



The origins of mindreading: how interpretive socio-cognitive practices get off the ground

Marco Fenici¹ · Tadeusz Wieslaw Zawidzki²

Received: 2 May 2019 / Accepted: 11 February 2020
© Springer Nature B.V. 2020

Abstract

Recent accounts of mindreading—i.e., the human capacity to attribute mental states to interpret, explain, and predict behavior—have suggested that it has evolved through cultural rather than biological evolution. Although these accounts describe the role of culture in the ontogenetic development of mindreading, they neglect the question of the cultural origins of mindreading in human prehistory. We discuss four possible models of this, distinguished by the role they posit for culture: (1) the standard evolutionary psychology model (Carruthers), (2) the individualist empiricist model (Wellman, Gopnik), (3) the cultural empiricist model (Heyes), and (4) the radical socio-cultural constructivist model, which we favor. We motivate model (4) by arguing that many forms of mental state ascription do *not* serve the function of simply describing inner states causally responsible for the behavior of a cognitive agent; rather, they relate the agent to her environment by characterizing her practical commitments. Making these practical commitments explicit has an important regulatory function in that it supports action coordination and alignment on joint goals. We propose a model of how the ascription of mental states may have evolved as a linguistic device to perform exactly this function of making agents' practical commitments explicit.

Keywords Mindreading · Theory of mind · Social cognition · Sociocultural psychology · Evolutionary psychology · Constructivism · Folk psychology · Mental state ascription · Propositional attitudes

✉ Marco Fenici
marco.fenici@bilkent.edu.tr
Tadeusz Wieslaw Zawidzki
zawidzki@gwu.edu

¹ Department of Philosophy, Bilkent University, Ankara, Turkey

² Department of Philosophy, George Washington University, Washington, USA

1 The debate on the origin of mindreading

Recent years have seen increasing debate in philosophy and cognitive science on the nature, function, and evolution of mindreading—or “Theory of Mind” (ToM)—that is, the human capacity to attribute mental states (such as desires, intentions, and beliefs) to interpret, explain, and predict behavior. According to a widely shared naturalist nativist view, mindreading is an intrinsic component of the human biological endowment. Because of its adaptive value, it has been specified by natural selection within particular neurocognitive structures (Cosmides and Tooby 1992; Dunbar 1998), that can be damaged selectively (Baron-Cohen 1995). It is also innate, and this explains why it purportedly manifests already in early infancy (Baillargeon et al. 2016), and becomes more flexible and sophisticated due to its progressive integration with executive and linguistic abilities (Baillargeon et al. 2010; Helming et al. 2016; Westra and Carruthers 2017).

Against nativism about mindreading, some have argued for a more empiricist approach, and claimed that children progressively develop their knowledge of other minds. In particular, advocates of the child-as-scientist view have argued that children discover other minds by following an individual process of hypotheses production and testing that is akin to that followed by scientists in developing scientific theories (Gopnik and Meltzoff 1996; Wellman 1990).

Others have similarly claimed that mindreading is acquired during childhood via domain-general learning mechanisms applied to social interaction, but also stressed that cultural and linguistic practices mediate its transmission. Thus, mindreading is the outcome of a process of cultural evolution that was carried out over generations by our ancestors, and children would not acquire it if they did not encounter it in their social environment (Heyes 2014, 2018; Heyes and Frith 2014).

The defenders of the latter two accounts differ as to whether individual domain-general learning capacities alone are (child-as-scientist view) or are not (cultural learning view) sufficient for acquiring mindreading capacities. However, both of these accounts endorse a form of *rational constructivism* about mindreading. By ‘constructivism,’ they refer to the claim that mindreading is not an innate capacity but is acquired through domain-general learning applied to the child’s social experience—what may more appropriately be defined as an empiricist rather than a constructivist claim (see Allen and Bickhard 2011, 2013). By ‘rational,’ they intend the fact that other minds constitute a natural cognitive domain that is independent of our mindreading capacities and practices, and children (child-as-scientist view) or our species (cultural learning view) developed these capacities and practices by focusing on this domain. Thus, mindreading is acquired and/or has evolved following a rational process of discovery and theorizing (either individual or cultural), aimed at mastering a domain (mental states) that exists independently of it.

In contrast, others have argued for a form of constructivism that draws more extensively on socio-cultural considerations (Fenici 2017; Fenici and Garofoli 2017, 2019; Hutto 2008; Zawidzki 2013). Unlike individual rational empiricism, and like the cultural learning view, they claim that mindreading depends essentially on cultural practices of talking about mental states, and has emerged in the course of human cultural evolution. Unlike the cultural learning view, however, they also claim that the

domain mindreading aims to master is *constituted* by these practices: the very states that mindreading aims to track, e.g., propositional attitudes, do not exist independently of this culturally evolved capacity (Clark 1994; Kusch 1997).

Although this claim seems extreme, it is merely a consequence of taking seriously the analogy between mindreading and print reading urged by Heyes (Heyes and Frith 2014; Heyes 2018). Just as text—i.e., the domain of reading and writing—does not exist independently of this sophisticated linguistic practice, so too mental states like propositional attitudes—i.e., the domain of mindreading—do not exist independently of what is likewise, according to the radical form of socio-cultural constructivist, the sophisticated *linguistic* practice of attributing such mental states. Moreover, in the same way in which reading is interdependent with its regulative counterpart, writing, which constructs the objects (texts) that constitute the domain of reading, mindreading is interdependent with practices of *shaping* interpretive targets to respect rational norms that are constitutive of these mental states (Mameli 2001; McGeer 2007, 2015; Zawadzki 2008, 2013).¹

Thus, there are four approaches to mindreading that can be distinguished in terms of their implications concerning the origins of mindreading, and in terms of their assumptions concerning the relations between its descriptive and regulative dimensions. According to the first, the *nativist evolutionary psychological* approach, mindreading is the function of an innate module, naturally selected in human prehistory for accurately tracking conspecific mental states. Therefore, it aims primarily at describing facts about minds, and it may contribute to extrinsic, social, regulative tasks, e.g., using accurate knowledge of minds to turn people into better cooperators. According to the second, the *individualist empiricist* approach, mindreading is (re)discovered by individual infants during ontogeny, largely unassisted, via domain-general, science-like theorizing and hypothesis-testing. As with the nativist approach, it aims primarily to describe facts about minds, though it may contribute to extrinsic, social, regulative tasks. According to the third, the *cultural empiricist* approach, infants rely on cultural scaffolding to acquire culturally evolved mindreading capacities. This cultural scaffolding consists in using domain-general mechanisms to learn from caretakers and the ambient culture more broadly. Again, mindreading aims primarily to describe facts about minds, though succeeding at this task is made more tractable by regulative pressures on interpretive targets to conform to social expectations. Finally, according to the fourth, the *radical socio-cultural constructivist* approach, mindreading belongs to a suite of simultaneously interpretive and regulative practices that evolved through human prehistory, and continues to evolve to facilitate linguistically-mediated, large-scale cooperation. Mindreading and mindshaping are entirely co-dependent, like reading and writing; hence, the descriptive success of mindreading is entirely dependent on the regulative effects of mindshaping.

While it is relatively straightforward to construct origin stories for the first three options, this task appears considerably more challenging for the fourth. The reason is that, according to the first three paradigms, the domain of mindreading is independent of and pre-exists the practice of mindreading, and the aim of mindreading is primarily

¹ Mindshaping practices include such distinctively human phenomena as “overimitation”, explicit pedagogy, norm institution and enforcement, and narrative (self-) constitution.

to track this domain: our conspecifics have mental states that are worth tracking, and some mechanism, whether natural selection, individual learning, or cultural evolution plus cultural learning, gives rise to the capacity to track them. However, on the fourth paradigm, if mindreading depends on mindshaping in the way that print reading depends on print writing, then the domain of mindreading is *not* independent of and does *not* pre-exist the practice of mindreading. In this case, it is not at all obvious how to construct an origin story. Our main task in the following is to propose such an origin story, i.e., an origin story that begins with the assumption that mental states and practices of attributing them are interdependent in something like the way texts and reading/writing are interdependent.

One might question the value of this project. After all, is it not just *prima facie* implausible that mental states are constituted by our practices of attributing them? Such conventionalism makes sense for economic value or institutions like marriage but, surely, facts about what we think exist independently of our practices of attributing such facts. The alternative seems barely intelligible: doesn't attribution require pre-existing thought?

Our full response to this worry must await the conclusion of the paper, since it relies on the relative merits of our origin story over those of other paradigms. However, the following consideration may go some way to deflating initial skepticism about the value of our project. Until the cognitive sciences establish that mental states *as the folk conceive them* are integral to the best *scientific* theory of human psychology, it is at least an open question whether the mental states attributed in folk interpretive practice pre-exist folk interpretive practice. One needn't follow Churchland (1981) in endorsing the wholesale replacement of folk psychology by a completed neuroscience to doubt that folk psychology will be integral to a true theory of natural cognition. For example, even if brains enter mental states that, like beliefs, function to represent facts, it seems unlikely, given currently influential scientific models of the cognition, that they bear many of the properties the folk take beliefs to have. They may, for instance, admit of degrees of credence, rather than requiring "flat-out" endorsement or rejection, as with the folk notion of belief (Frankish 2004, 2009; Schwitzgebel 2001). Given the current lack of empirical support for much of folk psychology, considered as a scientific theory of natural cognition, it is a live possibility that the states attributed in folk interpretive practice do not exist *except as constituted by this practice*. Hence, an account of the origins of this practice that is compatible with this possibility constitutes a useful contribution to the discussion.

Our discussion proceeds as follows. We begin, in Sect. 2, with a more detailed presentation of Heyes' (2018) cultural empiricist account. We argue that it is at the same time both promising and limited, in that it opens the door to radical socio-cultural constructivism without fully exploring its possibilities. In Sect. 3, we begin our origin story of how a practice of talking about mental states may have originated without there already being an independent domain of mental states for it to describe by drawing on Robert Gordon's simulationist account of mindreading. In light of this account, we argue in Sect. 4 that mindreading functions *not* to track independently constituted, pre-existing mental states, but rather to make explicit practical commitments—understood in a specific way that we will clarify—in the context of linguistically-mediated coordination on cooperative projects. Section 5 then completes our origin story, pro-

viding a plausible account of how mental state terms and concepts could arise in human prehistory, given this understanding of their function.

2 Limitations of Cecilia Heyes' socio-cultural constructivism

The analogy between mindreading and print reading is central to Heyes' (2018) argument that mindreading is a "cognitive gadget" rather than a "cognitive instinct": a culturally transmitted and evolved skill rather than a biologically evolved, neuro-cognitive module. There are several reasons for this. First and foremost, despite emerging too recently to count as a product of biological evolution, print reading has many of the features that typically support arguments that cognitive capacities constitute biologically evolved, cognitive instincts: it correlates with specific, localized neural activation, and it admits of genetically determined deficits (dyslexia). This undermines arguments for the hypothesis that mindreading is a biologically evolved cognitive instinct: just because mindreading correlates with specific, localized neural activation, and admits of genetically determined deficits (autism spectrum disorder), it does not follow that it is a biologically evolved cognitive instinct. Like print reading, it may be a culturally evolved and transmitted cognitive gadget.

Much as we endorse this point, our focus in what follows is the second reason Heyes emphasizes the analogy between print reading and mindreading. In line with a number of recent proposals (Mameli 2001; McGeer 1996, 2007; Zawidzki 2008, 2013), Heyes notes that mindreading (like print reading) bears both interpretive *and regulative* relations to its domain. There can be no print reading without print writing. That is, learning how to master print requires learning both how to interpret texts, *and how to produce* texts, i.e., how to regulate one's behavior (writing) in ways that yield legible text. Similarly, argues Heyes (2018, p. 148; Heyes and Frith 2014), learning how to master explicit mindreading requires learning both how to interpret behavior, and how to produce or regulate behavior such that it is seamlessly interpretable in terms of the interpretive resources of the ambient culture. It is typically assumed that cognitive instincts are independent of such socially-mediated regulation; hence, the fact that the domain of mindreading, like text, is in important respects subject to regulative social conventions, supports Heyes' conclusion that mindreading is quite unlike a biologically evolved cognitive instinct, but is instead a product of cultural evolution.

We find Heyes' proposal that mindreading plays both interpretive *and regulative* roles with respect to human behavior intriguing and promising. For example, it has the potential to explain how mindreading comes to be *reliable*, despite the complex, many-to-many mapping between observable behavior and mental states. Although any finite bout of observable behavior is compatible with indefinitely many distinct sets of hidden mental states, we can come close to accurate interpretations because those we interpret are products of the same regulative practices as ourselves (Zawidzki 2008, 2013).

However, although Heyes recognizes this regulative dimension of mindreading, she resists the seeming implication that the domain of mindreading is socially constructed in the same sense that the domain of print reading is socially constructed:

The analogy between print reading and mind reading is not perfect. ‘Sound symbolism’ shows that the relations between inscriptions and their corresponding speech sounds and referents often depend on structural features of the nervous system ... but it is likely that these relations are more arbitrary than the relations between observable behavior and mental states. (Heyes and Frith 2014, p. 1243091-5)

Heyes appears concerned to avoid making mindreading too radically socially constructivist:

As a collectivist account, the cognitive gadget view has much in common with social constructivist accounts of mind reading ... However, unlike many social constructivists, and like theory-theory, the cognitive gadget view assumes that the processes involved in the development of mindreading are broadly rational and yield conceptual knowledge about the mind. (2018, p. 267, n. 1)

It is not immediately clear why Heyes thinks an overly radical social constructivist account of mindreading makes it less than broadly rational, and incapable of yielding conceptual knowledge about the mind. Consider thoroughly socially constructed conventions, like marriage and traffic regulations. There is a long tradition in philosophy and game theory of explaining how such phenomena can constitute rational solutions to coordination problems (Lewis 1969), and our knowledge of such conventions is clearly conceptual.

We think that Heyes must have a different worry; namely, that mindreading is unlike print reading because it must somehow track psychological facts about human beings that are *natural and non-conventional*. Accordingly, these facts—unlike fully conventional facts, like the economic value of currencies, cultural institutions like marriage, board games like *Pandemic* and chess, and coordination norms like traffic laws—are *non-arbitrary* and exist independently of our mindreading capacities and practices.

Making a comparison with a different cultural practice, we may say that, for Heyes, mindreading is analogous to hunting. A successful hunt consists in capturing prey, and cultures evolve tools and practices for maximizing such success. Thus, in hunting, success can be defined independently of the culturally evolved tools that enable it. Similarly, it may be argued that successful mindreading consists in the accurate attribution of mental states that exist independently of it, and cultures evolve tools and practices for maximizing such successes. Perhaps there is some role for regulation, e.g., to ensure that individuals conform more reliably to behavioral assumptions implicit in our mindreading practices; but this is a relatively marginal phenomenon. For the most part, these behavioral assumptions track natural *non-arbitrary* facts about human beings. In contrast, successful print reading cannot be defined independently of the practice of print writing. Accordingly, if the analogy between mindreading and print reading is strict and expresses more than a metaphor, as argued by the radical socio-cultural constructivist, successful mindreading cannot be defined independently of practices of shaping interpretive targets to be readable in terms of prevalent concepts of mind. In our view, Heyes attenuates the analogy between mindreading and

print reading, expressly because she does not endorse such conclusions about the interdependence between successful mindreading and such mindshaping.

Why prefer the latter view to Heyes' less radical alternative? We proposed one reason earlier, in the introduction: there is little evidence that mental states, as the folk conceive of them, will constitute integral parts of an empirically adequate, scientific psychology; hence, there is little reason to think that they exist except as constituted by our mindreading practices.²

A second reason concerns Heyes' claim that "the relations between observable behavior and mental states" are less arbitrary than "the relations between inscriptions and their corresponding speech sounds and referents" (Heyes and Frith 2014, p. 1243091-5). There are now a number of vibrant empirical research programs that appear to confute this assertion. For instance, Lisa Feldman Barrett's work on emotions (Barrett 2012, 2017; Gendron et al. 2014) suggests significant cultural variability in emotion attributions triggered by facial expressions. Gendron et al. (2014) show that "Western perceptions (e.g., scowls as anger) depend on cues to US emotion concepts embedded in experiments" (251). Comparing performance of US to Himba subjects on a task requiring them to sort facial expressions by expressed emotion type, this study found that, without cues to emotion concepts, only the US subjects showed the presumed "universal" pattern. When cues were included, both sets of subjects responded more closely to the presumed "universal" pattern, though considerable cultural variability remained. Similarly, cross-cultural studies on folk psychology show that, even if members of different cultures deploy the same folk psychological concepts, their assumptions about the relations between these concepts and observed behavior can differ markedly (Heelas and Lock 1981; Lillard 1998; Macdonald 2003; Vinden 1999).

This problem has long been recognized at a conceptual level by philosophers of mind. Nobody has ever successfully identified regularities linking mental states to behavior because of the holistic relations that obtain between them: how a cognitive agent responds upon experiencing one mental state depends on indefinitely many other mental states she might experience at the same time (Davidson 1970; Gauker 2002, p. 240; Morton 1996; Putnam 1980). Sadness does not invariably lead to tears, nor does anger invariably lead to a raised voice, because these states can co-occur with desires to conceal one's emotions, or paralyzing fear, etc. Propositional attitudes like belief or desire appear even less closely tied to behavioral responses: precisely what behaviors can we predict of all and only subjects who believe Paris is the capital of France, or who desire world peace (Dennett 1987, pp. 54–55)? Here there is bound to be dramatic variation among individuals within the same culture, let alone across cultures. Hence, it is far from apparent that the relations between mental states and

² Of course, as one reviewer notes, folk conceptions of mental states will likely play *some* role in scientific psychology, e.g., as ways of describing some forms of psychological functioning to be the target of mechanistic explanations. However, this does not necessarily vindicate the claim that mental states as the folk conceive them exist independently of folk practices of attributing such mental states. Instead, the forms of psychological functioning that are well described in terms of folk psychological concepts may be artefacts of folk practices of attributing mental states that fall under these concepts. Thanks to the regulative or "mindshaping" mechanisms and practices that accompany such attributions, they may turn into self-fulfilling prophecies, structuring psychological functioning in ways that respect the constraints that constitute them.

behavior are any less arbitrary than the relations between print and the sounds and concepts it expresses.

Thus, one of Heyes' principal reasons for attenuating her analogy between print reading and mindreading, such that the latter is less conventional and arbitrary, is not well-motivated. In fact, given the potential for massive individual variation here, this is precisely the sort of domain we would expect to be governed by socially constructed, regulative conventions. If the mental states attributed in folk interpretation are *constituted* by socially constructed, conventional relations to behavior, to which most are shaped to conform, then massive individual variation can be mitigated, and behavioral prediction made more tractable.³

But what of Heyes' concern that endorsing a radical socio-cultural constructivism compromises the rationality of mindreading? Above, we noted that there is a long tradition in philosophy and game theory of explicating the origins and rationality of even the most arbitrary, socially constructed conventions (Lewis 1969). However, such proposals cannot account for the rationality of conventions governing mindreading and mental states on pain of circularity, because they usually *presuppose* that mindreading and mental states are necessary to understand these conventions. Roughly, on such accounts, conventions are solutions to coordination problems that all parties *intend*, and *believe* of each other that each intends, and believe of each other that each believes that each intends, etc. Given that mindreading and mental states act as unexplained explainers in such accounts, they cannot account for the rationality of mindreading, if this essentially involves socio-culturally constructed regulative conventions that help constitute the mental states it attributes.

Fortunately, there are resources available for explaining the rationality, objectivity, and even relative non-arbitrariness of mindreading *other than* the claim that it aims to track natural, independently constituted cognitive states. Consider, for example, Robert Brandom's (Brandom 1994) theory of discursive practice. On his account of propositional attitudes, their structure derives from the task of "deontic scorekeeping", roughly, keeping track of the varying commitments of conversation partners relative to some topic of common interest.⁴ On Brandom's theory, it is a rational, objective, non-arbitrary fact that belief has the structure it has, *not* because folk interpretation aims to track natural cognitive states, but because it aims to make explicit the varying discursive commitments of different interlocutors relative to the same subject matter. Importantly, Brandom's appealing to commitments does not raise a risk of circularity, since "commitment" here is understood as a *deontic*, rather than a psychological

³ Incidentally, we note that this is not in tension with our previous claim that there is variability in folk psychologies. In order for a folk psychological framework to help mitigate massive individual variation in the behavior of individuals, it is not necessary for different cultures to employ the same framework, but only that members of each culture apply *some* common framework to other members.

⁴ Brandom argues that belief attributions, for instance, inevitably have what philosophers call a "de re – de dicto" ambiguity, because their function is to make explicit the varying commitments of different interlocutors concerning the same subject matter. When we say that Lois Lane believes that Clark Kent is not Superman, we are not attributing a flatly self-contradictory belief, because we mean that Lois Lane believes of Clark Kent (the *de re* component) that he is not the same person as Superman (the *de dicto* component). Roughly, the *de re* component captures the common referent about which we might disagree with Lois, while the *de dicto* component captures Lois's idiosyncratic perspective, or set of commitments about that referent.

category: it is constituted by the set of entitlements and obligations that governs some discursive performance, not by the performer's psychological dispositions. Hence, Brandom's account shows how it is possible to explain the rationality, objectivity, and partial non-arbitrariness of mindreading without presupposing that the mental states it attributes are independent of, or pre-exist it.

It may be objected that this response does not fully address the charge of circularity, but only pushes it back from the level of explaining norm-governed behavior to the level of explaining communication. Following a Gricean picture (Grice 1957, 1969), it is indeed common in cognitive science to assume that human communication requires that listeners must understand the communicative intentions of the speaker, i.e., her mental states (Bloom 2000; Scott-Phillips 2014; Sperber and Wilson 1995, 2002). It would follow, then, that explaining the rationality of mindreading by appealing to a theory of discursive commitments such as Brandom's would be a non-starter for socio-cultural constructivism, as it would still presuppose mindreading at its core. Interestingly, however, critics of this version of the Gricean picture of communication have argued that Grice's original framework is far too intellectualized, as it presupposes cognitive capacities such as higher-order meta-representation that are largely unnecessary (Gómez 1994; Moore 2017, 2018). If these critics are correct, then Gricean communication is not as cognitively demanding as it is often supposed, and a framework such as Brandom's is well suited to explain the rationality of mental state discourse without presupposing sophisticated mindreading in the interpretation of language.

In short, Heyes' motivations for attenuating the analogy between mindreading and print reading do not rule out a more radical social constructivism that takes this analogy strictly. In what follows, we argue that it is possible to construct a detailed and plausible origin story for such socially constructed mindreading, showing how it could have culturally co-evolved gradually with the domain it describes, i.e., mental states, like beliefs and desires, attributed in folk interpretation.

3 Robert Gordon's distinctive brand of simulation theory

We want to introduce our proposal that mindreading originated as a social and linguistic practice in the course of cultural evolution, by exploring in depth one under-appreciated intuition underlying Robert Gordon's simulationist account of mindreading (1986, see also 1995, 1996, 2000, 2007). According to the "simulation theory," we attribute mental states by pretending to be in an interpretive target's situation, deciding what to do if we were there, and then applying this decision to the interpretive target. In acknowledging Gordon as one of the pioneers of the simulation theory, the mainstream literature on mindreading (Carruthers and Smith 1996; Davies and Stone 1995) has tended to conflate his view with that of other simulationist pioneers, like Goldman (Goldman 2006; Gallese and Goldman 1998). This is a mistake, since Gordon's simulation theory is a far more radical departure from traditional assumptions about mindreading than is often appreciated, as will become clear after a detailed review of his proposal.

Against the view that mindreading is a theoretical capacity to reason about mental states, Gordon (1986) observes that, in the case of one's own behavior, one does not

usually predict what one is immediately going to do (or, similarly, what one would likely do in imagined or hypothetical circumstances) by attributing mental states to oneself. On the contrary, the “declarations of immediate intention [or of intentions in hypothetical circumstances] ... are often products of *practical* reasoning: reasoning that provides the basis for a decision to *do* something” (Gordon 1986, p. 160, here and in the following quotations the emphasis is in the original).

Elaborating on this proposal, Gordon suggests that the ability to declare an intention on behalf of another agent—i.e., what, on standard views of mindreading, actually constitutes the capacity to attribute an intention—may be nothing more than another product of this practical reasoning capacity. In particular, “in one type of hypothetical *self*-prediction the hypothetical situation is one that some *other* person has actually been in, or at least is described as having been in. The task is to answer the question, ‘What would *I* do in *that* person’s situation?’” (Gordon 1986, p. 161). Thus, Gordon argues, what in everyday life are usually taken as cases of intention attribution may actually involve a capacity for mental simulation—i.e., imagining what one would plan to do if one were in the other’s situation.

Importantly, Gordon suggests, mental simulation can also explain how we might attribute beliefs. Indeed, he proposes that declaring one’s beliefs—i.e., what, in the mindreading debate, actually constitutes the capacity to attribute a belief to oneself—does not seem to depend on some ability to introspect one’s mental states but rather on a capacity to answer particular questions about what is true in one’s particular situation: “I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*” (Evans 1982; quoted in Gordon 2007).

Switching again from the case of a declaration of one’s own mental state to the ascription of the same mental state to another, Gordon argues that we can similarly conceive the attribution of a belief to another agent as a particular way to answer questions, or make related statements: “to attribute a belief to another person is to make an assertion, to state something as a fact, *within the context of practical simulation*. Acquisition of the capacity to attribute beliefs is acquisition of the capacity to make assertions in such a context” (Gordon 1986, p. 168). Thus, Gordon proposes that what we actually do when we attribute a belief to an agent really is again a particular case of mental simulation—i.e., we imagine ourselves in the other’s situation, and then try to make a statement from that particular perspective.

Gordon’s proposal is unique in advancing a promising intuition we want to explore. As it should be clear by now, it is indeed motivated by what, at its core, is a *semantic* intuition about how we evaluate the truth of verbal attributions of mental states. According to this intuition, evaluating the truth of such attributions does *not* require focusing on anything internal to people’s minds/brains. Hence, the verbal practice of ascribing mental states does not say much about the alleged internal psychological reality of an agent. The linguistic form of propositional attitude attributions does not convey any deep truths about the cognitive processes that we put in use when we engage in what is usually described as mindreading.

This is the sense in which Gordon’s simulationist analysis of the ascription of mental states may be taken to promote a deflationist view about the psychological reality of mindreading. In contrast with other simulation theorists, like Goldman, Gordon does

not share the assumption that the products of simulations are psychological descriptions of the interpretive target's mental states. Instead, he defends the deflationist view that words like “believes”, in sentences of the form “S believes that P”, are merely linguistic devices which receive their meaning from the attributer's capacity to activate her decision-making processes from a displaced perspective, *rather than* terms referring to natural, interpretation-independent mental states. Consequently, the expression ‘mindreading’ does not denote a cognitive mechanism; rather, it is no more than a commonsense label that *we*—i.e., philosophers and cognitive scientists in the last thirty years—use when describing typical, quotidian behavior interpretation.

At the same time, Gordon's ontological view—as far as one can extrapolate it from his brief remarks—is in many respects in line with mainstream assumptions. While he is mainly interested in the *ascription* of mental states, he never denies that mental states may be natural kinds and constitute the usual causes of intelligent behavior, and he certainly rejects radical socio-cultural constructivism. For instance, while discussing how children may learn to report their desires, he clearly distinguishes their *verbal expression* of a desire from their *psychological state* of desire—although he does not clarify what the latter is.⁵ Because he seems to assume that mental states are physical properties and dispositions with a causal role in producing behavior, Gordon also maintains that our ascriptions of mental states describe something real about causal and nomological relations. Accordingly, he does not deny the predictive function of mindreading, nor does he suggest that we should avoid attributing mental states in our folk psychological explanations.⁶

For the sake of our discussion, however, what is interesting to us is that Gordon's account of the capacity to ascribe mental states in language is completely independent of this traditional (and underspecified) ontological view on the nature of mental states. Consequently, while the causal/realist ontology of *mental states* endorsed by Gordon may be incompatible with a strict analogy between mindreading and text reading, since it rules out constitutive relations between mental states and their ascription, this incompatibility does not extend to Gordon's semantic analysis of *mental state ascription*, which is fully compatible with a radically, socio-culturally constructivist one.

As Gordon—building on Evans—clarifies, what is essential to his proposal is that, in ascribing mental states such as beliefs, “people optionally step up a semantic level from an assertion that *p* to a self-ascription of a belief that *p*.” (Gordon 2007). Accordingly, when we make a statement while simulating another agent, we attempt to answer a question “*that is not about oneself, nor about mental states at all*: [it is] an outward-looking question” (Gordon 2000, p. 111). Because of this, mindreading forces us to focus *on the external world* “with its emotional and motivational qualities, its affordances, and its various modal properties” (Gordon 2000, p. 107).

⁵ “Young children ... [utter ‘I] want ϕ ,’ in order to] request or demand a ϕ only when they actually want a ϕ ” (Gordon 1995, p. 358).

⁶ “I do not deny that explanations are often couched in terms of *beliefs*, *desires*, and other propositional attitudes; or that predictions, particularly predictions of the behavior of others, are often made on the basis of attributions of such states. Moreover, as functionalist accounts of folk psychology have emphasized, common discourse about beliefs and other mental states presupposes that they enter into a multitude of causal and nomological relations. I don't want to deny this either.” (Gordon 1986, p. 165).

It follows that, according to Gordon, one needs no *prior*, conceptual capacity to track independently constituted mental states in order to master their verbal ascription in conversation. All that is necessary is that one can grasp the relations (specific for every kind of mental state) that must hold between an agent and her environment in order for the agent to count as instantiating the relevant mental states.⁷ And in fact, Gordon argues that this is how toddlers and young children start learning the vocabulary of folk psychology. Rather than focusing on their own or other agents' mental state states, they start manipulating the verbal ascription of verbal states by mastering the external conditions in which they can be employed correctly. Only later do they realize the private, psychological nature of such uses (Gordon 1995, 2007; also see Carpendale et al. 2009; Fenici subm. for a more detailed suggestion about how children might assemble their understanding of mental states starting from their limited mastery of the pragmatics of mental expressions).

This perspective makes Gordon's view on the ascription of mental states a natural ally of radical socio-cultural constructivism, since, according to it, the mental states attributed in mindreading are conventional, in the sense that they are not independent of such attributions. In particular, it makes possible a new origin story about mindreading: just as toddlers and young children may master the verbal ascription of mental states while having no understanding of them as private psychological states, the lexicon of folk psychology may have similarly evolved in the history of our species to track other kinds of situations, and acquired its mentalistic significance only later.

We elaborate on this suggestion in the next two sections. The mainstream view appears to have an unassailable advantage: if mental state ascriptions describe inner states causally responsible for the behavior of a cognitive agent, it is clear why attributing them supports the interpretation, prediction, and justification of behavior. However, on the extension of Gordon's analysis of mental state ascription explored here, according to which it begins with normative links between an agent's behavior and her environment, one may wonder why we should ascribe mental states *at all* rather than talking directly about the behavior of the agent, and the environment in which it occurs. In Sect. 4, we argue that the solution to this problem is to appreciate how mindreading functions to express practical commitments—understood in a specific way that we will clarify—in the context of collective projects requiring complex coordination. In Sect. 5, we focus on the cultural evolution of the verbal and linguistic practices that we use to ascribe mental states in everyday social interaction to show how mental state concepts could have acquired the pragmatic functions we discuss.

⁷ These relations are normative ones, i.e., constituted by the norms of practical rationality, and hence mastering them is not necessarily the same as characterizing an agent's inner psychological reality. However, due to the regulative or "mindshaping" roles ascribing such normative relations play, they typically come to characterize agents' inner psychological realities, to the extent that such regulative and mindshaping regimes are effective. This is precisely how a radically socio-culturally constructivist ontology of mental states differs from the alternatives: as with text reading, mindreading tracks facts that are *artefacts of the regulative cultural practices with which it is interdependent*.

4 Ascribing mental states as a way of assuming and sharing practical commitments

We now elaborate on the intuitions motivating Gordon's analysis in the previous section, to clarify what the social function of mindreading may be if not that of describing inner states causally responsible for the behavior of a cognitive agent. We will get to our conclusion through presenting a series of increasingly complex examples. Let us start from a simple situation. Suppose that Andrew is planning to have dinner at an Indian restaurant. He may express this plan by declaring "I intend to go to the Indian restaurant," and hence by self-ascribing the intention to have dinner at a specific Indian restaurant. Following Gordon's analysis, this self-ascription does not *describe* a psychological entity (i.e., an intention) in Andrew's mind, but simply *expresses* the conclusion of a bout of practical reasoning in which has engaged: the fact that he is acting in order to have dinner at the Indian restaurant.

If all ascriptions of intentions were of this kind, it would raise the difficult question of why we ever started ascribing mental states in language at all. The intention ascription in the example seems redundant because it simply describes what Andrew is going to realize as a result of practical reasoning, which would exist and have the usual behavioral effects whether or not he expressed it.⁸ Why should Andrew declare "I intend to go to the Indian restaurant" rather than simply "I will go to the Indian restaurant," then?

We answer the question by showing with two examples the pragmatic function that the verbal ascription of mental states plays in supporting social interaction. Suppose that Barbara and Andrew are now trying to determine where to go out for dinner together. In this context, if Andrew declared "I will go to the Indian restaurant," Barbara would have a reason to be upset by his bossy stance in imposing his preference for Indian food. In contrast, his declaring "I intend to go to the Indian restaurant" expresses a practical commitment to go to a particular place while still being open to negotiation. That is, the declaration of an intention in the process of coordinating plans and sharing decisions functions to allow expressing one's commitment to a plan in order to share it in the attempt to secure another's endorsement for it.⁹

We can similarly analyze the practice of verbally ascribing beliefs to clarify its distinctive role in the activity of sharing and coordinating plans. Suppose Barbara and Andrew's attempt to jointly decide where to go out together for dinner faces an unexpected obstacle: the restaurant Andrew expressed the intention to visit may be closed without him being informed of it, while Barbara has acquired evidence of this. In this context: (i) Andrew's plan cannot be accomplished; (ii) the fact that the Indian restaurant is closed constitutes a good reason for Andrew to withdraw his proposal to

⁸ We may also call what Andrew is going to realize the *goal* of his action. Importantly, however, goals so conceived identify the state of affairs that must hold in order for the practical reasoning to be successful; they do *not* describe psychological states (see, e.g., our discussion in Fenici and Zawidzki (2016), as well as the discussion of Geurts below). As Canfield (2007, p. 50) argues: "We often act with an aim. But to speak of someone's aim is not to make a psychological claim".

⁹ As we argued in Sect. 2 (and we will repeat below), we use the word 'commitment' to indicate the set of practical and discursive entitlements and obligations that follow from one's playing a role in a social practice. In this sense, 'commitment' is a deontic category, which is independent of the psychological dispositions of the performer.

go to the Indian restaurant, and (iii) reporting “The Indian restaurant is closed” would be an acceptable conversational move for Barbara to make, in order to rule out the Indian option, and direct the conversation to a viable alternative.

Importantly, by declaring “The Indian restaurant is closed,” Barbara would incur a kind of obligation regarding the fact that the Indian restaurant is closed: more precisely, a commitment that the proposition expressed by “The Indian restaurant is closed” is true. For example, she would expose herself to sanction if it were later discovered that the Indian restaurant is open. Nevertheless, she may reasonably attempt to modify the joint commitment to go the Indian restaurant by stating, “I believe that the Indian restaurant is closed,” which expresses both (i) her commitment to rule out the Indian option under the assumption the restaurant is closed together with (ii) her request that Andrew explicitly agree to share this commitment. This linguistic device enables her to attempt to alter the joint plan with Andrew in the same way as asserting that the Indian restaurant is closed, *without incurring the same doxastic obligations*: she could not be blamed if it is discovered the restaurant is open, for expressing a mere belief rather than asserting the belief’s content presents this fact as a mere possibility that is compatible with Andrew’s original plan. This is because, compared with a pure assertion, in expressing a belief, a speaker makes a statement about the world while not fully committing to it. The declaration of a belief implies some “hedging” on the part of the speaker with respect to the content of the expressed belief (Simons 2007; Van Cleave and Gauker 2010).

So far, we have shown that the verbal practice of ascribing intentions and beliefs helps make explicit an agent’s practical commitments, and thus plays an important role in the context of creating, sharing, and negotiating joint plans and goals. This expands Gordon’s suggestion that the verbal practice of ascribing mental states fundamentally aims to characterize the relations between an agent and her environment. The practical commitments that an agent makes explicit while creating, sharing, and negotiating joint plans indeed express how an agent is inclined to act, and are thus characterized by how the agent is oriented toward the environment.

Nevertheless, merely showing that (self-)ascribed mental states manifest an endorsed commitment to act does not yet demonstrate that the verbal practice of ascribing mental states also constitutes the domain of folk psychology. On the contrary, according to a long tradition in philosophy, it is exactly the fact that mental state attributions identify inner states causally responsible for the behavior of a cognitive agent, which are independent from, and more fundamental than the practice of ascribing them, that allows them to characterize the behavioral dispositions of an agent, and thus to be used in the interpretation, prediction, and explanation of behavior (Davidson 1963).

Furthermore, a broader tradition not only in philosophy but also in game theory and cognitive science would expand on the previous view to claim that the whole context of shared planning—that is, the framework in which we developed our analysis of the social/pragmatic function of ascribing mental states—*presupposes* mindreading. Sharing plans is indeed a typical case of coordination, and the coordination of behavior poses a typical problem for cognitive agents: *prima facie*, it seems impossible to coordinate with another agent unless one already understands the (mental) goal the other agent is trying to realize. An agent’s capacity to reason about conventions and

common knowledge is the typically proposed solution: all parties *intend* a particular joint behavior, and *believe* of each other that each intends it, and believe of each other that each believes that each intends it, etc. (Lewis 1969).

Even more problematically for our view, creating, sharing, and negotiating plans involves verbal communication, and this is a specific form of coordination, which is traditionally assumed to require explaining how a speaker can transfer a meaning—information in her mind—to the listener by alignment/convergence on the use of the same meanings. The traditional solution posits that the participants in verbal communication continuously assume a mentalistic understanding of each other, in order to understand each other's communicative intentions (Grice 1975). Therefore, if, following these views, mindreading is proposed as a fundamental cognitive capacity that allows disentangling the problems that arise in coordination and communication, it cannot be explained in terms of socio-culturally constructed regulative conventions constituting the domain of mental states, as we suggest.

Fortunately, there are a number of explanatory tools that allow us to avoid this potential circularity at the heart of our account. First, evolutionary game theory provides an extremely rich set of formal resources to model the emergence of social conventions in the absence of sophisticated mindreading or even mental states (Skyrms 1996, 2004, 2010). According to these models, under certain parameters, evolutionarily stable strategies of coordination emerge among populations of interacting agents, even if these agents have no knowledge of each other's mental states.¹⁰ The same sorts of formal models can explain everything from the evolution of the social contract, to solutions to formal coordination games, like the “stag hunt”, to signaling conventions arising in natural populations of organisms.

Second, one need not even abandon the Gricean project, broadly construed, to explain human communication without presupposing sophisticated mindreading. For example, Richard Moore (2017, 2018) argues persuasively that functionally Gricean forms of communication are possible among agents who are incapable of attributing beliefs, or hierarchically nested mental states. This is compatible with Bart Geurts's (2019, p. 3) recent suggestion that “the chief purpose of speech acts is to enable speakers to share commitments that enable them to coordinate their actions: communication is coordinated action for action coordination.” Similarly to our proposal, he defines a commitment not as a mentally represented agent goal,¹¹ but as a three-place relation, $C_{a,b} p$, between two individuals, a and b , and a propositional content, p , according to which “to say that a is committed to b to act on p is to say that a is committed to b to act in a way that is consistent with the truth of p .”

Significantly, for the sake of the present discussion, it is possible, following Geurt's proposal, to consider communication as a way of sharing commitments in the context of action coordination, without assuming that communication requires sophisticated

¹⁰ In particular, Skyrms focuses on non-psychological facts like the network structure of populations, which make it more likely that individuals with complementary coordination strategies interact, leading to population-level equilibria where such strategies are stable against invasion by incompatible strategies. In such situations, there can be successful coordination without mindreading. Skyrms (2010) contains many real-world examples of biological signaling conventions that succumb to this form of analysis.

¹¹ “It is sometimes supposed that goals are psychological entities, but that is not how I understand the term” (Geurts 2019, footnote 4).

mindreading. Indeed, Geurt's definition of commitment implies, first, that commitments are defined in relation to states of affairs—i.e., situations *in the world* rather than *in the agent's head*—that make the propositional content of a commitment true, and constrain the possible behaviors of an agent, by prescribing that she acts in accordance with the truth of that content. Thus, saying that an agent is committed to p implies something about how the agent is supposed *to act*, not about what the agent is *intending* or *thinking*. Second, Geurt's definition conceives of commitments in terms of *normative relations* among individuals (similar to obligations, duties, and responsibilities), originating in the mutual expectations that the agents living within a community hold of one another. Thus, the necessity governing an agent who undertakes a commitment is neither physical nor causal, but a form of a social obligation.¹²

Thus, although we do not have the space here to work through such proposals in detail, it is clear that there are rich theoretical frameworks available for making sense of the structure of mental state attribution, and the roles it plays in coordination and communication, without assuming that these properties derive from the function of tracking independently constituted, pre-existing mental states. Coordination (Skyrms 1996, 2004, 2010) and communication (Moore 2017, 2018; see also Breheny 2006; Taylor 2012) can arise in populations lacking sophisticated concepts of mental states, and mental state attribution can be understood in terms of the tracking of normative commitments that arise in the context of such forms of coordination and communication (Geurts 2019; see also Brandom 1994).

We now turn to a phylogenetically plausible origin story that applies these notions: we discuss the kinds of coordinative challenges that likely faced our prehistoric ancestors, illustrate the useful role that commitment tracking could have played in such contexts, and explain how such commitment tracking could easily have gradually transformed into a form of *mental* state attribution.

5 The origin of mental state concepts

We are not the first philosophers to engage in what Sellars calls “anthropological science fiction” (Sellars 1956) in order to make conceptual points about the origins of our interpretive practices, including mental state attribution. Sellars famously constructs his “myth of Jones” to show how a population of humans who already spoke complex language might come to develop concepts of propositional attitudes that are in some sense derived from linguistic categories, like assertions. Gordon (2000) constructs a similar hypothetical population of “outlookers” to make the point that we can explain each other's behavior entirely in terms of external states of the world rather than states

¹² Interestingly, Geurts also argues that it is possible to explain the verbal ascription of mental states as particular commitments that one can assume (either with herself as private commitments or with others as social commitments). More precisely, a (self-)ascribed intention can be defined as a commitment $I_{a,b} p$ that is (i) private (if $a = b$) or social (if $a \neq b$), and (ii) *telic*, i.e., p refers to a state of affairs that a aims to realize. Similarly, a (self-)ascribed belief can be defined as a commitment $B_{a,b} p$ that is (i) private (if $a = b$) or social (if $a \neq b$), and (ii) *atelic*, i.e., the truth of p significantly constraints a 's actions without she aims to realize it.

of mind. Below, we construct a speculative, historical progression that is firmly rooted in this tradition.

Like Sellars, we begin with a population that already has sophisticated coordinative and communicative practices, and like Gordon, we think of the initial stages of this history of human interpretation as relying on appeals to states of the world, rather than states of mind, to explain behavior. However, our account extends these others in three important respects. First, it is constrained by plausible, empirical models of human prehistory. Specifically, we take as our starting point Sterelny's (2012) idea that *H. sapiens*' distinctive evolutionary arc is characterized and explained by sophisticated practices of inter-generational transfer of information (primarily imitation and master-apprentice-style pedagogy), relevant to complex forms of group foraging, like megafaunal hunts. Second, taking lessons from Sect. 4 above, we assume that the coordination and communication required for such complex forms of teamwork involved something like the capacity for *commitment* to linguistically articulated claims. And finally, as defended above, we assume that there are formal tools for explaining the emergence of sophisticated coordination and communication that do not presuppose mindreading, against mindreading-first accounts of coordination (Lewisian accounts) and communication (overly intellectualized Gricean accounts).¹³

Our goal is to show that Gordon's deflationary perspective on mindreading, and the related idea that mindreading derives from linguistic practices of expressing practical commitments, can be used to construct a plausible origin story for our mental state concepts. We begin with Sterelny's (2012) idea that complex coordination in group hunts of megafauna is the crucible in which the human socio-cognitive "syndrome" was forged. Appealing to persuasive paleontological evidence, Sterelny argues that the first hominins were distinguished from other hominids in successfully hunting megafauna with relatively rudimentary weapons. The only way this was possible was via complex coordination—a level of teamwork and group planning unparalleled in the natural world. And such techniques of coordination were possible only as a result of traditions of social learning enabling cumulative cultural evolution, where later generations could tweak the techniques of earlier generations into more effective forms, without "reinventing the wheel".

As Sterelny notes, this socio-cognitive syndrome would require subtle and complex forms of communication, to manage complex coordination in group hunts, as well as forms of pedagogy that transferred these techniques to new generations. To plan and engage in such complex joint activities, there had to be a way of publicly articulating goals, and of publicly sharing information relevant to these goals: something

¹³ In addition, it is no part of our claim that human coordination and communication do *not* require sophisticated, perhaps species-specific cognitive capacities. In particular, we are persuaded that our natural abilities to track actions, their goals, and the information that rationally constrains them, precede and make possible our coordinative and communicative feats, and are in many respects unique among animals. Human beings are also likely unique in various capacities to learn from each other, and these also make human-style coordination and communication possible. However, this is *not* equivalent to the claim that mindreading, *in terms of the concepts of folk psychology*, precedes and makes possible human-style coordination and communication. There is little reason to think that such folk concepts can be used to accurately characterize the complex, natural, socio-cognitive states that make human coordination and communication possible (see Fenici 2015a; Fenici and Zawidzki 2016, and the discussion in Fenici 2015b for our account of the nature and the ontogenetic development of these abilities).

analogous to commands and assertions, respectively. And to transfer these techniques to new generations of novices, such communicative acts had to be characterized by *differentials in authority*: the utterances of experts, both command-like and assertion-like, had to carry a kind of weight in the eyes of novices that their own utterances did not.

Thus, consider a population of human ancestors that routinely engage in complex joint activities, made possible through communicative acts conveying what each is to do (goals expressed in commands), together with relevant information (states of the world expressed in assertions). The commands and assertions of a select minority, i.e., the experts, are, in some sense, self-justifying: they carry a kind of authority in virtue of the fact that they are made by experts. These experts are typically leaders of such complex joint activities. For example, we can imagine a senior member of a hunting party assert that the mammoth's pace has increased (based on some sign that goes unnoticed by novice members), and command the other members of the hunting party to likewise pick up the pace. Novices would immediately comply. But novices do not stay novices forever. They eventually become experts themselves. And it is plausible to suppose, with the Vygotskian tradition, that this process involves *internalizing* initially social forms of regulation, i.e., the commands and assertions of experts, so that the novice gains a form of endogenous self-regulation derivative in form and efficacy from these social acts of communication (Vygotsky 1934, 1978; Wertsch and Stone 1985; Martínez-Manrique and Vicente 2015; Fernández Castro 2016). Such internalization, on the Vygotskian view, is how pedagogy works: novices learn what to do in different situations, and what to attend to in order to categorize their situations, by internally rehearsing the verbal commands and assertions of their teachers.

Now consider how such late-stage novices might report on their own behavior, or describe and predict the behavior of others. They would appeal to the kinds of expert commands and assertions that, from their earliest memories, *justified* certain behavioral decisions (because they were made by experts), which they had now internalized to enable endogenous self-regulation. It is quite natural to think of such reports and descriptions as providing *practical reasons* for engaging in certain acts: behavior is explained in terms of practical commitments to pursuing certain goals (articulated as internalized expert commands) in the light of commitments about the way the world is (articulated as internalized expert assertions). But here we are already in the realm of folk psychology: these patterns of report, description, and explanation in terms of practical commitments that justify behavior are exactly the kinds of practices that Gordon argues philosophers and cognitive scientists have come to call “mindreading”, or “folk psychology”, or “mental state attribution”. And, as in Gordon's and our own conception, there is no sense in which these practices aim at tracking the psychological states of interpretive targets. Rather, what they track are practical reasons: commitments to goals and concerning facts that can justify behavior.

Considerations of space prevent us from filling in the details of this sketchy account. However, we think the broad outlines are clear and persuasive. Given the distinctive socio-ecology that there is good reason to think gave rise to the human socio-cognitive syndrome, i.e., coordination on complex joint foraging like group hunts of megafauna, it is plausible that our ancestors relied on capacities to publicly share goals and informa-

tion about states of the world, together with deference to experts at such coordination. These ingredients would be enough, we claim, to give rise to an interpretive practice not unlike Gordon's conception of mindreading: behavior would be reported, described, explained, and regulated in terms of the kinds of linguistically articulated commitments to goals and concerning facts that novices internalized from communicative interactions with experts.¹⁴

6 Conclusion

Drawing on an anthropological thought experiment, constrained by reasonable conjectures about human prehistory, we have argued that mindreading, i.e., the ascription of mental states, may have originated in order to track the practical commitments cognitive agents assume when creating, sharing, and negotiating shared projects and goals, rather than to describe causal entities in their brains/minds.¹⁵ The narrative we told in Sect. 5 supports the analysis of the pragmatic function of the verbal ascription of intentions and beliefs we introduced in Sect. 4, and shows how these cultural and linguistic practices may be grounded inter-subjectively even without presupposing more fundamental, meta-representational capacities. Our analysis has thus vindicated the important intuitions that we adopted from Gordon, in Sect. 3: (1) the expression "mindreading" actually identifies a way of describing kinds of social interaction, but does not denote a genuine cognitive mechanism, and (2) ascribed mental states ultimately relate cognitive agents to external states of the world rather than tracking their inner psychological reality.

Accordingly, our analysis provides a cultural-evolutionary origin story for a socio-cultural constructivist model of mindreading—as a verbal practice of ascribing mental states—that is an alternative to Heyes' account. Furthermore, it shows how mindreading and its domain, i.e., the kinds of mental states ascribed in folk interpretation, like beliefs and desires, may have gradually, culturally co-evolved. This solves the puzzle raised by radical socio-cultural constructivism, or any view that takes Heyes' analogy between mindreading and print reading strictly. We needn't presume that mindreading aims to describe an independently constituted, natural domain of psychological facts in order to explain how it originates and culturally evolves through human prehistory.

We cannot directly compare our origin story to Heyes', since she does not explicitly provide an account of the origins and cultural evolution of mindreading from a cultural empiricist perspective. However, we doubt that any such account could accord anything but a marginal role to culture. Perhaps storytelling and other forms of pedagogy can explain how a culturally empiricist folk psychology is preserved and passed down

¹⁴ Some may notice that our proposal strikingly resonates with that of Pettit (2018). We greatly thank an anonymous reviewer for indicating this work to us, and remark that we reached our conclusion independently of it.

¹⁵ As we noted above, due to the regulative roles of such commitments, such assumptions often end up impacting the psychologies of such agents. Our point, however, is that the use of mental state ascriptions in order to track such constructed psychologies *derives from* their prior, commitment-expressing uses, and this shows how mental states and their ascription can be interdependent, in the way that reading and constructing texts is, i.e., it constitutes a radically socio-culturally constructivist account of the origins of mindreading and mental states.

to subsequent generations, but the origins of folk psychological concepts, for Heyes, would presumably differ little from what an individualist empiricist like Wellman would hypothesize: individual, science-like hypothesis generation and test seems to be the only mechanism either account can provide for explaining the *origins* of folk psychological concepts. In contrast, on our account, the origins of folk psychological concepts are socio-cultural through and through, because these concepts originate in a socio-linguistic practice of commitment sharing in the context of coordination on joint projects.

References

- Allen, J. W. P., & Bickhard, M. H. (2011). Emergent constructivism. *Child Development Perspectives*, 5(3), 164–165.
- Allen, J. W. P., & Bickhard, M. H. (2013). Stepping off the pendulum: Why only an action-based approach can transcend the nativist–empiricist debate. *Cognitive Development*, 28(2), 96–133. <https://doi.org/10.1016/j.cogdev.2013.01.002>.
- Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, 67, 159–186. <https://doi.org/10.1146/annurev-psych-010213-115033>.
- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14(3), 110–118.
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: The MIT Press.
- Barrett, L. F. (2012). Emotions are real. *Emotion*, 12(3), 413–429. <https://doi.org/10.1037/a0027555>.
- Barrett, L. F. (2017). *How emotions are made: The secret live of the brain*. New York: Houghton Mifflin Harcourt.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: The MIT Press.
- Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Cambridge: Harvard University Press.
- Breheny, R. (2006). Communication and folk psychology. *Mind and Language*, 21(1), 74–107. <https://doi.org/10.1111/j.1468-0017.2006.00307.x>.
- Canfield, J. V. (2007). *Becoming human. The development of language, self and consciousness*. Basingstoke: Palgrave Macmillan.
- Carpendale, J. I. M., Lewis, C., Susswein, N., & Lunn, J. (2009). Talking and thinking. The role of speech in social understanding. In A. Winsler, C. Fernyhough, & I. Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-regulation* (pp. 83–94). Cambridge: Cambridge University Press.
- Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78(2), 67–90.
- Clark, A. (1994). Beliefs and desires incorporated. *Journal of Philosophy*, 91(8), 404–425.
- Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). Oxford: Oxford University Press.
- Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60(23), 685–700.
- Davidson, D. (1970). Mental events. In L. Foster & J. W. Swanson (Eds.), *Essays on actions and events* (pp. 207–224). Oxford: Clarendon Press.
- Davies, M. K., & Stone, T. (Eds.). (1995). *Folk psychology: The theory of mind debate* (1st ed.). Oxford: Wiley-Blackwell.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.
- Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5), 178–190. [https://doi.org/10.1002\(SICI\)1520-6505\(1998\)6:5%3c178::AID-EVAN5%3e3.0.CO;2-8](https://doi.org/10.1002(SICI)1520-6505(1998)6:5%3c178::AID-EVAN5%3e3.0.CO;2-8).
- Evans, G. (1982). *The varieties of reference*. New York: Oxford University Press.

- Fenici, M. (subm.). *How children approach the false belief test: Social development, pragmatics, and the assembly of Theory of Mind*.
- Fenici, M. (2015a). A simple explanation of apparent early mindreading: Infants' sensitivity to goals and gaze direction. *Phenomenology and the Cognitive Sciences*, 14(3), 497–515. <https://doi.org/10.1007/s11097-014-9345-3>.
- Fenici, M. (2015b). Social cognitive abilities in infancy: Is mindreading the best explanation? *Philosophical Psychology*, 28(3), 387–411. <https://doi.org/10.1080/09515089.2013.865096>.
- Fenici, M. (2017). What is the role of experience in children's success in the false belief test: Maturation, facilitation, attunement or induction? *Mind and Language*, 32(3), 308–337. <https://doi.org/10.1111/mila.12145>.
- Fenici, M., & Garofoli, D. (2017). The biocultural emergence of mindreading: Integrating cognitive archaeology and human development. *Journal of Cultural Cognitive Science*, 1(2), 89–117.
- Fenici, M., & Garofoli, D. (2019). Cultural evolutionary psychology is still evolutionary psychology. *Behavioral and Brain Sciences*, 42, e176.
- Fenici, M., & Zawidzki, T. W. (2016). Action understanding in infancy: Do infant interpreters attribute enduring mental states or track relational properties of transient bouts of behavior? *Studia Philosophica Estonica*, 9(2), 237–257.
- Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.
- Frankish, K. (2009). Partial belief and flat-out belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 75–93). Netherlands: Springer. https://doi.org/10.1007/978-1-4020-9198-8_4.
- Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12), 493–501.
- Gauker, C. (2002). *Words without meaning (The MIT Press)*. Cambridge, MA: Christopher Gauker.
- Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion (Washington, D.C.)*, 14(2), 251–262. <https://doi.org/10.1037/a0036052>.
- Geurts, B. (2019). Communication as commitment sharing: Speech acts, implicatures, common ground. *Theoretical Linguistics*, 45(1–2), 1–30.
- Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York: Oxford University Press.
- Gómez, J. C. (1994). Mutual awareness in primate communication: A Gricean approach. In S. T. Parker, R. W. Mitchell, & M. L. Boccia (Eds.), *Self-awareness in animals and humans: Developmental perspectives* (pp. 61–80). Cambridge University Press. <https://doi.org/10.1017/CBO9780511565526.007>.
- Gopnik, A., & Meltzoff, A. N. (1996). *Words, thoughts, and theories*. Cambridge, MA: The MIT Press.
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language*, 1(2), 158–171.
- Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In M. K. Davies & T. Stone (Eds.), *Mental simulation: Evaluations and applications. Reading in mind and language* (pp. 53–67). Oxford: Wiley-Blackwell.
- Gordon, R. M. (1996). “Radical” simulationism. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 11–21). Cambridge: Cambridge University Press.
- Gordon, R. M. (2000). Sellars's Ryleans revisited. *Protosociology*, 14, 102–114.
- Gordon, R. M. (2007). Ascent routines for propositional attitudes. *Synthese*, 159(2), 151–165.
- Grice, H. P. (1957). Meaning. In *Reprinted in H. P. Grice (1989). Studies in the way of words* (pp. 213–223). Cambridge, MA: Harvard University Press.
- Grice, H. P. (1969). Utterer's meaning and intention. In *Reprinted in H. P. Grice (1989). Studies in the way of words* (pp. 86–116). Cambridge, MA: Harvard University Press.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech acts (Syntax and semantics 3)* (pp. 41–58). New York: Academic Press.
- Heelas, P., & Lock, A. (Eds.). (1981). *Indigenous psychologies the anthropology of the self*. New York: Academic Press.
- Helming, K. A., Strickland, B., & Jacob, P. (2016). Solving the puzzle about early belief-ascription. *Mind and Language*, 31(4), 438–469. <https://doi.org/10.1111/mila.12114>.
- Heyes, C. M. (2014). False belief in infancy: A fresh look. *Developmental Psychology*, 17, 647–659.
- Heyes, C. M. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Cambridge, MA: Belknap Press: An Imprint of Harvard University Press.

- Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science (New York, N.Y.)*, 344(6190), 1243091. <https://doi.org/10.1126/science.1243091>.
- Hutto, D. D. (2008). *Folk psychological narratives*. Cambridge, MA: The MIT Press.
- Kusch, M. K. (1997). The sociophilosophy of folk psychology. *Studies in History and Philosophy of Science*, 28(1), 1–25.
- Lewis, D. (1969). *Convention: A philosophical study*. Oxford: Wiley-Blackwell.
- Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, 123, 3–32. <https://doi.org/10.1037/0033-2909.123.1.3>.
- Macdonald, P. S. (2003). *History of the concept of mind: Volume 1: Speculations about soul, mind and spirit from homer to hume* (1st ed.). Aldershot, Hants, England; Burlington, VT: Routledge.
- Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy*, 16(5), 595–626. <https://doi.org/10.1023/A:1012203830990>.
- Martínez-Manrique, F., & Vicente, A. (2015). The activity view of inner speech. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2015.00232>.
- McGeer, V. (1996). Is “self-knowledge” an empirical problem? Renegotiating the space of philosophical explanation. *The Journal of Philosophy*, 93(10), 483–515. <https://doi.org/10.2307/2940837>.
- McGeer, V. (2007). The regulative dimension of folk-psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk-psychology reassessed*. Berlin: Springer.
- McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18(2), 259–281. <https://doi.org/10.1080/13869795.2015.1032331>.
- Moore, R. (2017). Gricean communication and cognitive development. *The Philosophical Quarterly*, 67(267), 303–326. <https://doi.org/10.1093/pq/pqw049>.
- Moore, R. (2018). Gricean communication, language development, and animal minds. *Philosophy Compass*, 13(12), e12550. <https://doi.org/10.1111/phc3.12550>.
- Morton, A. (1996). Folk psychology is not a predictive device. *Mind*, 105(417), 119–137.
- Pettit, P. (2018). *The birth of ethics: Reconstructing the role and nature of morality* (K. Hoekstra, Ed.). Oxford: Oxford University Press.
- Putnam, H. (1980). Brains and behavior. *Readings in Philosophy of Psychology*, 1, 24–36.
- Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly (1950-)*, 51(202), 76–82. **(Retrieved from JSTOR)**.
- Scott-Phillips, T. (2014). *Speaking our minds: Why human communication is different, and how language evolved to make it special* (2015th ed.). Houndmills, Basingstoke, Hampshire; New York, NY: Red Globe Press.
- Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, 1, 253–329.
- Simons, M. (2007). Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6), 1034–1056. <https://doi.org/10.1016/j.lingua.2006.05.006>.
- Skyrms, B. (1996). *Evolution of the social contract*. Cambridge, MA: Cambridge University Press.
- Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge, MA: Cambridge University Press.
- Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Retrieved from <https://www.wiley.com/en-us/Relevance%3A+Communication+and+Cognition%2C+2nd+Edition-p-9780631198789>.
- Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language*, 17(1–2), 3–23. <https://doi.org/10.1111/1468-0017.00186>.
- Sterelny, K. (2012). *The evolved apprentice*. Cambridge: The MIT Press.
- Taylor, T. J. (2012). Understanding others and understanding language: How do children do it? *Language Sciences*, 34(1), 1–12. <https://doi.org/10.1016/j.langsci.2011.07.001>.
- Van Cleave, M., & Gauker, C. (2010). Linguistic practice and false-belief tasks. *Mind and Language*, 25(3), 298–328. <https://doi.org/10.1111/j.1468-0017.2010.01391.x>.
- Vinden, P. G. (1999). Children’s understanding of mind and emotion: A multi-culture study. *Cognition and Emotion*, 13(1), 19–48. <https://doi.org/10.1080/026999399379357>.
- Vygotsky, L. S. (1934). *Thought and Language*. The MIT Press.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wellman, H. M. (1990). *The child’s theory of mind*. Cambridge, MA: The MIT Press.

- Wertsch, J. V., & Stone, C. A. (1985). The concept of internalization in Vygotsky's account of the genesis of higher mental functions. In J. V. Wertsch (Ed.), *Culture, communication, and cognition: Vygotskian perspectives* (pp. 162–179). Cambridge University Press.
- Westra, E., & Carruthers, P. (2017). Pragmatic development explains the Theory-of-Mind Scale. *Cognition*, 158, 165–176. <https://doi.org/10.1016/j.cognition.2016.10.021>.
- Zawidzki, T. W. (2008). The function of folk psychology: Mind reading or mind shaping? *Philosophical Explorations*, 11(3), 193–210.
- Zawidzki, T. W. (2013). *Mindshaping. A new framework for understanding human social cognition*. Cambridge, MA: The MIT Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.