgeable adults; but another possibility is that these findings reflect a more immature capacity to disengage visual attention from a target's actual location in deaf toddlers than in hearing toddlers (cf. Southgate & Vernetti, 2014). This is clearly a topic where further research is needed.

language) for mental states. There can be little doubt that knowledge of linguistic terms for mental states depends on exposure to others' conversations about mental states, and so in order to succeed on verbal false belief-tasks, participants must also understand when mental state attribution is pragmatically relevant to conversations, including questions about others' actions. Thus, one straightforward interpretation of these findings is that they show how performances on verbal false-belief tasks and tasks about knowledge of linguistic terms for mental states can depend on exposure to others' speech about mental states. If so, then performances on either verbal false-belief tasks or on tasks about knowledge of linguistic terms for mental states should not be taken to directly reflect mindreading competences.

As an example, consider the findings by Taumoepeau & Ruffman (2006, 2008) showing that young children are less exposed to adults' speech about beliefs than about desires. It has been argued that as a result, young children are not likely to take the utterance of a complex belief-sentence (such as 'Mara believes that the concert starts at 8:00PM') as an attribution of belief to an agent. Instead they are likely to interpret such utterances as a speaker's cautious endorsement of the truth of the embedded clause on the model of 'I believe that the concert starts at 8:00PM' (cf. Helming et al., 2014; Lewis et al., 2012; Westra, 2017; Westra & Carruthers, 2017). If so, then what exposure to others' speech about beliefs is likely to teach language learners is not mindreading proper (the mill), but rather the conditions in which belief attribution is pragmatically relevant and appropriate to conversation.

Following what McGeer (2007) has called "the regulative role of folk psychology," Heyes (2018: 159) also points out that reading minds and reading words share a dual function. On the one hand, they have the interpretive function of extracting meaning from either behavior or written signs. On the other hand, they have a regulatory or regulative function. Just as children are taught the spelling and writing conventional norms governing their written language, they are similarly taught social norms and conventions, for example that human

behavior (including their own) should be produced by rational interactions among beliefs and desires and should be justifiable by appropriate reasons.[7] Heyes (2018: 41-42) further illustrates this regulative role of mindreading by reference to what she calls an "honor theory of mind": "a theory in which the desire to retaliate against insults is an important source of motivation." As shown by social psychological studies (cf. Cohen & Nisbett, 1994; Cohen, Nisbett, Bowdle & Schwartz, 1996), wide cultural variations are to be expected to underlie respectively the acceptance and the rejection of an "honor theory of mind." For example, such an "honor theory of mind" has been shown to be significantly more prevalent among individuals from the Southern rather than the Northern states in the US. And indeed, what Heyes calls the "honor theory of mind" includes social and behavioral norms whose acceptance varies widely across different cultures.

But in any case, explicit theories in this sense are not part of mindreading proper. Instead they belong to what McGeer (2007) and philosophers generally call commonsense or *folk psychology*, i.e. explicit psychological theories, some of which deal with social norms or conventions and others with the ontological relations between individuals' minds, brains and bodies. In his seminal (1956) essay, the philosopher Wilfrid Sellars entertained the influential idea that the commonsense ontology of mental states was created by some early modern human who invented folk psychology, i.e. a set of explicit theories whose purpose is to explain an agent's observable behavior by appeal to the agent's unobservable mental states. Folk psychology underlies ontological disputes about such topics as the mind-body problem and about whether reason explanations of human actions are causal explanations as well. But philosophers engaged in such controversies all share the same mindreading capacities. Nor is there any good reason to think that knowledge of explicit folk psychological theories is

[7] While norms governing writing are likely to be conventional, it is quite unlikely that norms of reasoning and justifications are equally conventional.

required to attribute mental states to others. As Heyes (2018: 10-12, 38-39) reminds her readers on several occasions, explicit folk psychological theories are the grist, not the mill, of human social cognition. But her treatment of mindreading betrays her own admonition. The mill is the mindreading capacity to attribute mental states to self and others: its job is neither to stipulate social norms or conventions nor to solve the mind-body problem.

In short, the evidence that Heyes brings to the table shows that children learn from adults much of their folk psychology (e.g. their "honor theory of mind" or their commitment to either ontological materialism or dualism), but it fails to show that they learn mindreading itself in the same way.

2.3. The threat of circularity

We now turn to the threat of circularity that lurks behind Heyes's proposal. According to the mindreading-as-cognitive-gadget hypothesis, the cultural evolutionary emergence of both non-mental literacy (reading words) and mental literacy (reading minds) rests on the presence of spoken natural languages. This is clearly uncontroversial in the case of reading words: if humans did not speak natural languages the capacities to read and write could simply not emerge. The case of reading minds, however, is far less straightforward. As Heyes herself recognizes (2018: 166), the idea that children learn mindreading through conversation with competent others must be at least somewhat controversial because there is a clearly viable alternative (which we actually favour), namely that word learning rests on mindreading (see e.g. Bloom, 2000, 2002; Sperber, 2000). Heyes (2018, chapter 8) also rejects the Chomsky-Pinker view that the human language faculty is an instinct (not a gadget). As a result of these joint commitments she faces the twin burden of showing that (i) children can learn to read minds by conversation and that (ii) language can be a cognitive gadget. In what

follows, we will ignore claim (ii) and highlight the peculiar circularity involved in claim (i) that children are taught to read minds in the way they are taught to read words.[8]

Indeed, Heyes herself notes that "mindreading is important in relation to cultural evolution because it plays a crucial role in teaching [literacy]" (2018: 145) in at least three respects. First, teaching is a process whereby an agent acts "with the intention of producing an enduring change in the mental states […] of another agent." Secondly, mindreading "allows teachers to represent the extent and limits of a pupil's current knowledge" (*ibid*.). Thirdly, "in a complementary way, mindreading by pupils enables them to isolate what it is that a teacher intends them to learn" (*ibid*.). In other words, teaching literacy involves a sequence of verbal and non-verbal communicative acts, whereby the teacher's goal is to cause her pupil to acquire new beliefs about e.g. the visual shapes of spoken words by various verbal and non-verbal demonstrations. The teaching succeeds if the pupil fulfills the teacher's intention by acquiring the relevant beliefs that the teacher wishes (or intends) him to acquire, namely beliefs about what the shapes and contours of spoken words visually look like when they are written. In short — and as Heyes comes close to saying herself — young children could not learn to read words unless they could read their teachers' minds. But if so, then Heyes's proposal would turn out to be circular. In any case, children could *not* learn to read minds the way they learn to read words.

In fact, Heyes's position about the role of mindreading in the acquisition of literacy (and cultural learning, more generally) is demonstrably ambivalent. As noted, Heyes (2018, p. 145) clearly grants mindreading to knowledgeable adults by recognizing that mindreading "allows teachers to represent the extent and limits of a pupil's current knowledge." What

[8] Notice that by the age most children pass verbal false-belief tasks (which Heyes takes to be a hallmark of mindreading), they are still unable to read words fluently — a disparity that surprisingly does not seem to worry Heyes.

about pupils? At this point, Heyes faces the following dilemma: either pupil-mindreading is required for learning literacy or it is not. There is textual evidence (mentioned above) that Heyes (2018, p. 145) is willing to acknowledge that mindreading enables pupils "to isolate what it is that a teacher intends them to learn." If so, then clearly young children cannot learn to read minds the way they learn to read words and Heyes's analogy breaks down.

Suppose now that Heyes does not require pupil-mindreading for learning literacy. Then Heyes's burden is twofold. First, by denying pupil-mindreading, she commits herself to an asymmetrical picture of learning literacy via a process of direct instruction whereby teacher-mindreading is required, but pupil-mindreading is not. If so, then she must explain how pupils could fulfill their teachers' communicative intention without mindreading.[9]

Secondly, she must also explain away growing experimental evidence showing not only that preverbal infants are sensitive to ostensive non-verbal communicative actions directed to them, but also that when presented with a non-verbal ostensive communicative interaction between an agent and a recipient, from a third-person perspective, preverbal infants are sensitive to the agent's communicative intention.

Regarding this second burden, we focus on studies by Martin et al. (2012) and Voulomanos et al. (2014) with 12- and 6-month-old infants respectively (but see also the work of Tauzin & Gergely (2018), among others, for similar findings). In familiarization trials, infants first saw the speaker alone repeatedly select one of a pair of toys. Then they saw the addressee alone play with both toys. In the test trials, the speaker's head appeared through a narrow window while only the addressee could reach for the toys. In the test trials of the speech condition, the speaker uttered 'koba' while looking at the addressee. In the test trials

[9] Heyes (2016) offers some reasons for thinking that eye contact, contingencies, infant-directed speech, gaze cuing, and rational imitation might not be genetic adaptations for teaching. But she does not meet the specific challenge of explaining how pupils could fulfill their teachers' communicative intention without mindreading.

of the cough condition, the speaker coughed while looking at the addressee. An utterance of 'koba' is not an utterance of an English word, but unlike coughing it shares many acoustic properties with utterances of English words. Furthermore, while coughing is not a reliable cue of a speaker's communicative intention, speech is. What Martin et al. (2012) and Voulomanos et al. (2014) report is that only in the speech condition did the infants look longer when the addressee gave the speaker the toy she did not select rather than the toy she earlier selected. This suggests that by the age of 6 months infants can detect the presence of an agent's communicative intent and form expectations based on the familiarization trials about what it takes for the addressee to fulfill the speaker's request. If this is so, then it is likely that when they learn the meanings of spoken words of their native tongue, and a fortiori when they learn to read words, children can read their teachers' minds.

To recap. On Heyes's explicit account, mindreading capacities on the teacher's side are crucial for ensuring the success of the communicative interaction whereby she contributes to the pupil's learning to read words. If Heyes denies the role of pupil-mindreading, then she seems committed to a puzzling asymmetrical picture of learning literacy and she also faces counter-evidence. If Heyes grants that mindreading capacities on the pupils' side play some role — any role at all —, then she must recognize that human children cannot learn to read minds in exactly the way they learn to read words. This much seems uncontroversial. But it also opens the door to the admission that learning to read words might differ in indefinitely many ways from learning to read minds. Furthermore, if teaching others to read minds requires mindreading capacities on the teacher's side, then it is a puzzle how mindreading could have evolved at all from a phylogenetic point of view.[10]

[10] We are grateful to an anonymous reviewer for highlighting this latter problem and also for his or her very useful comments that have helped us clarify the dilemma faced by Heyes about the role of pupil-mindreading.

When she was earlier presented with Strickland & Jacob's (2015) first version of this argument about circularity (introduced in her 2014 *Science* paper with Chris Frith), Heyes' (2015b) response was to acknowledge that her "cultural evolutionary account of the origins of mindreading is essentially a bootstrapping story, …" and that "… from a distance, bootstrapping looks an awful lot like circularity." So far as we know, Heyes has not yet supplied any of the details needed to disentangle the relevant bootstrapping strategy from circularity, when looked at from close-up.

3. How likely is it that infants submentalize?

Mindreading has been a topic of much experimental investigation for over 40 years and in diverse groups, including healthy human adults, adults with autism spectrum disorder, human children, preverbal human infants, and non-human animals (including non-human primates). Thus, in addition to presenting positive arguments in favour of the mindreading-as-cognitive-gadget hypothesis (see above), Heyes also offers alternative low-level interpretations of many of the findings from these literatures, which others take to support views inconsistent with Heyes's own. In particular, Heyes has reinterpreted key experimental studies that other psychologists have taken to provide evidence of mindreading in preverbal infants and non-human primates: two groups that should not possess mindreading abilities, if the mindreading-as-cognitive-gadget hypothesis is correct. As part of these reinterpretations, Heyes has introduced and developed the idea of *submentalizing*, according to which infants' (or apes') responses to stimuli that appear to be the product of mindreading are in fact driven by genetically inherited, domain-general mechanisms that simulate the effects of mentalizing in social contexts but do not in fact involve the representation of others' mental states.

Here, we critique this proposition in three stages. First, we supply general background to present debates in developmental psychology, which constitute the main context for

Heyes's thesis (§3.1). Second, we highlight a fundamental tension in Heyes's reinterpretation of one famous study in this area (§3.2). Third, we argue that findings from a key study by Kovács, Téglás and Endress (2010) fail to support one crucial component of Heyes's account, namely retroactive interference (§3.3). We conclude that the mindreading-as-gadget hypothesis cannot perspicuously explain extant data.

3.1. Submentalizing within the developmental study of mindreading

In recent years, the developmental investigation of mindreading has focused on a major discrepancy in reported findings between verbal and non-verbal false-belief tasks. On the one hand, preschoolers reliably fail verbal false-belief tasks, in which they learn the location of a toy that a mistaken agent wants to retrieve and are then directly asked to predict where the mistaken agent is likely to look for it. On the other hand, data from non-verbal tests, in which infants observe a scene and the dependent variables are measures of surprise or expectation (e.g. looking time, anticipatory looking) suggest that preverbal infants seem to expect an agent to act in accordance with the content of her belief, regardless of whether that belief is true or false.

It is highly debated among developmental psychologists whether or not the findings of non-verbal false-belief tests should be regarded as reliable evidence for false-belief understanding. Aside from the different theoretical accounts, to which we shall turn in a moment, one key reason why these debates have fuel is that the replicability and validity of several non-verbal false-belief tests has been called into question. There is growing recognition — in psychology as a whole — of the negative long-term impact of the so-called "file-drawer problem" (i.e. negative results going unpublished) and other sources of publication bias, and as part of this wider movement there are now many attempted replications of key studies in the literature on infant mindreading. While these attempted replications have produced mixed results (see e.g. the papers collected by Sabbagh & Paulus,

2018), we think that the reproducibility of scientific findings is a delicate issue.[11] Agreeing with the authors of many of the failed replications, we consider the possibility of infant mindreading as "firmly theoretically grounded and motivated" (Poulin-Dubois et al., 2018, p. 308) but empirically open: "The stance we are advocating is not skeptical in the sense of *denying* early rich ToM [mindreading], but in the sense of considering it an open empirical question" (*ibid.*, italics in original). As the current exchanges among developmental psychologists clearly show, further empirical work is critically needed to explore the possible reasons for failed replications (cf. Baillargeon et al., 2018).[12]

With the empirical issue open, the theoretical burdens are distributed. Researchers whose inclination is to take the results of non-verbal false-belief tests at face value (i.e. as evidence of false-belief attribution in human infancy) incur the burden of explaining why verbal false-belief tasks are so challenging for preschoolers (cf. Baillargeon et al., 2010). Meanwhile those who are skeptical about the results of non-verbal false-belief tests face the burden of offering adequate non-mentalistic accounts of the infant data (cf. Perner and Ruffman, 2005; Perner and Roessler, 2012). Depending on the results of future studies, the skeptics' burden might be light, if the supposed findings of false-belief attribution in infancy turn out to be driven largely by false positives, or else heavy, if these findings turn out to reveal a real effect. Either way, Heyes's idea of submentalizing is an attempt to address this issue and its cogency can be assessed on its own terms.

---

[11] Cf. Baker (2016).

[12] For example, Powell et al. (2018) report that they attempted and failed to replicate the results of Onishi and Baillargeon's (2005) seminal study based on the violation-of-expectations, which was the first to provide evidence of early false-belief understanding in human infancy. In their response, Baillargeon et al. (2018) highlight several aspects of the design of the familiarization trials of Powell and colleagues' (2018) study that might have prevented infants from forming expectations about the goal of the agent's action.

Thus, before we address Heyes's submentalizing approach, we wish to say a few words about how we propose to meet the burden (i.e. our own burden) of explaining why children below the age of 4,5 should reliably fail verbal false-belief tasks (in the non-random systematic way that they do), if they are in fact able to represent the contents of others' minds? Advocates of the two-systems approach to mindreading (Apperly & Butterfill, 2009 and Butterfill & Apperly, 2013) have proposed to answer this question and resolve the discrepancy between the developmental findings by crediting human infants with minimal (not full) mindreading capacities, i.e. domain-specific cognitive resources dedicated to the attribution of a restricted range of relational mental states such as an agent's registration. Heyes (2018: 157) is unconvinced that such minimal mindreading enables infants to represent genuine "mental states, rather than purely observable, behavioral states." In our view, we should look to the pragmatics of verbal false-belief tasks to explain why children below the age of 4,5 reliably and non-randomly fail these tasks, in the systematic way that they do, i.e. they point to the toy's actual location.

It is uncontroversial that false-belief understanding cannot be sufficient for success on verbal false-belief tasks. In other words, verbal false-belief tasks probe more than just false-belief understanding (Bloom & German, 2000). But what else? Trivially, success on these tasks requires knowledge of the syntax and semantics of the language used by the experimenter to first tell the false-belief story and then to ask the question about the mistaken agent's likely action — but it also requires understanding the pragmatics of the key test question used in verbal false-belief tasks (cf. Carruthers, 2013; Helming et al., 2014; Westra, 2017; Westra & Carruthers, 2017). Children are the recipients of information that is verbally conveyed by an experimenter and they are also directly asked a question by the experimenter. How do they conceive of the motives and functions of these interactions? It is quite possible that young children might be pragmatically misled into wrongly thinking that all of the

information that has been verbally conveyed to them by the experimenter is relevant to answering the experimenter's question. In particular, it is possible that they assume that information about the toy's actual location, which is strictly irrelevant for the prediction of the mistaken agent's action, must be relevant since this information has been verbally conveyed to them by the experimenter (*ibid.*). If so, then young children's answers in verbal false-belief tasks might be a product of their trying to interpret this irrelevant information as relevant in some way, causing them to fail to correctly answer the experimenter's question, even though they have in fact correctly represented the agent's (false) belief. For example, they might interpret the experimenter's question, not as a prediction question, but instead as a normative question "Where should the mistaken agent look for her toy?" to which the right answer is the toy's actual location (cf. Helming et al., 2014).

In several publications addressing these issues, Heyes (2014a, 2014b, 2015a, 2018) has taken on the opposite burden, namely to offer a non-mentalistic deflationary account of findings based on non-verbal false-belief tests consistent with infants' presumed inability to attribute false beliefs to others. She has argued that not only mentalizing (whether full-blown or minimal in the sense of Apperly and Butterfill's two-systems approach), but also behavior-reading, all rest on domain-specific cognitive resources (dedicated to processing *social* stimuli), which are not part of the domain-general "start-up kit" genetically inherited by human infants. This "start-up kit" is domain-general in the sense that its computations are taken to apply to social and non-social stimuli alike. This, then, is what she calls *submentalizing*: the human capacity to predict human behavior on the basis of low-level domain-general psychological processes that simulate the effects of mentalizing in social contexts. Thus, it is of crucial importance to Heyes's cognitive gadget view of mindreading that far from reflecting infants' expectations about an agent's mental states (or even behavior), all of the infants' responses in non-verbal false-belief tests should be explainable

by some low-level, non-social, perceptual processes.

Attempting to meet this requirement, Heyes (2014a) has argued in details for her *perceptual novelty* hypothesis, according to which the infants' looking behavior in non-verbal false-belief tests reflects the degree to which the low-level properties (the colors, shapes and movements) of test events depart from the low-level properties of the earlier events encoded and remembered by the infants. More recently, Heyes (2017) has proposed to apply her perceptual novelty hypothesis to findings based on anticipatory gaze and reported by Krupenye et al. (2016), which they interpret as evidence that non-human apes have some understanding of others' false beliefs. Krupenye et al. (2017) have subsequently tested and claimed to refute Heyes's (2017) proposal.

Some of Heyes's interpretations of false-belief tests involving very young children sound like requests for further controls. For example, in a famous false-belief study based on anticipatory gaze by Southgate et al. (2007), 25-month-olds watched videos depicting a human adult whose head appeared above a pair of windows at the bottom of each of which was an opaque box. In the familiarization trials, the agent watched as a puppet bear placed a ball into one of the boxes and then the puppet disappeared. As soon as the puppet disappeared, both windows above the boxes were illuminated and a chime sounded simultaneously, providing children with a cue that the actor was about to open a window and retrieve the ball from its box. Less than 2 seconds later, the agent reached for the ball. In the test trials, as the agent was watching, the puppet placed the ball into the left box, then into the right box; then the puppet closed the lid of the left box and disappeared. At this point, the sound of a phone rang and the agent turned her head away. While the agent was not looking, the puppet reappeared, opened the right box, retrieved the ball, closed the lid, and left. After the puppet disappeared, the phone stopped ringing, the agent turned back, the windows were illuminated and the chime sounded. Toddlers' first anticipatory eye-movements were

recorded. They were found to reliably gaze at the right window above the box where the agent had last seen the puppet place the ball. Southgate et al. (2007) take this as evidence that 25-month-olds attribute a false belief to the agent. Heyes (2014a, 2018: 159-160) provocatively argues that the toddlers' attention might have been so disrupted by hearing the phone ring and seeing the agent turn her head away that they were not aware that the puppet had taken the ball away. If so, then they might have gazed at the right box because they thought that the ball was there, not because that is where they thought the agent thought the ball was.

While Heyes's non-mentalistic account of the findings based on anticipatory gaze is a request for further controls, her explicit account of the infant data based on looking time rests on the interplay between two kinds of domain-general cognitive processes. On the one hand, Heyes's low-level perceptual novelty account commits her to the assumption that infants can only encode a highly *restricted* subset of the full set of properties (the *low-level* properties) that are instantiated by both the test events and the earlier events perceived by the infants. On the other hand, Heyes appeals to the phenomenon of so-called *retroactive interference*, whereby the infants' perception of some of the specific *later* events *impairs* their *memory* of some of the *immediately preceding* events in the sequence.[13]

As Heyes (2014a, p. 647) rightly says about the infant data, "the devil is in the details." The question is: can the details of her own low-level account perspicuously explain the relevant data?

3.2. The water-melon, the green and the yellow box

Onishi and Baillargeon's landmark study (2005) involves three temporal stages: a set of familiarization trials, a set of belief-induction trials, and a pair of test trials. In the first of

[13] For an earlier critique of Heyes's (2014a) explanation of the infant data based on her perceptual novelty account, cf. Scott and Baillargeon (2014).

three successive familiarization trials, 15-month-olds first saw a melon toy surrounded by a pair of opaque boxes (a green and a yellow box) against the background of a closed window. After the window was opened, the infants saw a human (female) agent manipulate the toy before placing it into the green box with her right hand. In the two following familiarization trials, they first saw the two opaque boxes against the background of the closed window. Then the window was opened and they saw the human (female) agent reach into the green box with her right hand (cf. Figure 1).

In a series of four distinct belief-induction trials generating a pair of true-belief (TB) conditions and a pair of false-belief (FB) conditions, the infants saw the toy either move (by its own self-propelled motion) from the green to the yellow box or not, either in the presence of the agent or not (cf. Figure 2). In the TB-green condition, the yellow box moved towards the green box and back to its initial position (as indicated by the arrow in Figure 2a), but the toy did not move and the agent was present. In the TB-yellow condition, the toy moved from the green to the yellow box in the agent's presence (as indicated by the arrow in Figure 2b). In the FB-green condition, the toy moved to the yellow box in the agent's absence (cf. Figure 2c). In the FB-yellow condition, the toy first moved from the green to the yellow box in the agent's presence and moved back into the green box in the agent's absence (cf. Figure 2d). Finally, in the test trials, the infants first saw the two opaque boxes at their previous location against the background of the closed window. The window opened and they saw the agent reach either for the green box with her right hand (green test event) or for the yellow box with her left hand (yellow test event) (cf. Figure 3).

Onishi and Baillargeon found that infants in the TB-green condition looked reliably longer at the yellow rather than at the green test event; infants in the TB-yellow condition looked reliably longer at the green rather than at the yellow test event; infants in the FB-green condition looked reliably longer at the yellow rather than at the green test event; and infants in

the FB-yellow condition looked reliably longer at the green rather than at the yellow test event. According to Onishi and Baillargeon's mentalistic interpretation, infants looked reliably longer either when the agent reached to the empty location with a true rather than a false belief or when she reached to the toy's actual location with a false rather than a true belief.

Heyes asks two questions about this interpretation of the data, and each of her answers has two components: the low-level perceptual novelty account and retroactive interference. There is, we believe, a deep tension between these two components.

Heyes's first question is: Why did infants in both the TB-green and the FB-green conditions look reliably longer at the yellow than at the green test event? On the mentalistic account, the answer to this question is: because infants expected the agent to reach for the box in which she believed (truly or falsely) the toy to be. Of course, this cannot be Heyes's non-mentalistic answer. Heyes's own answer is two-tiered. On the one hand, she argues that to infants in the TB-green condition (in which the toy stayed in the green box), the yellow test event must have looked perceptually more novel than the green test event, relative to their *familiarization* experience (in which on three repeated occasions they saw the agent reach for the green box). On the other hand, she must explain away the fact that infants in the FB-green condition saw the toy move to the yellow box (in the agent's absence).

The way Heyes cancels the effect of infants' seeing the toy move to the yellow box in the FB-green condition is by appealing to the phenomenon called *retroactive interference*. In the FB-green condition, the infants see the toy move to the yellow box, *in the agent's absence*. Heyes argues that the agent's unexpected reappearance in the test event retroactively interferes with infants' memory of the immediately preceding event in the FB-green condition: "their memory for this event was impaired because it was immediately followed by a salient distractor event – the unexpected reappearance of the agent at the beginning of the

test phase." In effect, she argues that retroactive interference cancels the difference between the TB-green and the FB-green condition. So in both cases, the yellow test event must have looked to infants perceptually more novel than the green test event, relative to their familiarization experience.

Heyes's second question is: why did infants in the TB-yellow and the FB-yellow conditions look reliably longer at the green than at the yellow test event? On the mentalistic account, the answer to this question is: because infants expected the agent to reach for the box in which she either truly or falsely believed the toy to be. Here too Heyes's answer, based on her low-level perceptual novelty account, is two-tiered. On the one hand, she argues that in the TB-yellow condition, "after familiarization and before the test, these infants saw an event (movement of the toy-shape towards yellow), that was visually similar to the yellow test event (movement of the agent-shape towards yellow), and therefore reduced the novelty of the yellow test event." As a result, these infants looked longer at the green than at the yellow test event. On the other hand, Heyes must also explain away the fact that in the FB-yellow belief-induction trial, after seeing the toy move to the yellow box in the agent's presence, the infants saw the toy move back to the green box in the agent's absence.

Here again, as with the first question, Heyes appeals to retroactive interference: she hypothesizes that the unexpected reappearance of the agent at the beginning of the test phase impairs infants' memory of the immediately preceding event whereby the toy moved back to the green box in the agent's absence. As a result, the difference between the TB-yellow and the FB-yellow conditions vanishes: in both cases, what matters is that after familiarization and before test, the infants saw "an event (movement of the toy-shape towards yellow), that was visually similar to the yellow test event (movement of the agent-shape towards yellow), and therefore reduced the novelty of the yellow test event."

We can now highlight the internal tension between Heyes's appeal to perceptual

novelty and her appeal to retroactive interference in her account of Onishi and Baillargeon's findings.

On the one hand, her low-level perceptual novelty account can only be satisfied if the novelty generating the surprise is present at a low representational level "where the events witnessed by the infants are represented as colors, shapes and movements, rather than as actions on objects by agents." For example, it is a critical assumption of the low-level perceptual novelty account that to infants the movement of the *toy* to the *yellow* box (in the TB-yellow belief-induction trial) is visually *similar* to the yellow test event whereby the *agent* reaches for the *yellow* box. In the TB-yellow belief-induction trial, infants see the toy move to the yellow box; they see the agent watch the movement of the toy; but they only see the agent' head, not her full upper body parts. In the yellow test event, however, they see the agent reach for the visible yellow box with her fully visible left arm, but they do not see the toy. So it is crucial to the low-level perceptual novelty account that infants be *blind* to the many differences between the event of the toy moving to the yellow box in the TB-yellow belief-induction trial and the yellow test event of the agent's reaching for the yellow box. Although it is logically possible that infants are indeed blind to these many differences, the psychological thesis that they are as a matter of fact blind to them clearly requires further independent experimental support.

But on the other hand, Heyes appeals to retroactive interference in both the FB-green and the FB-yellow conditions. So, infants must be vulnerable to the relevant instance of retroactive interference whereby the *agent*'s unexpected reappearance in the test event impairs their memory of the immediately preceding event in which the agent was absent. Only if the infants can encode the property of *being an agent* (capable of executing e.g. reaching arm movements) could they be vulnerable to relevant instances of retroactive interference. But according to the low-level perceptual novelty account, infants should not be

able encode the property of *being an agent*. (Remember: by Heyes's own criteria, even *behavior-reading* rests on domain-specific cognitive mechanisms that are unavailable to human infants who are limited to domain-general cognitive resources.) So there is a conflict between the two components of her non-mentalistic explanation of the findings reported by Onishi and Baillargeon (2005).

We are, of course, aware that our argument here, that there is a deep tension between the two components of Heyes's account of these findings is a request for further clarification rather than a straightforward refutation. But there is more to come.

3.2. The Blue Smurf and the ball

Another key study for the mentalistic interpretation of the false-belief literature in human infancy is Kovács, Téglás and Endress's (2010) smurf study. Here too the data seem to suggest that infants represent false-beliefs, to which Heyes has offered counter-interpretations (2014a, 2014b). Here we argue that her account entails a prediction, which, so far as we know, has not yet been tested and which we take to be unlikely to be confirmed. Here, we explain why.

In these studies with adults and infants, participants watch one of four versions of a video. In all four conditions, an agent (a blue Smurf) places a ball on a table in front of an occluder and then the ball rolls behind the occluder of its own self-propelled motion. From then on, the four different conditions diverge: participants see the ball either stay behind the occluder or leave and they see the agent leave the scene either before or after the ball reaches its final location. Finally, in all four conditions, the agent comes back and the occluder is lowered in his presence (cf. Figure 4).

In the adult study, participants were instructed to press a button as fast as possible in the final stage when the agent is back, if they saw the ball when the occluder was lowered down (which was the case 50% of the time in all four conditions). Kovács and colleagues

measured participants' reaction times in performing this task. Not surprisingly, they found that adults were significantly faster to respond in the P+A+ condition (when both participants P and the Smurf agent A expected the ball to be behind the occluder) than in the P-A- condition (when neither participants P nor the agent A expected the ball to be behind the occluder). The surprising finding was that they were also significantly faster in the P-A+ condition (when A falsely expected the ball to be there, but participants did not) than in the P-A- condition. Kovács and colleagues argue that this surprising finding is evidence that participants automatically computed the content of the agent's false belief.[14]

Heyes (2014b, p. 137) has offered a non-mentalistic alternative explanation of the adults' findings based on retroactive interference. This shows that Heyes treats putative evidence for automatic mindreading in adults under time pressure on a par with putative evidence for mindreading in infants. (This assumption will matter shortly.) In both the P-A- and the P+A+ conditions, the Smurf is present during the last event that is relevant to the participants' own expectations about the location of the ball. However, in the P-A+ condition, the Smurf is absent during the last event that is relevant to the participants' own expectations about the location of the ball, i.e. when the ball finally leaves the scene. Heyes argues that the "perceptually salient" reappearance of the Blue Smurf in the final stage of condition P-A+ (when the occluder is being lowered) is likely to have impaired adult participants' memory of the immediately preceding event (i.e. the last motion of the ball) by a process of retroactive

[14] Kovács and colleagues' mentalistic interpretation of their adult study based on reaction times has been criticized by Phillips et al. (2015) on grounds independent from Heyes's own interpretation based on retroactive interference. Phillips and colleagues have proposed that Kovács and colleagues' findings in their adult study reflect differences in the temporal intervals between two attention checks, not participants' computation of another's false belief. On the one hand, Phillips and colleagues' non-mentalistic critique does not extend to the infant study, but Heyes's interpretation does. On the other hand, Phillips and colleagues' non-mentalistic attention-check hypothesis has been tested and refuted in a recent study by El Kaddouri et al. (2019).

interference. Heyes's non-mentalistic account makes a straightforward prediction: in the P+A- condition, when participants see the ball finally move back behind the occluder, the Blue Smurf is also absent. According to Heyes's account, the salient reappearance of the Blue Smurf when the occluder is lowered should also impair participants' memory for the immediately preceding event where the ball returns behind the occluder. So Heyes's account predicts that participants should not expect the ball to be behind the occluder and therefore be significantly slower in the P+A- condition than in the P+A+ condition. But they are not.

The same critique extends to Heyes's (2014a) account of the infant study. Here Kovács and colleagues used infants' looking time and found that when in the final stage, the Smurf is back, the occluder is lowered down and *there is no ball*, 7-month-olds look longer in the P-A+ than in the P-A- condition. Kovács and colleagues argue that this is evidence that infants' looking time is influenced by their computation of the Smurf's false expectation that the ball should be there. Heyes on the other hand surmises that in the P-A+ condition, the Smurf was away when the infants last saw the ball leave the scene. So; she hypothesizes that far from reflecting infants' computation of the Smurf's false belief, the Smurf's unexpected reappearance impaired the infants' memory of this last event by retroactive interference. As a result, infants looked longer in the P-A+ condition than in the P-A- condition because they had forgotten the ball's last motion and expected the ball to be there.

This account predicts conversely that in the P+A- condition, infants should also forget the last event whereby the ball rolled back behind the occluder. If so, then they should *not* expect the ball to be there and should *not* look longer in the P+A- condition than in the P-A- condition upon finding out that there is no ball. So far as we know (but we may be wrong), this comparison has not been tested. It would be interesting to test Heyes's prediction that infants should not look longer in the P+A- than in the P-A- condition. On the assumption that the mechanisms that underlie the adults' responses are the signature of the mechanisms that

underlie the infants' responses,[15] we would expect the infants' looking time to be consistent with the adults' reaction times. Since adults were faster to detect the ball when it was there in the P+A- than in P-A- condition, we would conversely expect the infants to look longer when there is no ball in the P+A- condition than in the P-A- condition. This is why we take the prediction entailed by Heyes's account to be quite unlikely to be confirmed.

To sum up. Heyes's appeal to retroactive interference in her account of the adult data entails that adults should not expect the ball to be behind the occluder in the P+A- condition and therefore to be significantly slower in this condition than in the P+A+ condition — but this prediction is refuted by the data. Her account of the infant data entails that infants should look longer when there is no ball also in the P+A- condition rather than in the P-A- condition. However, in accordance with Heyes's own assumption that the adult data should be treated as the signature of infants' non-mentalistic responses, the refutation of her prediction in the P+A- condition of the adult study entails the improbability of her own prediction in the very same condition in the infant study.

Summarizing and concluding remarks

Heyes's intriguing idea of a cognitive gadget is best exemplified by human literacy and human numeracy: both are unquestionably inherited by human children through a combination of vertical, oblique and even horizontal routes of linguistically mediated cultural transmission, not by children's genetic inheritance from their biological parents. Moving beyond these uncontroversial examples, the gadget approach to human social cognition must be addressed case by case. In this paper, we have argued that the case for human mindreading

[15] An assumption Heyes (2014a, 2014b) must accept since she applies retroactive interference to both the infant and the adult studies.

is unconvincing. We have first dealt with Heyes's analogy between learning to read words and learning to read minds in four major steps:

• The fact that the specificity of the neural basis of mindreading and the heritability of autism spectrum disorders are compatible with cultural inheritance is not evidence *in favour of* the gadget approach to mindreading (§2.1).

• The evidence for cultural variation in mindreading is far weaker than Heyes presents it to be, because evidence for cultural variation in folk theories of mindreading is *not* evidence for cultural variation in mindreading itself (§2.1).

• The evidence taken to show that children learn to read minds by conversation can also — and we think more parsimoniously — be interpreted as evidence that children learn the pragmatic conditions in which belief attribution is relevant to conversation (§2.2).

• It is furthermore unlikely that children could learn to read minds as they learn to read words by being explicitly taught by knowledgeable adults, as Heyes herself recognizes that children could not learn to read words unless their teachers could read their pupils' minds. And it is at best unclear how adults with mindreading capacities could successfully perform their teaching communicative actions unless their pupils could recognize their teachers' intentions (§2.3).

Second, we agree with Heyes that mindreading could not be a cognitive gadget unless infants respond to social stimuli by submentalizing rather than by mentalizing (notwithstanding unresolved issues regarding the replication of key experimental studies). However, we have argued that the submentalizing approach leads to a deep tension between perceptual novelty and retroactive interference (§3.2) and that it also entails a prediction about infant mindreading that, although it has so far not been tested, is unlikely to hold because its adult counterpart has been refuted (§3.3).

In short, the human mind may well be filled with gadgets — but mindreading is

unlikely to be one of them.

References

Apperly, I. & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116, 4, 953–970.

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533, 26 May, 452-454.

Baillargeon, R., Scott, R.M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review in Psychology*, 67, 159-186.

Baillargeon, R., Buttelmann, D., & Southgate, Victoria (2018). Invited Commentary: Interpreting failed replications of early false- belief findings: Methodological and theoretical considerations. *Cognitive Development*, 46, 112-124.

Barkow, J. H., Cosmides, L., & Tooby, J. (eds.) (1992). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Oxford: Oxford University Press.

Baron-Cohen, S., Leslie, A. & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21, 37-46.

Barrett, H.C. (2014). *The Shape of Thought: How Mental Adaptations Evolve*. Oxford: Oxford University Press.

Bloch, M. (2012). *Anthropology and the cognitive challenge*. Cambridge: Cambridge University Press.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Bloom, P. (2002). Mindreading, communication and the learning of names for things. *Mind and Language*, 17, 1-2, 37–54.

Bloom, P. & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77, B25-B31.

Boyd, R., & Richerson, P. J. (1988). *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

Boyer, P. (2018). *Minds Make Societies*. New Haven: Yale University Press.

Brem, S., Bucher, K., Halder, P., Summers, P., Dietrich, T., Martin, E., & Brandeis, D. (2006). Evidence for developmental changes in the visual word processing network beyond adolescence. *NeuroImage, 29*(3), 822–837.

Butterfill, S. & Apperly, I. (2013). How to construct a minimal theory of mind? *Mind and Language*, 28(5), 606-637.

Butterworth, B. (2005). The development of arithmetical abilities. *Journal of Child Psychology and Psychiatry*, 46, 3–18.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Carruthers, P. (2013). Mindreading in infancy. *Mind and Language,* 28, 141–172.

Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon Books.

Cohen, D. & Nisbett, R. E. (1994). Self-protection and the culture of honor: explaining southern violence. *Personality and Social Psychology Bulletin*, 20, 551-567.

Cohen, D. Bowdle, B.F., Nisbett, R.E., & Schwarz, N. (1996). Insult, aggression, and the Southern culture of honor: an "experimental ethnography". *Journal of Personality and Social Psychology*, 70, 5, 945-960.

Dehaene, S. & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, 56, 384–398.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences, 15*(6), 254–262.

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development, 46*, 12–30.

Dretske, F. (1988). *Explaining Behavior*. Cambridge, MA: MIT Press.

El Kaddouri, R., Bardi, L., De Bremaeker, D., Brass, M., & Wiersema, J.R. (2019). Measuring spontaneous mentalizing with a ball detection task: putting the attention-check hypothesis by Phillips and colleagues (2015) to the test. *Psychological Research*,

Fodor, J.A. (1975). *The Language of Thought*. New York: Crowell.

Gallistel, C.R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford: Oxford University Press.

Harris, P. (1992). From simulation to folk psychology: the case for development. *Mind and Language*, 7, 120-144.

Helming, K.A., Strickland, B. & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences,* 18, 167–170.

Henrich, J. (2015). *The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.* Princeton: Princeton University Press.

Heyes, C. (2005). Imitation by association. In Hurley, S. & Chater, N. (eds.) (2005) *Perspectives on Imitation: From Neuroscience to Social Science*, vol. 1, Cambridge, MA: MIT Press, pp. 157-175.

Heyes, C. (2012). Grist and mills: on the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society* B2012, 367, 2181-2191.

Heyes, C. (2014a). False belief in infancy: a fresh look. *Developmental Science*, 17:5, 647–659.

Heyes, C. (2014b). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, 9(2) 131–143.

Heyes, C. (2015a). Animal mindreading: what's the problem? *Psychonomic Bulletin & Review*, 22:313–327.

Heyes, C. (2015b). The bigger picture. http://cognitionandculture.net/why-reading-minds-is-not-like-reading-words/

Heyes, C. (2016). Born pupils? Natural pedagogy and cultural pedagogy. *Perspectives on Psychological Science*, 11(2) 280–295.

Heyes, C. (2017) Apes Submentalise. *Trends in Cognitive Sciences*, 21(1):1-2.

Heyes, C. (2018). *Cognitive Gadgets, the Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press.

Heyes, C. and Frith, C. (2014). The cultural evolution of mindreading. *Science*, 344(6190), 1243091.

Hughes, C., Jaffee, S.R., Happé, F., Taylor, A., Caspi, A. & Moffitt, T.E. (2005). Origins of individual differences in theory of mind. From nature to nurture? *Child Development*, 76(2), 356-370.

Hyde, D.C, Simon, C.E., Ting, F. & Nikolaeva, J. (2018). Functional organization of the temporal-parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *Journal of Neuroscience,* 2 May 2018, 38 (18) 4264-4274.

Kovács, A., Téglás, E. & Endress, A. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, 330, 1830-1834.

Krupenye, C., Kano, F., Hirata, S., Call, J. & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354 (6308):110-4.

Krupenye, C., Kano, F., Hirata, S., Call, J. & Tomasello, M. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology*, 10(4) http://dx.doi.org/10.1080/19420889.2017.1343771

Kulke, Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development, 46*, 97–111.

Lewis, S., Hacquard, V. & Lidz, J. (2012). The semantics and pragmatics of belief reports in preschoolers. *Proceedings of SALT*, 22, 247–267.

Martin, A., Onishi, K.H., & Vouloumanos, A. (2012). Understanding the abstract role of speech in communication at 12 months. *Cognition*, 123(1), 50–60.

Mayer, A & Träuble, B.E. (2013). Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *International Journal of Behavioral Development*, 37, 21-28.

McGeer, V. (2007). The regulative dimension of folk psychology. In Hutto, D. & Ratfcliffe, M. (eds.) *Folk Psychology Re-assessed*, New York: Springer, pp. 137-156.

Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L. & Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science*, 15, 5, 633-640.

Millikan, R.G. (2004). *The Varieties of Meaning*. Cambridge, MA: MIT Press.

Morin, O. (2015). *How Traditions Live and Die.* Oxford: Oxford University Press.

Morin, O. (in press). Did social cognition evolve by cultural group selection? *Mind and Language.*

Onishi, C. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255–258.

Peterson, C. & Wellman, H. (2009). From fancy to reason: Scaling deaf and hearing children's understanding of theory of mind and pretence. *British Journal of Developmental Psychology*, 27, 297-310.

Peterson, C., Wellman, H., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development*, 76, 502–517.

Phillips, J., Ong, D.C., Surteees, A.D.R., Xin, Y., Williams, S., Saxe, R. & Franck, M.C. (2015). A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, 26(9), 1-15.

Pinker, S. (1994). *The Language Instinct*. New York: Morrow.

Pinker, S. (1997). *How the Mind Works*. New York: Norton.

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Horst Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B.,

Rafetseder, E., & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet – A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, 48, 302-315.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46, 40-50.

Pyers, J. & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science*, 20, 805-812.

Ramus, F. & Fisher, F.E. (2009). Genetics of language. In Gazzaniga, M. (ed.) (2009) *The Cognitive Neurosciences* (Fourth Edition). Cambridge, MA: MIT Press, pp. 855-872.

Saxe, R. & Kanwischer, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in **"**theory of mind." *Neuroimage*, 19(4):1835-42.

Saxe, R. & Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43(10):1391-9.

Scott, R.M., & Baillargeon, R. (2014). How fresh a look? A reply to Heyes. *Developmental Science*, 17:5, 660–664.

Sellars, W. (1956). Empiricism and the philosophy of mind. In Feigl, H. & Scriven, M. (eds.) *Minnesota Studies in the Philosophy of Science*, vol. I, Minneapolis, MN: University of Minnesota Press, 1956: 253–329.

Shahaeian, A., Peterson, C., Slaughter, V., & Wellman, H. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, 47, 1239–1247.

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

Southgate, V. & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, 130, 1-10.

Sperber, D. (1996). *Explaining Culture, a Naturalistic Approach*, Oxford: Blackwell.

Sperber, D. (2000.) Metarepresentations in an evolutionary perspective. In Sperber, D. (ed.) *Metarepresentations: A Multidisciplinary Perspective*. New York: Oxford University Press, pp. 117-137.

Sperber, D. (in press). Instincts or gadgets? Not the debate we should be having. Commentary on Cecilia Heyes (2018) *Cognitive Gadgets: The Cultural Evolution of Thinking* (To appear in BBS).

Strickland, B. & Jacob, P. (2015). Why reading minds is not like reading words. http://www.cognitionandculture.net/home/blog/44-pierre-jacobs-blog/2669-why-reading-

minds-is-not-like-reading-words

Taumoepeau, M. & Ruffman, T. (2006). Mother and infant talk about mental states relates to desire language and emotion understanding. *Child Development*, 77, 465-481.

Taumoepeau, M. & Ruffman, T. (2008). Stepping stones to others' minds: maternal talk relates to child mental state language and emotion understanding at 15, 24, and 33 months. *Child Development*, 79, 284-302.

Tauzin, T. & Gergely, G. (2018). Communicative mindreading in preverbal infants. *Scientific Reports, 8:9534* | DOI:10.1038/s41598-018-27804-4

Voulomanos, A, Martin, A. & Onishi, K.H. (2014). Do 6-month-olds understand that speech can communicate? Dev*elopmental Science*, 17, 6, 872-879.

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

Wellman, H., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75, 523–541.

Westra, E. (2017). Pragmatic development and the false belief task. *Review of Philosophy and Psychology*, 8, 2, 235-257.

Westra, E. & Carruthers, P. (2017). Pragmatic development explains the Theory-of-Mind

Scale. *Cognition*, 158, 165-176.

Young, L., Doddel-Feder, D. & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, 48, 2658-2664.