Mindshaping

Tadeusz Wiesław Zawidzki

George Washington University

0. Introduction

Mindshaping is a hypothesis about what makes human social cognition distinctive.[1] There are persuasive reasons for holding that distinctively human social cognition is key to our evolutionary success, i.e., the fact that our species dominates the planet, while our closest extant cousin-species teeter on the brink of extinction. Our technological feats rely heavily on our superb abilities at social organization. Unlike any other mammalian species, we routinely coordinate on complex, cooperative projects with complete strangers, often comprising groups of thousands of individuals, many of whom are often even unaware of each other (Seabright, 2004). Our dazzling technological capacities would be impossible without traditions of social learning, enabling the preservation and gradual improvement, over historical time, of techniques of resource extraction, processing, and distribution, and of social communication and interaction. This "cumulative cultural evolution" (Boyd & Richerson, 1996) or "ratchet effect" (Tomasello, 1999) is the only known mechanism for generating the complex and sophisticated tool-kits on which the biological success of human populations depends. The social capacities enabling this "human cooperation syndrome" (Sterelny, 2012) thus

[1] Mameli (2001) introduces the concept of mindshaping as a potential function of mental state attribution. McGeer (1996; 2007) defends a similar notion: the regulative dimension of folk psychology. Zawidzki (2008; 2013) argues that it is the key to understanding the phylogeny of distinctively human social cognition.

distinguish us from other species, and are among the most important pieces to the puzzle of human evolution.

There is a widely held and persuasive theory of the social capacities that make all of this possible (Humphrey, 1980; Tooby & Cosmides, 1995; Baron-Cohen, 1999; Leslie, 2000; Mithen, 2000; Sperber, 2000; Dunbar, 2000, 2003, 2009; Siegal, 2008). According to this received view, our species is distinct in its capacity to correctly represent mental states. It is because we have a neurally implemented capacity to correctly ascertain each other's beliefs, desires, and other propositional attitudes, that we are able to coordinate and cooperate with, and learn from each other so well. There are many disagreements within this broad paradigm. Some concern the form that this neurally implemented cognitive capacity takes: Is it something akin to a scientific theory (re)discovered during human ontogeny (Wellman, 2014)? Or is it instead an innately specified, domain-specific computational module that gradually comes online during human ontogeny, as domain-general capacities for attention and working memory improve (Baillargeon et al., 2010)? Or is it more akin to a skill at simulating the perspectives of others, gradually acquired during human ontogeny (Goldman, 2006)? There are also lively debates about how and where the human brain implements our capacities to read one another's minds (Saxe, 2010). But all parties to these debates share a basic assumption: what sets us apart from other species, and explains the complex social capacities on which our evolutionary success depends, is a neurally implemented, individual capacity to correctly ascertain each other's mental states, especially propositional attitudes, like beliefs and desires.

The mindshaping hypothesis rejects this assumption, and proposes an alternative. According to this alternative, our social accomplishments are not due to an individual, neurally implemented capacity to correctly represent each other's mental states. Rather, they rely on less intellectualized and more embodied capacities to shape each other's minds, e.g., imitation, pedagogy, and norm enforcement. We are much better mindshapers, and we spend much more of our time and energy engaged in mindshaping than any other species. Our skill at mindshaping enables us to insure that *we come to have* the complementary mental states required for successful, complex coordination, without requiring us to solve the intractable problem of *correctly inferring* the independently constituted mental states of our fellows.

Of course no champion of the received view would deny the importance of human mindshaping. Instead, they would claim that mindreading and mindshaping are complementary components of distinctively human social cognition. However there remains a difference in emphasis. On the received view, mindreading is the key innovation on which the rest of the distinctively human socio-cognitive syndrome depends. Without a capacity to correctly represent independently constituted beliefs and desires we could not shape each other's minds as effectively as we do. The mindshaping hypothesis reverses this priority: without appropriate mindshaping, attributing propositional attitudes is pointless and intractable. Furthermore, sophisticated versions of the kinds of *behavior* reading available to our closest primate cousins are sufficient to support the sophisticated mindshaping practices necessary for successful mutual interpretation and coordination in human populations.

The mindshaping hypothesis is a natural ally of "4e" approaches to human social-cognition. Rather than conceptualize distinctively human social cognition as the accomplishment of computational processes implemented in the brains of individuals, involving the correct representation of mental states, the mindshaping hypothesis conceptualizes it as emerging from embodied and embedded practices of tracking and molding behavioral dispositions in situated, socio-historically and culturally specific human populations. Our socio-cognitive success depends essentially on social and hence extended facts, e.g., social models we shape each other to emulate, both concrete ones, e.g., high status individuals, and "virtual" ones, e.g., mythical ideals encoded in external symbol systems. And social cognition, according to the mindshaping hypothesis, is in a very literal sense enactive: we succeed in our socio-cognitive endeavors by cooperatively enacting roles in social structures.

The mindshaping hypothesis is also an ambitious attempt at theoretical integration. It seeks to reconcile insights about human social life from traditions often thought to be unrelated or even antithetical. Most dramatically, it rests on a neo-Darwinian justification for a broadly Nietzschean understanding of human sociality. Rather than conceive of human-specific biological adaptations to social life as neurally implemented computational systems aiming to correctly represent unobservable mental states, we should think of such adaptations as capacities to institute and enact social roles in social structures, and otherwise shape each other in ways that make coordination possible. Put another way, according to the mindshaping hypothesis, culturally specific ideologies to which members of human populations try to conform are

the most adaptive way to solve the coordination problems that characterize distinctively human socio-ecology.

Theoretical stances in the sciences are often motivated by background metaphors that are seldom defended explicitly. For example, the idea that the universe is a blind mechanism replaced the idea that it is a collection of agencies hierarchically organized by a supernatural power during the scientific revolution of the Seventeenth and Eighteenth Centuries. Since the publication of Wilfrid Sellars's *Empiricism and the Philosophy of Mind* (1997), the philosophy and psychology of social cognition have accepted, largely uncritically, the metaphor that forms the centerpiece of that work: social cognition is conceptualized on the model of theoretical inference in science. On this metaphor, the success of our quotidian interactions depends largely on our ability to infer concrete, unobservable mental states that are causally responsible for observable behavior. The mindshaping hypothesis is above all an attempt to introduce, articulate, and defend an alternative metaphor. Rather than conceive of successful human social agents as scientific psychologists, it proposes that we conceive of them as engineers, teachers, pupils, actors, and advocates. Our social success depends on capacities to engineer social environments by teaching and learning roles to play in social structures, and defending our status within such structures through reasoned advocacy. As with all such highly abstract and metaphorical construals, it is difficult to identify empirical tests that vindicate this metaphor over the older one. However, science thrives when different paradigms are brought to bear on the same phenomena, and mindshaping should be understood as a new and viable alternative to the metaphor of human interpreters as scientific psychologists.

In what follows, I put some flesh on this skeletal outline. In Section 1, I make clearer what mindshaping is supposed to be, focusing on how it can be independent of sophisticated mindreading, and illustrating its different varieties. In the process, I identify some ways in which human mindshaping is distinctive. In Section 2, I motivate the mindshaping hypothesis by identifying some puzzles about human social cognition that it seems better suited to address than the received, mindreading view. Section 3 responds to the most persuasive criticism of the mindshaping hypothesis: our best theories of the various capacities on which sophisticated mindshaping relies claim these capacities presuppose the ability to accurately represent mental states. I conclude in Section 4.

1. What Is Mindshaping?

Any attempt to define mindshaping in a way that coheres with the spirit of the hypothesis articulated above immediately runs into a problem. It is unclear how social agents can purposefully and intelligently shape each other's minds without first accurately representing them. Surely, to intelligently shape a mind, whether one's own or another's, one must, at the very least, represent the current state of this mind, the state one desires for it, and some means of minimizing the difference between these. But this way of conceptualizing mindshaping makes it parasitic on sophisticated mindreading, and hence immediately jeopardizes its role in formulating an alternative to the received theory of what makes human social cognition distinctive. Fortunately, Darwin's theory of natural selection provides the resources necessary to make sense of

intelligent behavior, including mindshaping, without assuming sophisticated representation, like the attribution of mental states.

Following a philosophical framework articulated and defended by Ruth Millikan (1984), I define mindshaping by appeal to the proper functions of cognitive mechanisms and the normal conditions on their operation. This definition is neutral on the precise means by which mindshaping occurs, and hence leaves room for the possibility of mindshaping mechanisms that do not rely on the representation of mental states. On Millikan's theory, evolved mechanisms have proper functions, i.e., effects that explain their selection in evolution. So, for example, the proper function of the heart is to pump blood. There are also normal conditions on the execution of such proper functions, e.g., hearts can perform their proper functions only if certain arteries are unobstructed. Millikan applies this framework to cognitive states in order to specify naturalistic conditions for individuating them in terms of their contents. For example, desires are individuated in terms of their proper functions: they aim to get organisms to bring about certain states of affairs; this is what explains their selection in evolution. Mechanisms giving rise to desires to ingest food were selected because they led to the ingestion of food. Beliefs also have proper functions: they aim to combine with desires in order to give rise to practically rational behavior. Mechanisms giving rise to beliefs were selected because they produced beliefs that were accurate enough to guide organisms in ways that led to the satisfaction of their desires, e.g., by representing where food was.

Although Millikan's framework is unabashedly representationalist, when it is applied to the case of mindshaping, it need not be *metarepresentationalist*. That is, it provides the resources required to define mindshaping without presupposing a capacity

to represent mental states. I define mindshaping as a relation between a target mind (the mind being shaped), a cognitive mechanism (the proper function of which involves shaping that mind), and a model that the mindshaping mechanism works to make the target mind match. Thus, mindshaping occurs when a cognitive mechanism selected in evolution for making target minds match models performs its proper function, in Millikan's sense. Clearly, normal conditions on this must include representing the model accurately, but this need not involve the attribution of mental states. The reason is that mindshaping can occur simply in virtue of making the target mind disposed to match a pattern of behavior. Thus, all that needs to be represented is the model's behavior. If there are mechanisms that use such representations to alter the dispositions of target minds in ways that make them more likely to match model behavior, they constitute mindshaping mechanisms that require no representation of mental states.

Let us make this more concrete by applying it to a specific example. A human infant observes an adult model turn on a light panel resting on a table by leaning over and touching it with her forehead (Meltzoff, 1988). After seeing this, the infant is disposed to do the same when put in similar circumstances. This early form of infantile imitation clearly fits the definition of mindshaping. There is some cognitive mechanism in the infant that treats the behavior of the adult as a model to be matched, and disposes the infant to match it. However, there appears to be no reason to assume that the infant need represent the adult's intentions or other mental states in order to shape its mind in this way.[2] On this definition, mindshaping is widespread among non-human

[2] This case, however, is much more complicated than I suggest here. Infants respond very differently to subtle variations in such scenarios, and this leads many researchers to conclude that even such apparently simple imitation relies on sophisticated mindreading. I respond to this claim below, in Section 3.

animals. For example, it applies to baby rats learning which foods to favor based on odors they smell on their mothers' breath (Galef et al., 1983). In all such cases, it is arguable that there are cognitive mechanisms involved that alter behavioral dispositions to approximate behavioral patterns observed in social models.

Such a minimalist understanding of mindshaping raises another problem however. If mindshaping is so widespread among nonhuman animals, how can it be used to explain what is distinctive about human social cognition? Here, there is again a temptation to collapse the distinction between the mindshaping hypothesis and the received view that human social cognition is distinctive in its reliance on sophisticated mindreading. How else can human-specific mindshaping be distinguished from other varieties? A brief survey of recent empirical work on human social learning shows that there are actually at least four ways of distinguishing human-specific mindshaping from other varieties, without assuming that it relies on sophisticated mindreading.

First, the developmental and comparative literature on imitation provides overwhelming evidence of a clear distinction in the *scope* of human vs. non-human imitation. Most non-human species are limited to acquiring *new goals* from observing the behavior of others, while selecting their *own methods* of accomplishing those goals. For example, many bird species can learn from observing conspecifics that food can be extracted from a particular location, but then go onto discover their own method of extracting it, ignoring the method used by their model (Zentall, 2006). The one non-human exception to this appears to be chimpanzees (Horner & Whiten, 2005). They can sometimes acquire both goal and method from a model, but only when there is no alternative method available to them. If they come to discover a different, more efficient

method to accomplish the goal, chimpanzees immediately switch to it, ignoring the model's method. Surprisingly, this is *not* the case with human children. When shown a method to accomplish some goal by an adult model, human children persist in using that method, even after they are made aware of a more efficient method, through demonstrations that components of the modeled method are superfluous or irrelevant to accomplishing the goal. They persist in the modeled method, even when the adult model is not present, and they think they are alone and unobserved; so fear of contradicting an adult cannot explain this phenomenon. Such "overimitation" (Lyons et al., 2007; Nielsen & Tomaselli, 2010) is a distinctively human form of mindshaping. Yet, it does not appear to require sophisticated mindreading, like the attribution of propositional attitudes. Human children need only represent the goal of an adult model's behavior, and the precise sequence of behavioral steps used by her in accomplishing the goal.

A second distinctive feature of human mindshaping is a plausible explanation of phenomena like overimitation. Matching model behavior, for humans but not non-humans, appears to be its own reward. Nonhumans will imitate a model to the extent that it helps accomplish some further goal, like extracting food from a novel location (Zentall 2006). Humans, on the other hand, seem to find matching a model's behavior intrinsically rewarding. This explains overimitation: children imitate the precise means of accomplishing a desired goal, even if they are aware of more efficient means of accomplishing the same goal. It is plausible that this is due to the fact that they experience some kind of reward signal for matching model behavior precisely that outweighs the value of accomplishing the goal as efficiently as possible. There are other

forms of mindshaping that also appear to show intrinsic motivation. For example, the costly punishment of norm flouters appears widespread in human populations (Henrich et al., 2006). Since this involves incurring a cost in order to punish counter-normative behavior, it suggests that shaping minds to respect norms is intrinsically motivating (Sripada & Stich, 2006). Thus, the fact that human mindshaping appears intrinsically motivating is another feature that sets it apart from non-human varieties.

A third distinctive feature of human mindshaping is the socially extended nature of many human mindshaping mechanisms. For example, although there are some limited examples of pedagogy among non-human species (Thornton & McAuliffe, 2006), none come close to the sophistication and pervasiveness of pedagogy in human populations. Unlike imitation, pedagogy relies on extra-mental components, e.g., active guidance by a teacher. Perhaps the most pervasive form of pedagogy in human populations takes the master-apprentice form, in which experts provide subtle, behavioral guidance to novices (Sterelny, 2012). This kind of pedagogy is possible without a sophisticated language or even, arguably, a sophisticated theory of mind. It seems to involve the gradual tuning of novice behavioral dispositions via skilled expert demonstrations and interventions. There is evidence that human infants are innately adapted to this style of “natural pedagogy” (Csibra & Gergely, 2011). For example, from a very young age they interpret certain stereotyped, adult communicative behaviors, such as eye contact, as overtures to demonstrating novel, generalizable information about referential objects specified by subsequent stereotyped behaviors, such as eye saccades. As human civilization grew more complex, socially extended mindshaping mechanisms became more sophisticated. We now have institutions of formal education

and sanctioning to shape group members to play highly specific roles in very complex social structures.

A fourth distinctive feature of human mindshaping concerns its use of abstract, fictional models. All non-human mindshaping involves matching some aspect of the observable behavior of another, actual, concrete individual. But many of the most sophisticated forms of human mindshaping involve matching the behavior of fictional models like protagonists of myths, or morally ideal agents. This is possible due to the representational power of public language. We can formulate public representations of non-actual states of affairs, including non-actual patterns of behavior by fictional agents that we go on to imitate. This form of mindshaping does not obviously require sophisticated mindreading, only a public language for representing the *behavior* of fictional models.

Thus, it is possible to describe forms of mindshaping unique to humans without assuming that they rely on sophisticated mindreading, like propositional attitude attribution. Distinctively human mindshaping is (1) intrinsically motivating, (2) maximally flexible in the aspects of model behavior that it seeks to match (as in overimitation), (3) often reliant on external components (e.g., expert guidance and sophisticated pedagogical practices and institutions), and (4) often seeks to match the behavior of fictional models. Of course, it is empirically possible that these varieties of mindshaping presuppose, in practice, sophisticated mindreading, including the attribution of propositional attitudes. I respond to variants on this objection below, in Section 4. However, there are good empirical reasons to doubt this. First, as I argue next, in Section 3, it is puzzling how full-blown propositional attitude attribution can be accurate,

timely, and computationally tractable at the same time; so it is not clear that it can support the mindshaping practices described above. Second, there is already empirical evidence of low-level mindshaping mechanisms in humans that appear independent of propositional attitude attribution. For example, a recent fMRI study identified low-level mechanisms of social conformism that make use of basic reward circuits involved in behavioral conditioning (Klucharev, et al., 2009). Signals that one's behavior fails to conform to group behavior can play the same role as prediction errors in individual learning (i.e., when a planned behavior does not have its intended effect), driving individuals to conform to their groups. There is no evidence that such mechanisms require the representation of propositional attitudes.

I now turn to a discussion of various explanatory advantages of the mindshaping hypothesis over the received, mindreading hypothesis. It turns out that there are a number of deep problems about distinctively human social cognition that arise for the received view that can be solved or avoided on the mindshaping hypothesis.

3. The Advantages of Mindshaping

The basic story motivating the received view that human social cognition is distinctive in its reliance on sophisticated mindreading is well known and persuasive. The idea is that our prehistoric ancestors faced strong selection pressures for Machiavellian intelligence (Humphrey 1980). Due to unusually large and complex groups, they had to learn both how to take advantage of others and how to prevent others from taking advantage of them. This triggered an evolutionary "arms race" the result of which was an advanced theory of mind, supporting the reliable attribution of propositional attitudes

like belief and desire. It is easier to take advantage of others, e.g., by deceiving them, if one can correctly represent their mental states, especially false beliefs. Once this capacity is prevalent in a population, it pays to detect deception, so the capacity to attribute more complex mental states, like intentions to deceive or (higher order) beliefs about false beliefs, is incentivized. Once the capacity to detect deception is widespread in a population, a new incentive arises: knowing when one's deception is likely to be detected. This requires an even more sophisticated theory of mind. In this way, on the received view, social complexity in human prehistory triggered Machiavellian adaptations, in the form of capacities to attribute increasingly complex types of mental states.

However, there are a number of serious problems with this picture. First, it appears that our closest non-human cousins, chimpanzees, live in groups large and complex enough to incentivize deception and deception-detection. The Machiavellian nature of "chimpanzee politics" is well known (de Waal, 2000). Yet the consensus in experimental, comparative psychology is that chimpanzees are incapable of attributing full-blown propositional attitudes, like beliefs and sophisticated desires or intentions (Call & Tomasello, 2008). They can certainly interpret conspecific behavior in terms of its goals, and the perceptions or knowledge by which it is guided. But there is little evidence that they conceive of their conspecifics as animated by unobservable states of mind with complex relations to each other and observable behavior. Given that chimpanzees and other intelligent non-human animals living in complex social groups appear to manage Machiavellian intelligence without the attribution of full-blown propositional attitudes, the evolutionary story behind the received view that human

social cognition is distinctive in this capacity seems unmotivated. It seems that sophisticated mindreading, like the accurate attribution of propositional attitudes, is unnecessary for Machiavellian success in socially complex groups.

A second problem with the received, mindreading view concerns the computational tractability of accurate mindreading. The attribution of full-blown propositional attitudes is holistically constrained. There is no simple, one-to-one mapping between the observable behavior and circumstances of a target of attribution and the propositional attitudes that she tokens. It is a familiar philosophical point that any finite set of beliefs and desires is compatible with any observable behavior or circumstance, given appropriate adjustments to background beliefs and desires (Morton 1996, 2003; Bermúdez 2003, 2009). An agent may want to stay dry and believe that it is raining, while standing in the rain with an unopened umbrella, due to certain background beliefs, like the belief that opening the umbrella will trigger a bomb, or that the umbrella does not work, etc. Given this holism, it is hard to see how interpreters can come to accurate attributions and predictions in time to react appropriately to rapidly evolving, dynamic, real-world social situations.

This is particularly problematic if, as seems to be the case with non-human Machiavellian intelligence, successful social cognition can make do with far less complicated cognitive capacities. If tracking the goals and information access of conspecifics is enough to support adaptive behavior in real-world social contexts, then why would natural selection support a further capacity to correctly represent the actual psychological causes of conspecific behavior, together with the complex “ceteris paribus” laws that link them (Gauker 2003, p. 240), when this appears to be too

computationally demanding to make a difference in real time? This point is often lost on empirical researchers because they equate the attribution of beliefs and desires with far simpler capacities like tracking goals and information access. However, it is certainly possible to perceive a bout of behavior as aiming at a goal, and informed by a (potentially non-actual) worldly situation, without conceiving of it as caused by an unobservable mental state with complex connections to behavior, encoded in "ceteris paribus" laws. Such an embodied, perceptual attunement to relational properties of bouts of behavior is more likely to support timely and accurate behavioral anticipation, than inference over hidden mental states with tenuous connections to behavior. And there is increasing consensus that the quotidian social cognition of non-humans, human infants, and even human adults in most circumstances takes something like this less intellectualized form (Hutto, 2008; Gallagher & Hutto, 2008; Apperly & Butterfill, 2009; Apperly, 2011; Butterfill & Apperly, 2013).

A final set of problems with the received view of distinctively human social cognition as dependent on reliable propositional attitude attribution concerns the kinds of social situations likely faced by our prehistoric ancestors, especially opportunities to deceive and otherwise free ride on the cooperative dispositions of group-mates. Given the holism problem, and the relatively long time-course for developing the capacity to attribute higher order propositional attitudes in ontogeny (Perner & Wimmer, 1985), it is unlikely that deception in prehistory was checked through more sophisticated mindreading. But it had to be checked somehow. After all, our capacity to coordinate on extremely complex cooperative projects with large numbers of individuals, of whom we

have little personal knowledge, is one of the most important distinguishing marks of human sociality.

Furthermore, even assuming largely cooperative dispositions among early human populations, it is not obvious how sophisticated mindreading could help them solve *coordination problems*. You might think that knowing what your partner in a coordination problem is thinking would help you select behavior that leads to successful coordination. However, things are not so simple. The problem is that your partner is in exactly the same situation as you: she must know what you are thinking. But if the mindreading is accurate, she will learn only that you are thinking about what she is thinking, just as you will learn only that she is thinking about what you are thinking (Gilbert, 1996; Bacharach, 2006). Trying to reconnect a disconnected telephone conversation is a good example of this. If both parties call back at the same time then they will not reconnect. If both parties wait for the other to call, then they will not reconnect. They must somehow figure out who is to call back and who is to wait. But mindreading appears to be of no help here, since A's accurate mindreading of B reveals only that B is trying to read A's mind, and vice versa. Thus, it is not even clear that sophisticated, accurate mindreading is sufficient for solving simple coordination problems of the kind likely faced by our prehistoric ancestors.

The mindshaping hypothesis, coupled with some plausible conjectures about the role of "cultural group selection" (Henrich, 2004) in human prehistory offer a way of avoiding these problems. The basic idea is simple: if members of human populations are shaped via the kinds of mechanisms discussed above to routinely adopt similar or complementary mental states and behavioral dispositions when coordinating on

cooperative tasks, then the holism problem should never arise, and deception and other forms of free riding should be rare. If there are mindshaping practices in human populations that insure that most of one's potential interactants react in familiar, coordination and cooperation enhancing ways, e.g., by conforming to norms promulgated via the behavior of well-known real and fictional models, then our social cognition should succeed even if it relies on relatively unsophisticated socio-cognitive mechanisms. We should manage to track the behavior of our group mates simply by attributing to it goals by which we think people ought to be motivated in such circumstances, and assuming it is guided by information we think people ought to find relevant to such goals. This should work because, as a matter of fact, most people with whom we interact are products of similar mindshaping regimes as we are; such mindshaping prevents the radical cognitive heterogeneity that might thwart such simple heuristics. Furthermore, groups composed of members shaped to favor cooperation over free riding and to respect coordination norms should outcompete groups composed of members not so shaped, and this could explain, via cultural group selection, the evolution of human capacities for coordination on cooperative projects (Henrich, 2004).

On the face of it, tracing distinctively human social cognition to virtuosity at the sorts of mindshaping practices I described in Section 2 appears to avoid the major shortcomings of the received view that it depends on sophisticated mindreading, like the attribution of propositional attitudes. However, many will find this unconvincing. It is not obvious that the mindshaping mechanisms described in Section 2, e.g., overimitation, natural pedagogy, and the emulation of fictional agents encoded in public language,

require no sophisticated mindreading, no attribution of propositional attitudes. In fact, many theories of such mechanisms posit precisely such capacities. This is the most serious problem for the mindshaping account, considered as an alternative to the received view. I address it next.

4. Sophisticated Mindshaping without Sophisticated Mindreading

The received view is not just a hypothesis about the phylogenetically most important component of distinctively human social cognition. It is also central to most current explanations of most sophisticated, human social capacities. The distinctively human mindshaping mechanisms and practices I discussed in Section 2 are no exception. Consider overimitation. The capacity of human infants to imitate adult models who switch on light panels lying on tables with their foreheads is a classic example of overimitation: they learn an inefficient method of accomplishing a goal which they could accomplish much more easily, i.e., by switching the light on by hand. Subsequent experiments show that this is not mere blind copying (Gergely, et al., 2002). If the adult model's hands are otherwise occupied or out of view when she switches on the light panel with her forehead, infants learn to switch on the light panel using the most efficient method available to them: with their hands. A natural interpretation of this is that infant imitators rely on the attribution of intentions to adult models. When an adult model switches on the light panel with her forehead while her hands are free, she must intend specifically to use her forehead, since she could more easily switch it on with one of her hands. But, when an adult model switches on the light panel with her forehead

while her hands are occupied, she must intend to switch it on by the most efficient method available to her.

Natural pedagogy is also typically explained in terms of sophisticated infant mindreading. For example, Csibra (2010) argues that it relies on the capacity to attribute higher order intentions. On this explanation, infants interpret eye contact as expressing the communicative intention that immediately ensuing behavior be interpreted as intending to inform the infant of novel information concerning some salient object. On this view, natural pedagogy relies on infant capacities to attribute second-order propositional attitudes.

Finally, in Section 2 I suggested that our capacity to copy non-actual patterns of behavior by fictional agents encoded in public language is one of the most sophisticated forms of distinctively human mindshaping. But mastering a public language is routinely explained in terms of capacities to attribute complex propositional attitudes. For example, according to Sperber and Wilson (2002), all linguistic communication presupposes the capacity to attribute nested intentions and beliefs. And, according to Bloom (2002), word learning requires the capacity to attribute referential intentions to adult models. Thus, it would seem that any mindshaping reliant on the representation of model behavior in public language presupposes sophisticated mindreading. If these theories of overimitation, natural pedagogy, and language use are correct, then the distinctively human mindshaping practices and mechanisms discussed in Section 2 presuppose sophisticated mindreading, and hence cannot constitute an alternative to the received view of what makes human social cognition distinctive.

This whole question turns on what we mean by "sophisticated mindreading" and "propositional attitude attribution". Most philosophers of psychology follow Wilfrid Sellars (1997) when interpreting these concepts. Propositional attitudes are treated as states of an unobservable causal nexus responsible for an agent's behavior: the agent's mind. Furthermore, as I noted above, their relations to observable circumstances and behavior are holistically constrained: what one does in specific circumstances depends on indefinitely broad networks of propositional attitudes; hence, it should be difficult to determine an agent's propositional attitudes based on observations of finite bouts of behavior, and an agent's future behavior based on attributions of finite sets of propositional attitudes. Finally, if we take the Sellarsian picture seriously, and think of propositional attitude attribution on the model of scientific hypotheses about unobservable causal factors, then propositional attitude attribution should involve a strong appearance/reality distinction. Think of medical diagnosis here. Because the causes of symptoms, e.g., bacteria, are unobservable factors independent of the symptoms, it is always possible that two qualitatively similar sets of observable symptoms are products of radically different unobservable factors. Appearance does not determine reality. If propositional attitude attribution is supposed to be like this, then it requires an appreciation of the possibility that two qualitatively indistinguishable patterns of observable behavior are caused by radically different sets of propositional attitudes.

If we conceive of propositional attitude attribution along these lines, there is good reason to doubt that sophisticated human mindshaping, like overimitation, natural pedagogy, and language-assisted mindshaping presuppose propositional attitude attribution. For one thing, the speed and fluency with which infants overimitate, interpret

pedagogical interactions, and engage in linguistic interactions suggest that they are not engaging in scientific reasoning about unobservable causes with tenuous connections to observable behavior. Secondly, it is very unlikely that such mindshaping capacities rely on an appreciation of a strong behavioral appearance / mental reality distinction. There is no evidence that human infants can conceptualize the possibility that qualitatively indistinguishable patterns of behavior might be products of radically different sets of propositional attitudes. Typically, when tested for capacities to interpret behavior, infants and children show no hesitation: they see behavior as unambiguously directed at specific goals and informed by specific situations. Thus, if we think of sophisticated mindreading and propositional attitude attribution along Sellarsian lines, there is no reason to suppose that distinctively human mindshaping depends on them.

How else might we conceive of the socio-cognitive capacities underlying distinctively human mindshaping? One possibility is to think of human mindshapers and "mindshapees" as operating with an ontology of informed, goal-directed bouts of behavior. To be goal-directed, a bout of behavior must be predictable on the assumption that it selects the most efficient of observable means to some observable end-state. To be informed by some (possibly non-actual) situation, a bout of behavior must count as the most efficient of observable means to some observable end-state *relative to that situation*. One can perceive bouts of behavior as goal directed and informed in these ways, without thinking of them as caused by representations of goals and information within the unobservable minds of agents. A number of theorists have defended the hypothesis that human infant and even most human adult social cognition relies on such minimalist assumptions about behavior (Gergely & Csibra 2003; Apperly

& Butterfill 2007; Apperly 2011; Butterfill & Apperly 2013). Although these are still early days, this is a viable hypothesis about the socio-cognitive capacities underlying sophisticated human mindshaping that does not presuppose a capacity for sophisticated mindreading, at least not in a Sellarsian sense.

Furthermore, there is evidence that our closest non-human cousins, i.e., chimpanzees, also sometimes rely on the assumption that their conspecifics engage in goal-directed bouts of behavior that are informed by (possibly non-actual) worldly states (Crockford, et al., 2011). Thus, if sophisticated human mindshaping relies exclusively on similar assumptions, then distinctively human mindshaping does not presuppose *distinctively* human means of interpreting behavior. Motivations to treat conspecific behaviors as models for one's own seem more important to distinguishing human from non-human social cognition, than assumptions about goal-directedness or informedness. Of course, it is true that humans deploy such assumptions about behavior in far more subtle, complex, and diverse ways than chimpanzees or any other nonhuman species. But we can conceive of such differences as products of the gradual evolutionary accumulation of tweaks to a basic socio-cognitive capacity we share with other social primates, aimed at improving mindshaping practices, i.e., making us better overimitators, pupils, teachers, and conformers to linguistically encoded, fictional models. On this view, though there are differences between humans and nonhumans in the scope and sophistication of our means of interpreting behavior, these are products of evolution for better mindshaping, which was necessary in our distinctive, cooperative socio-ecology. Mindshaping remains the source of our socio-cognitive distinctiveness.

This perspective raises another worry, however. If our most important socio-cognitive feats are products of motivations to use others as models for our own behavior, guided by enhanced versions of ancient primate capacities to interpret behavior as goal-directed and informed by (possibly non-actual) worldly states, then why do we engage in sophisticated mindreading at all? What function is left for the attribution of full-blown propositional attitudes? On the mindshaping hypothesis, this is a late-arriving capacity involved in the *justification* rather than the prediction of behavior. Given our highly inter-dependent, cooperative socio-ecology, one's social status, and hence, ultimate biological success, is heavily dependent on being perceived as a competent and reliable potential partner in coordination on complex, cooperative projects. Therefore, it is unsurprising that anomalous behavior jeopardizing one's reputation for such competence and reliability, e.g., misinforming someone or reneging on an explicit commitment, immediately puts one's social status at risk. In fact our status may be even more precarious. Assuming that our potential cooperation partners see behavior that resembles their own, is familiar, and, in general, respects prevalent norms as a signal of general trustworthiness,[3] any kind of deviant behavior might put status at risk. In such circumstances, it is useful to have a practice of rehabilitating status, of explaining away apparent deviance by showing the behavior to be reasonable in the light of propositional attitudes of which witnesses might be unaware. On this view, the attribution of full-blown propositional attitudes first gains traction as a means of normalizing apparently deviant behavior (Bruner 1990). This proposal even has some empirical support (Malle et al. 2007).

[3] An assumption for which there is some empirical evidence (Wiltermuth & Heath, 2009).

If full-blown propositional attitude attribution functions to mitigate the social fall-out from apparently deviant behavior, many of the properties that make it unsuitable as a prediction device start to look adaptive. Holism seems tailor-made for this function, as any behavior can be made to accord with any set of propositional attitudes, given appropriate adjustments to background propositional attitudes. A behavioral appearance / mental reality distinction also begins to make sense: the whole point of rationalization is that behavior seemingly caused by one set of mental states might actually be caused by a different set. Once such a practice prevails in a population, one would expect members to start interpreting all behavior, both their own and others', in terms of potential rationalizations. There would also be incentives to police behavior to insure that it stays rationalizable in terms of widely tolerated propositional attitudes. In such populations, individuals would actively shape themselves and each other to conform to expectations generated by propositional attitude attributions. On the mindshaping hypothesis, this dynamic characterizes many modern, human populations. The idea that propositional attitude attribution is the most important component of human social cognition is an illusion born of these relatively recent, socio-historical circumstances.

5. Conclusion

Of necessity, this has been a relatively superficial exploration of the mindshaping hypothesis. But the broad contours, I hope, are clear. On this hypothesis, what sets human social cognition apart from other varieties is our capacity to shape each other and ourselves into the kinds of agents that can coordinate successfully on cooperative

projects. This capacity does not presuppose sophisticated mindreading, like propositional attitude attribution. Instead, it relies on sophisticated versions of behavior reading strategies also present in non-human species, coupled with unusually strong motivations to copy the behavior of others. This variant of the basic primate socio-cognitive tool kit was adaptive in the distinctively cooperative socio-ecologies of pre-historic hominins. Eventually it gave rise to such human-specific phenomena as overimitation, pervasive pedagogy, and the imitation of fictional agents. The contrast with the received view is clear: we are not scientific psychologists first, and only later social engineers, teachers, pupils, actors, and advocates. In fact, this gets things almost exactly backward. It is only relative to a social environment engineered via sophisticated practices of mindshaping that the practice of justifying behavior in terms of full-blown propositional attitudes makes sense.

Also clear from the foregoing are the rich affinities between the mindshaping hypothesis and "4e" approaches to social cognition. With many champions of such approaches, the mindshaping hypothesis rejects the common assumption that solving the "other minds" problem via inference to unobservable mental states is the basis for human socio-cognitive competence (Hutto, 2008; Gallagher & Hutto, 2008). The kinds of low-level, behavior tracking and shaping assumed in my characterization of mindshaping mechanisms plausibly qualify as examples of embodied and embedded cognition. They involve attunement to low-level, bodily dispositions, contextualized to specific socio-cultural embeddings. Furthermore, external or extended structures are central to many human-specific forms of mindshaping, including the roles of teachers and pedagogical institutions, as well as both concrete and abstract (fictional) models,

the latter encoded in public systems of representation. It is true that, in assuming Millikan's teleosemantics in the definition of mindshaping, I accept a kind of representationalism that some champtions of "4e" approaches to cognition might find anathema. However, there are moderate varieties of "4e" approaches that allow for the possibility of some forms of representation (Clark 1997), and at least one prominent defender of "4e" approaches endorses Millikan's teleosemantics, though as a theory of biosemiotics rather than biosemantics (Hutto 2008). Thus, there is much potential for a productive mutualism between the mindshaping hypothesis and "4e" approaches to human social cognition.

References

Apperly, I. A. (2011). Mindreaders. Hove: Psychology Press.

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

Bacharach, M. (2006). Beyond individual choice. Princeton: Princeton University Press.

Baillargeon, R., Scott, R., & He, Z. 2010. False-belief understanding in infants. Trends in Cognitive Sciences 14(3): 110-118.

Baron-Cohen, S. (1999). The evolution of a theory of mind. In M. C. Corballis & S. E. G. Lea (Eds.), The descent of mind. New York: Oxford University Press.

Bermúdez, J. L. (2003). The domain of folk psychology. In A. O'Hear (Ed.), Minds and persons. Cambridge: Cambridge University Press.

Bermúdez, J. L. (2009). Mindreading in the animal kingdom. In R. Lurz (Ed.), The philosophy of animal minds. Cambridge: Cambridge University Press.

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

Boyd, R., & Richerson, P. J. (1996). Why culture is common but cultural evolution is rare. Proceedings of the British Academy, 88, 73–93.

Bruner, J. (1990). Acts of meaning. Cambridge, MA: Harvard University Press.

Butterfill, S. & Apperly I.A. (2013). How to construct a minimal theory of mind. Mind and Language 28(2), 606-637.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences, 12, 187–192.

Clark, A. (1997). Being there: Putting brain, body and world together again. Cambridge, MA: MIT Press.

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2011). Wild chimpanzees inform ignorant group members of danger. Current Biology, 22(2), 142–146.

Csibra, G. (2010). Recognizing communicative intentions in infancy. Mind and Language, 25, 141–168.

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. Philosophical Transactions of the Royal Society of London: Series B, 366, 1149–1157.

de Waal, F. B. M. (2000). Chimpanzee politics. Baltimore: Johns Hopkins University Press.

Dunbar, R. (2000). On the origin of the human mind. In P. Carruthers & A. Chamberlain (Eds.), Evolution and the human mind: Modularity, language, and meta-cognition (pp. 238–253). Cambridge: Cambridge University Press.

Dunbar, R. (2003). The social brain: Mind, language, and society in evolutionary perspective. Annual Review of Anthropology, 32, 163–181.

Dunbar, R. (2009). Why only humans have language. In R. Botha & C. Knight (Eds.), The prehistory of language (pp. 12–35). Oxford: Oxford University Press.

Galef, B. G., Wigmore, S. W., & Kennett, D. J. (1983). A failure to find socially mediated taste aversion learning in Norway rats (R. norvegicus). Journal of Comparative Psychology, 97(4), 358–363.

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev et al. (Eds.), The shared mind. Amsterdam: John Benjamins.

Gauker, C. (2003). Words without meaning. Cambridge, MA: MIT Press.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. Nature, 415, 755–756.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. Trends in Cognitive Sciences, 7(7), 287–292.
Gilbert, M. (1996). Living together: Rationality, sociality, and obligation. Lanham, MD: Rowman & Littlefield.

Goldman, A. I. (2006). Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford: Oxford University Press.

Henrich, J. (2004). Cultural group selection, coevolutionary processes, and largescale cooperation. Journal of Economic Behavior and Organization, 53, 3–35.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., …, Ziker, J. (2006). Costly punishment across human societies. Science, 312, 1767–1769.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). Animal Cognition, 8, 164–181.

Humphrey, N. (1980). Nature's psychologists. In B. D. Josephson & V. S. Ramachandran (Eds.), Consciousness and the physical world (pp. 57–80). Oxford: Pergamon Press.

Hutto, D. D. (2008). Folk psychological narratives: The sociocultural basis of understanding reasons. Cambridge, MA: MIT Press.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. Neuron, 61, 140–151.

Leslie, A. M. (2000). How to acquire a "representational theory of mind." In D. Sperber (Ed.), Metarepresentations: A multidisciplinary perspective (pp. 197–223). Oxford: Oxford University Press.

Lyons, et al. (2007). The hidden structure of overimitation. Proceedings of the National Academy of Sciences of the United States of America, 104(50): 19751–19756.

Malle, B. F., Knobe, J., & Nelson, S. E. (2007). Actor-observer asymmetries in behavior explanations: New answers to an old question. Journal of Personality and Social Psychology, 93, 491–514.

Mameli, M. (2001). Mindreading, mindshaping, and evolution. Biology and Philosophy, 16, 597–628.

McGeer, V. (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. Journal of Philosophy, 93(10), 483–515.

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), Folk psychology re-assessed (pp. 137–156). Dordrecht: Springer.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. Developmental Psychology, 24, 470–476.

Millikan, R. G. (1984). Language, thought, and other biological categories: New foundations for realism. Cambridge, MA: MIT Press.

Mithen, S. (2000). Palaeoanthropological perspectives on the theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), Understanding other minds. Oxford: Oxford University Press.

Morton, A. (1996). Folk psychology is not a predictive device. Mind, 105(417), 119–137.

Morton, A. (2003). The importance of being understood: Folk psychology as ethics. London: Routledge.

Nielsen, M., & Tomaselli, K. (2010). Overimitation in Kalahari Bushman children and the origins of human cultural cognition. Psychological Science, 21(5), 729–736.

Perner, J. & Wimmer, H. (1985). “John *thinks* that Mary *thinks* that…” attribution of second-order beliefs by 5- to 10-year-old children. Journal of Experimental Child Psychology, 39(3), 437-471.

Saxe, Rebecca. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. In Alan Leslie & Tamsin German (Eds.), *Handbook of Theory of Mind*. Psychology Press.

Seabright, P. (2010). The company of strangers. Princeton: Princeton University Press.

Sellars, W. (1997). Empiricism and the philosophy of mind. Cambridge, MA: Harvard University Press.

Siegal, M. (2008). Marvelous minds: The discovery of what children know. Oxford: Oxford University Press.

Sperber, D. (Ed.). (2000). Metarepresentations. New York: Oxford University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity, and mind-reading. Mind and Language, 17(1 & 2), 3–23.

Sripada, C., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), The innate mind: Culture and cognition (pp. 280–301). New York: Oxford University Press.

Sterelny, K. (2012). The evolved apprentice. Cambridge, MA: MIT Press.

Thornton, A., & McAuliffe, K. (2006). Teaching in wild meerkats. Science, 313, 227–229.

Tomasello, M. (1999). The cultural origins of human cognition. Cambridge, MA: Harvard University Press.

Tooby, J., & Cosmides, L. (1995). The language of the eyes as an evolved language of mind. In S. Baron-Cohen (Ed.), Mindblindness: An essay on autism and theory of mind. Cambridge, MA: MIT Press.

Wellman, H. (2014). Making minds. New York: Oxford University Press.

Wiltermuth, S. S., & Heath, C. (2009). Synchrony and cooperation. Psychological Science, 20(1), 1–5.

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. Cambridge, MA: MIT Press.

Zentall, T. R. (2006). Imitation: Definitions, evidence, and mechanisms. Animal Cognition, 9, 335–353.