Chapter Seven: The Role of Sophisticated Mindreading in Human Mindshaping

1. Preamble

In Chapter Four and Chapter Five, I proposed and defended a hypothesis about the phylogeny of three of the four components of the human socio-cognitive syndrome: sophisticated mindshaping, cooperation, and complex language. I argued that, contrary to the received view, none of this requires sophisticated mindreading, understood as the attribution of full-blown propositional attitudes, with tenuous, holistically constrained causal influence over behavior. For the most part, these components of the human socio-cognitive syndrome are products of changes in social motivations rather than social cognition. Our ancestors found themselves in a socio-ecological niche in which motivations to cooperate, learn from and conform to each other, and keep track of and enforce normative statuses yielded biological advantages. I do acknowledge that this socio-ecological niche likely selected for improvement in some socio-cognitive capacities that we share with non-human primates, especially the capacity to adopt the intentional stance.[1] We are better at this than other primates, especially at factoring differences in information access into our judgments of means-ends rationality. The cognitive homogeneity that results from pervasive mindshaping in human populations makes our virtuosity at adopting the intentional stance possible, because it makes it more likely that interpreters and their targets attend to similar information and make similar judgments of means-ends rationality. As I argued

[1] As the account I defend makes clear, there are also other, distinctively human socio-cognitive capacities selected for because of the changes to human socio-ecology wrought by pervasive mindshaping. As I noted in Chapter Two, from a very young age, human infants interpret certain stereotyped adult gestures as preludes to pedagogical interactions. In Chapter Five, I argued that our capacity to produce and process structurally complex, communicative performances evolved due to the importance, in human prehistory, of reliable signals of commitment to and capacity for coordination on cooperative projects.

in Chapter Five and Chapter Six, improved and more powerful versions of the intentional stance can also support more sophisticated mindshaping practices, especially those involving language.

Assuming that this is on the right track, it raises a difficult question. If pervasive, sophisticated mindshaping, together with sophisticated versions of the intentional stance are sufficient to explain human cooperation, coordination and language, what explains the phylogeny of sophisticated mindreading, as I have characterized it? If dispositions toward conformity, imitation, pedagogy, and the tracking and enforcement of normative statuses, together with the capacity to parse behavioral sequences into goals and rationally/informationally-constrained means of achieving them are all that is required to maintain the human socio-cognitive syndrome, then why do we need to attribute full-blown propositional attitudes, with tenuous, holistically-constrained causal influence on behavior? Answering this question is the main burden of this final chapter. I shall argue that such sophisticated mindreading derives from the practice of undertaking and attributing discursive commitments (Sellars 1963; Brandom 1994; Frankish 2004). Chapter Five and Chapter Six neutralized some obvious problems with making sophisticated mindreading parasitic on language in this way. With such worries out of the way, I can now turn to a more detailed exploration of this suggestion.

In section 2, I consider two alternative views that make sophisticated mindreading parasitic on discursive competence: Hutto (2008) and Bermudez (2003a; 2003b; 2009). I argue that their reasons for seeing propositional attitude attribution as parasitic on language are not compelling, and introduce my own reasons for this. In section 3, I explain the increasingly important role that the capacity to track and undertake discursive commitments likely played in human prehistory, given the phylogenetic story defended in Chapter Four and Chapter Five. I

then argue that, rather than enabling accurate mindreading, the primary raison d'être of full-blown propositional attitude attribution has always been what Malle et al. (2007) call "impression management": the maintenance, diminution, or rehabilitation of status in the wake of apparently counter-normative behavior, like apparent reneging on discursive commitments. In section 4, I provide more detail about how the capacity for discursive commitment is implemented, focusing on the sorts of mindshaping mechanisms that enable us to conform to our discursive commitments. I argue that self-constitution in terms of publicly available narrative plays a central role in this, and also explains how our capacity to shape ourselves can play both private, cognitive roles and public, coordinative roles. Section 5 wraps up the entire project, relating it to some prominent traditions in philosophy and psychology.

2. Why Propositional Attitude Attribution Depends on Language

A number of theorists maintain that propositional attitude attribution presupposes competence in a public language (Davidson 2001; Clark 1998; Bermudez 2003b; 2009; Hutto 2008). However, I do not agree with most of the reasons that have been offered for this claim. Hutto (2008), following Davidson (2001), argues that both the capacity to *token* propositional attitudes and the capacity to *attribute* them presuppose competence in a public language. Davidson's reasons for this are largely epistemological. If there is no principled way of determining the precise content of an agent's propositional attitudes, then there is no fact of the matter regarding which propositional attitudes the agent tokens. The only principled way of determining the precise content of an agent's propositional attitudes is by interpreting the agent's utterances of public

language sentences. Hence only speakers of public languages token and attribute propositional attitudes.[2]

Hutto provides supplementary arguments in support of Davidson's position. His key contention is that the intensionality of propositional attitude attributions that I discussed in Chapter Six requires that targets of such attributions represent their contents under linguistic modes of presentation. For example, consider a dog barking up a tree into which it has just chased a squirrel. It is tempting to attribute to the dog the belief that the squirrel is up the tree. But, in what sense does the dog think of the squirrel *as a* squirrel? Clearly, the dog does not conceptualize the squirrel as language users do: it does not think of it as a member of a species of mammal that semi-hibernates in the winter, spends the non-winter months stocking up on nuts, etc. So, how does the dog think of the squirrel? It seems impossible to say, since we can specify the dog's mode of presentation only using words drawn from a public language, all of which have connotations of which the dog is unaware. This is problematic because specifying the mode of presentation under which believers represent the contents of their beliefs is key to linking their beliefs to their behavior; this is why propositional attitude attributions are intensional. We predict that Lois Lane will not kiss Clark Kent because she represents him as a dorky reporter, rather than as the super hero she loves.

This argument, considered on its own, is not persuasive. Just because it is difficult to express in language the mode of presentation under which a non-lingual agent represents the contents of her beliefs, does *not* mean that she does *not* represent the contents of her beliefs under, presumably, non-linguistic modes of presentation. This is the problem with Davidson's

[2] If only speakers of public languages token propositional attitudes the, *a fortiori*, only speakers of public languages attribute them, since to attribute a propositional attitude one must token a higher order one, i.e., the belief that one's interpretive target has the attributed propositional attitude. Also, on a Davidsonian approach, one cannot determine which propositional attitude to attribute to a target without first interpreting her public language utterances.

general approach: epistemological constraints related to our typical evidence for and means of attributing propositional attitudes, i.e., public language utterances, have no direct implications for metaphysical questions related to the nature of non-lingual cognition. Furthermore, several philosophers have proposed rigorous methods for identifying the modes of presentation under which non-lingual agents represent the contents of their beliefs (Bermudez 2003a; Allen 1992). However, there is more to Hutto's point. Not only must there be a mode of presentation under which a believer represents the content of her beliefs, this mode of presentation must have certain formal properties, of a kind that characterize only linguistic vehicles. The reason is that the central role of propositional attitudes consists in the rational guidance of behavior. In order for a propositional attitude to rationally guide behavior, it must combine with other propositional attitudes in rational, practical inference. The desire to catch the squirrel must combine with the belief that the squirrel is in the tree and the belief that barking at it will help flush it out to yield the intention to bark at the tree. But such practical inference is possible only if it involves syntactically articulated modes of presentation. They must consist of components that can recur in different propositional attitudes. For example, in the bout of practical reasoning sketched above, the "squirrel" component of the belief must recur in the desire for the practical inference to work. This suggests that modes of presentation of propositional attitude content must be syntactically constituted, i.e., they must be linguistic.

This is still not sufficient to show that tokening and attributing propositional attitudes presuppose competence in a *public* language, for two reasons. First, as Hutto grants, it is possible that the medium of thought is a kind of language: the so-called "language of thought" (Fodor 1975). According to this hypothesis, both human and non-human cognition consists in computation over mental representations with the syntactic properties of public languages. If

this is true then competence in a public language is not necessary for a cognizer to represent contents under syntactically articulated modes of presentation. So, competence in public language is not required for a cognizer to token full-blown propositional attitudes or attribute them. However, I agree with Hutto (2008, ?), and others (Bermudez 2003a, ?) that the Language of Thought Hypothesis is problematic in a variety of ways. Unfortunately, there is a second, deeper problem with Hutto's argument. It is not clear to me that the concept of belief requires that the modes of presentation under which believers represent contents must be syntactically articulated. It is true that the forms of practical inference studied by philosophers since Aristotle require linguistic vehicles. However, the commonsense concept of belief, like most commonsense concepts, has many dimensions, not all of which are consistent with each other. To many, it seems obvious that individuals incapable of using a language can nonetheless be capable of belief. Such intuitions must be weighed against the philosophical motivations for restricting the class of believers to the class of sentence users. Hutto himself grants that non-linguals are capable of iconic or imagistic takes on situations that they can imaginatively manipulate to facilitate the planning of rational action (2008, ?). Perhaps this does not technically count as practical inference, in the sense in which philosophers have understood this term. However, I see no reason to deny that it comes close enough to qualify as cognition involving beliefs and other propositional attitudes. And, in fact, there are influential philosophical analyses of belief that are explicitly neutral on whether or not its instantiations need be sentential in form (Lewis ?; Stalnaker ?). Thus, Hutto's argument for the claim that only cognizers competent in public language can token propositional attitudes comes up short.

This is still not sufficient to show that the capacity to *attribute* propositional attitudes is independent of competence in public language. Since I have been discussing Davidson and

Hutto, I have followed them in not distinguished strongly between the alleged dependence on public language of, (1) the capacity to *token* propositional attitudes, and (2) the capacity to *attribute* them. However, Bermudez (2003a; 2009), addressing this issue from an entirely different perspective, does distinguish between these. He argues that non-linguals are capable of tokening propositional attitudes yet not attributing them. His argument for the former claim depends on a well-worked out method for attributing determinate propositional contents to the mental states of non-linguals, based on providing adequate explanations of behavior. According to Bermudez (2003a), non-human animals and pre-linguistic infants engage in behavior that can be explained only by positing mental states with determinate propositional contents that combine in "protological" inferences capable of guiding behavior (140-9). This is sufficient to show that they can token beliefs and desires, without competence in a public language, or even in a language of thought (Ibid ?). However, argues Bermudez, such cognitive resources are insufficient to allow for propositional attitude *attribution* (2003a, ?; 2009).

His reasons are the following. Propositional attitude attribution must take place at the "personal" or conscious level, rather than at the "sub-personal" level. The reason is that reasoning about the propositional attitudes of others often plays an important role in a person's practical reasoning, and practical reasoning takes place at the personal/conscious level, not at the sub-personal level (Bermudez 2009, p. 159). Given that propositional attitude attribution takes place at the personal/conscious level, it cannot involve sentences in the language of thought, which is supposed to be the medium of sub-personal cognition (Ibid 163).[3] Bermudez follows consensus in assuming that we are conscious of only two kinds of representational vehicle at the personal level: public language sentences and iconic representations like images or maps. So

[3] Bermudez (2003a, ?) also has reasons for general skepticism about the Language of Thought Hypothesis.

these are the only two candidates for the medium in which propositional attitudes can be attributed. But iconic representations are inadequate to this task because they lack the formal structure necessary to model the practical inferences that lead from beliefs and desires to action (Ibid, pp. 161-2). In contrast, linguistic representations necessarily possess this formal structure. It follows that public language is the only medium capable of supporting propositional attitude attribution (Ibid, pp. 162-3).

I think this argument fails for a number of reasons. First, I do not see why practical reasoning, and hence, in Bermudez's view, propositional attitude attribution need take place at the conscious/personal level. It is true that the conclusions of bouts of practical reasoning are often conscious, i.e., we tend to be aware of the decisions at which we arrive. However, it does not follow from this that the processes that lead to these decisions need take place at the conscious/personal level. In fact, Carruthers (2006, ?) argues that practical reasoning is typically unconscious, and involves multiple practical reasoning processes competing for control of behavior. Since Bermudez's argument for the claim that propositional attitude attribution must take place at the personal/conscious level depends entirely on his claim that practical reasoning must be conscious, the fact that the latter claim is unwarranted undermines the former claim. At the very least, Bermudez owes an argument that practical reasoning must be conscious, and critiques of models like Carruthers', according to which it is not.

Furthermore, there are well worked-out models of sub-personal mechanisms of propositional attitude attribution. For example, Nichols & Stich (2003) argue that sub-personal belief attribution might co-opt a mechanism dedicated to representing counterfactual states of affairs. Representing another agent's discrepant beliefs is a lot like representing states of affairs that are contrary to how the belief attributor takes the world to be. None of this need involve a

public language. Bermudez himself grants that non-human animals and pre-linguistic infants have beliefs and are capable of representing both how the world is and how it could be (2003a, ?). Once this capacity is integrated with other mindreading capacities in the way Nichols & Stich (2003, pp. 93-4) envision, attribution of beliefs independently of competence in public language could be possible.[4]

There is also something strange about Bermudez's claim that the "canonical structure of a proposition is only revealed when propositions are represented in a linguistic format" (2009, p. 162). Recall the reason that Bermudez thinks this matters to propositional attitude attribution: the medium of propositional attitude attribution must have the formal structure necessary to model the inferences that lead from beliefs and desires to actions. As Bermudez puts it, this "is a matter of reasoning about the logical and inferential relations between propositional attitudes. Accurate predictions depend upon the predictor being able in some sense to track the reasoning that the agent might themselves engage in" (Ibid, p. 160). This is strange because, in other work, Bermudez (2003a, pp. 140-9) defends the view that agents incapable of public language engage in a kind of propositional-attitude-involving inference, called "protologic," that differs substantially from the kind of inference made possible by language. It follows that, were language used to model such protologic it would be systematically misleading. For this reason,

[4] Of course, I actually think it is not possible, but for reasons other than Bermudez's. On my view, it simply does not pay to attribute full-blown propositional attitudes in order to predict behavior. There are far less computationally intensive and time consuming ways of succeeding at behavioral prediction, especially if one is part of a community shaped by robust and reliable mindshaping practices. Nichols & Stich's model of belief attribution is not necessarily at odds with this perspective, since it can be used to explain how language users do it, and, on my view, once language is on the scene, full-blown propositional attitude attribution can be useful (more on this below). The key point here is that Bermudez fails to show why propositional attitude attribution *requires a linguistic medium*: Nichols & Stich's model shows that it doesn't. In contrast, on my view, propositional attitude attribution *requires a linguistic motivation* (roughly, the tracking and undertaking of discursive commitments). This is compatible with the possibility that much of the machinery of propositional attitude attribution functions the way Nichols and Stich's model proposes, without necessarily employing a linguistic medium. The products of such machinery can be attributions of discursive commitment, even if the medium employed in generating them does not involve a public language.

the claim that the medium of propositional attitude attribution must have the formal structure necessary to model behavior-guiding inferences does not imply that it must be linguistic. By Bermudez's own lights, some propositional attitudes, i.e., those of agents incapable of public language, guide behavior via protological inferences that differ from language-based inferences. Hence, such practical inferences are not aptly modeled in language. The "off-line" application of protologic to representations of counterfactual situations, in order to model another agent's perspective, as suggested by Nichols & Stich's (2003) model of discrepant belief attribution, seems a much better way of modeling another agent's application of protologic to propositional attitudes.

Thus, Bermudez's reasons for holding that propositional attitude attribution presupposes competence in a public language, like Davidson's, and Hutto's, are inadequate. Still, I agree with this view: I think that full-blown propositional attitude does presuppose competence in a public language. However, my reasons for this are entirely empirical. There is no deep, philosophical connection between language and propositional attitude attribution. Rather, propositional attitude attribution presupposes competence in a public language because there is simply nothing to gain from attributing full-blown propositional attitudes independently of language mastery. As I have argued, human cooperation and other feats of social coordination can be explained in terms of sophisticated mindshaping practices, together with sophisticated application of the intentional stance, i.e., the capacity to parse behavioral sequences into goals and rationally/informationally-constrained means of achieving them. Applying the intentional stance to interpretive targets that have been shaped to reason and behave similarly to interpreters should be enough for interpreters to anticipate their behavior. In such circumstances, there is no

need for concepts of full-blown propositional attitudes, i.e., mental states with tenuous, holistically constrained causal influence on behavior.

Language changes things because it gives us both the means and the need to constantly signal a diverse and open-ended range of commitments, many of which, we might not realize as we signal, conflict with each other. In these circumstances, there arises the need for a practice of keeping track of commitments, and excusing failures to abide by them when we lose track of them. As I argue in the next section, propositional attitude attribution is a powerful mechanism for maintaining or rehabilitating social status in the wake of apparent reneging of publicly expressed commitments. This, and not behavioral prediction based on accurate mindreading, is the central function of full-blown propositional attitude attribution. And it is a function that makes sense only once a complex, commitment-signaling system, like natural language, is on the scene. This is why, in my view, full-blown propositional attitude attribution presupposes competence in a public language.

3. The Rise of Sophisticated Mindreading

Since Chapter One, I have reserved the term "sophisticated mindreading" for the attribution of full-blown propositional attitudes. As I have made clear, full-blown propositional attitudes are, at a minimum, states of a mind – understood as an enduring, unobservable causal nexus hypothesized to explain the behavior of enduring agents – with tenuous, holistically-constrained causal influence on behavior. I take this to be an uncontroversial understanding of full-blown propositional attitudes. It is certainly what most philosophers mean by beliefs, desires and their ilk, and many psychologists have adopted this understanding (Malle et al. 2007, p. 493; Apperly & Butterfill 2009, p. 957). I have also argued that most components of the human socio-

cognitive syndrome require less sophisticated capacities: various forms of mindshaping, complex signaling, and the capacity to adopt the intentional stance.

The intentional stance is a remarkably flexible and efficient tool for anticipating the behavior of rational agents. The reason is that it is focused entirely on behavioral appearances, with no concern for the mental reality behind them. Behavioral sequences and contexts are parsed into goals, rational means of achieving them, and available information. As long as everything goes smoothly and predictions are borne out, there is no need to second-guess interpretations by looking for evidence whether an interpretive target *really* believes, i.e., mentally represents, information an interpreter assumes is available to her, or whether an interpretive target *really* desires the goal an interpreter thinks rationalizes her behavior. If a goal and a set of available information rationalize some behavior, and if this leads to a prediction that is borne out, there is no further evidence required to justify an interpretation from the intentional stance. This is what I mean by the claim that the intentional stance is focused entirely on behavioral appearances, with no concern for an underlying mental reality, i.e., what the target of the interpretation really thinks.[5]

Of course, most interpretive acts that adult humans consciously perform seem to involve more than this. We usually take ourselves to attribute full-blown beliefs and desires, understood as unobservable, causally implicated, states of mind. However, just because many adult humans tend to conceptualize our interpretive practice in this way, does not mean that it is best

[5] One needn't even have a concept of mind – understood as an enduring, unobservable causal nexus that explains behavioral appearances – to adopt the intentional stance. In fact, an interpreter needn't even distinguish between herself and her interpretive targets, or even treat behaviors as products of enduring agents in order to adopt the intentional stance. An interpreter adopting the intentional stance can be restricted to an ontology of disjoint bouts of behavior, with no concern for the enduring agents that produce them. This is not to say that the more sophisticated versions of the intentional stance on which most human beings rely for their quotidian interpretations have such minimal presuppositions. It is likely that human applications of the intentional stance presume that the behavior they are used to interpret issues from an enduring agent that is distinct from the interpreter. However, this is likely an ontogenetic development; it is plausible that the earliest applications of the intentional stance, by infants as young as 6.5 months of age (Csibra 2008), rely on no such presuppositions.

characterized in this way. First, if Dennett is right, then, despite the rhetorical gloss, the beliefs and desires we consciously attribute are nothing but abstract posits that help compress and track observable patterns of behavior (1991). Second, even if Dennett is wrong about this, in previous chapters, I have defended the intentional stance as the best characterization of our low-level, unconscious, automatic interpretive capacities. As Chapter Five argued, this best explains how quickly we arrive at interpretations in dynamic, communicative contexts. As Chapter Six argued, this best explains the interpretive feats of very young infants who are surely not yet capable of sophisticated, conscious, reflective interpretation. Furthermore, as I have noted, non-human animals are capable of rapid interpretation from the intentional stance, though, presumably, they are incapable of the kind of sophisticated, reflective interpretation of which adult humans are (Wood & Hauser 2008). So, when I speak of interpretation from the intentional stance, I have in mind, primarily, what some have called "System-1" social cognition (Carruthers 2009a, ?): our automatic, unconscious, rapid interpretive responses to observed behavior.

I have argued that a System-1 version of the intentional stance is sufficient to explain most components of the human socio-cognitive syndrome. The sophisticated mindreading that I claim is the latest arriving, mindshaping-dependent component of this syndrome, i.e., full-blown propositional attitude attribution, is best understood as a "System-2" capacity for slow, conscious, and reflective interpretation.[6] The question that concerns me here is why such a

[6] This categorization of sophisticated mindreading as a reflective, conscious, slow, System-2 capacity is consistent with influential characterizations of System-2 reasoning. For example, Carruthers (2006, ?), claims that System-2 capacities form the basis for scientific reasoning, i.e., speculation about the unobservable causes of observed phenomena. On the other hand, I do not want to rule out the possibility that some full-blown propositional attitude attribution might be the product of, fast, automatic, unconscious System-1 capacities. The reason is that human beings can often learn to apply sophisticated concepts in an almost "perception-like" way. For example, scientists can just visually recognize instantiations of abstract concepts: a physicist might see a "hooked vapor trail" as the presence of a sub-atomic particle. Similarly, with suitable training in culturally specific signs of mental states, adult human interpreters may learn to perceptually recognize the presence of even full-blown propositional attitudes.

capacity evolved. If what matters most for successful mindshaping, cooperation, coordination, and linguistic communication is anticipating the behavior of one's conspecifics, then why would anything beyond an extremely flexible and efficient way of tracking behavioral patterns be required? Parsing an interpretive target's observable behaviors and their contexts into available information, intuitively obvious goals, and rational means of achieving them in light of the available information should be sufficient to support highly reliable behavioral prediction. Because the notion of rationality at work is so flexible, such interpretative heuristics can very easily be adapted to accommodate predictive failure. An interpreter need only revise her parsing, treating some other behavioral component as the goal, or looking for some difference in information access between her and her target. Such a basic interpretive capacity, when supplemented by mindshaping practices that insure that interpreters and their targets tend to attend to similar information, pursue similar goals, and make similar judgments of rationality, should suffice to explain most human coordinative and communicative feats. If one's main concern in interpretation is the anticipation of behavior, why worry about whether or not one's interpretive targets *really* represent the information one takes to be obviously available, or *really* desire the goals one thinks the context obviously affords? If, as a matter of fact, they are likely to act *as if* they do, due to effective mindshaping, then such shallow interpretations should work well enough for most purposes. There is no need to limn the mental reality behind the behavioral appearance.

According to influential proposals in comparative and developmental psychology, concern with the unobservable reality behind appearances is one of the most important

---

Still, such entrained, reflex-like applications of sophisticated concepts presuppose prior, System-2 capacities, both in science and in quotidian, sophisticated mindreading. In addition, applying the intentional stance needn't always be restricted to System-1 inference. For example, someone who accepts Dennett's understanding of beliefs and desires might reflectively interpret behavior using the intentional stance.

distinctions between human and non-human cognition. For example, Povinelli (?) explains the failure of chimpanzees to learn both social and physical tasks that human beings find trivial in terms of their lack of a distinctively human capacity to understand observable phenomena as effects of unobservable causes. Keil (?) has, over the last three decades, marshaled impressive evidence that human children are default essentialists: they identify and categorize objects in terms of their unobservable essences rather than in terms of the way they appear. For example, to human children, a horse painted to look exactly like a zebra remains a horse (?). Sophisticated mindreading, understood as the attribution of full-blown propositional attitudes, requires the deployment of a similar, essentialistic appearance/reality distinction to the domain of behavior.

Just as an animal's appearance does not fix its kind, or a physical object's behavior does not determine its unobservable causes, an agent's behavior does not determine the mental causes responsible for it. This explains the key difference between the intentional stance and sophisticated mindreading. If a goal and a set of available information rationalizes behavior and enables successful behavioral prediction, then there is no other, deeper fact of the matter of any interest from the intentional stance. But true mindreading must countenance the possibility that the same behavior needn't always issue from the same mental causes. Even a counterfactually robust pattern of behavior, i.e., the same set of responses to a variety of hypothetical circumstances, may issue from radically different mental states. This is why attributing such mental states encounters the holism problem. The mental reality can be entirely independent of behavioral appearance. Why and how would our prehistoric ancestors have developed such an interpretive framework? If their socio-cognitive goals concerned behavioral prediction exclusively then why would they waste time and energy wondering about mental realities that can be entirely independent of behavioral appearance?

Neither Povinelli nor Keil provide any detailed hypotheses regarding the phylogeny of the distinctively human concern with the reality behind appearances. Of course, once this distinction is appreciated and deployed in reasoning about the world, it can be very useful. For example, understanding the true, unobservable causes behind appearances might support creative interventions in the typical course of events, in order to accomplish novel ends. This explains the technological power that science has unleashed.[7] However, this does not explain how or why the distinction first came to be appreciated by our prehistoric ancestors. After all, appreciating that there is an unobservable reality behind appearances does not produce instant practical dividends. One must first develop methods for formulating and verifying accurate models of unobservable reality, and this is a very difficult task. Our species has mastered it only in the last few hundred years. For most of human history and prehistory, our speculations about the unobservable reality behind appearances have been woefully misguided, and supported very few practical dividends. Thus, it is unlikely that there was some kind of instant technological boon that explains how or why our prehistoric ancestors first developed an appreciation of the appearance–reality distinction.[8]

However, it is possible that applying the appearance/reality distinction to human behavior paid immediate, *non-epistemic, social* dividends in the socio-ecology that, I argued in Chapter Four and Chapter Five, likely characterized late prehistoric human populations. Explaining how the concept of a mental reality behind behavioral appearance may have emerged in these

---

[7] E.g., diseases with similar symptoms must often be treated very differently, and such differences can be appreciated only once differences in their unobservable causes are.

[8] Nor is there any reason to expect that an appearance-reality distinction applied to the social domain is immune to the general epistemological challenges of accurately modeling unobservable causes. This is the whole point of the holism problem: it is extraordinarily difficult to quickly and frugally formulate accurate representations of unobservable causes that do not make obvious differences to typical behavioral patterns, whether in the social domain or any other.

circumstances requires a short detour through some more detail about the phylogeny of language, as I understand it. In Chapter Five, I addressed the problem of identifying viable partners for coordination on cooperative projects, given increasing interaction with unfamiliar individuals. I argued that costly signaling of commitment to group endeavors could help solve this problem, especially if individuals disposed toward and competent at coordination on cooperative projects found these signals less costly to produce than other individuals. This would drive the evolution of ever more costly, complex signaling systems, as a way of filtering reliable and competent cooperation partners from less desirable mimics. Chapter Five traced the evolution of the human capacity for structurally complex communication, e.g., recursive language, to such complex signaling systems, which, initially, probably involved the use of rhythmic display in rituals.

Given such a socio-ecology, it is plausible that complex signaling could take on a life of its own. Better signalers would be more trusted, hence engage in more cooperative projects, from which they would gain material benefits, but also a good reputation, and hence higher social status, and better or more mates. This would drive the evolution of even more complex signaling capacities, as these became means to gain status and sexual access, for which humans, like other primates, compete intensely. Furthermore, as the variety of cooperative projects on which individuals could coordinate grew, commitment signaling systems would have to become more complex, to accommodate the growing variety of commitments, i.e., to play different roles in radically different cooperative endeavors. Finally, as I argued in Chapter Five, if, as is likely, this commitment signaling co-existed with an earlier, unstructured, lexical protolanguage, devoted to signaling referential and predicative information about salient objects and events, of the kind hypothesized by Bickerton (1990), these two communicative systems would likely become integrated. Thus, commitment signaling would gradually evolve into something like

contemporary language: a semantically flexible, structurally complex means of signaling intentions related to salient objects and their properties.

On this view, the earliest uses of language were likely something akin to promises, i.e., commitments to courses of behavior involving salient objects and properties.[9] An utterance roughly translatable as "I'll set the trap" is a good example: it expresses a commitment to performing a specific role in a cooperative endeavor, using a structurally complex signal[10] consisting of lexical items that refer to relevant objects and actions. But how do we get from such expressions of commitment to courses of behavior to straightforward assertions of facts? Promises, like "I'll set the trap," have what Searle (1985) calls a "world-to-word" direction of fit: they are fulfilled when the world is made to match the promise. But assertions have what Searle calls a "word-to-world" direction of fit: they are true when the assertion is made to match the world. On my view, the earliest uses of complex language involved a world-to-word direction of fit: they regulated behavior such that it conformed to expressions of commitment. How might such promise making have spawned assertive uses of complex language, involving a word-to-world direction of fit?

It is obvious that one of the most important kinds of roles that individuals can play in coordination on cooperative projects is an epistemic one: ascertaining facts to which other team-members have no access. For example, one's role in a hunt might be to scout out the location and disposition of a herd of prey. This would naturally encourage the evolution of linguistic constructions for reporting facts. Instead of promising just to engage in courses of behavior,

---

[9]See Brandom (1994, 163-5) for a characterization of promising that fits with the commitment-signaling story I have proposed.

[10] Of course, short sentences like this do not *seem* structurally complex, but as Chomsky has taught us, such surface simplicity conceals deep, structural complexity. The structure of any sentence requires complex, hierarchical representations, unlike, say, lists of lexical items.

such constructions could be used to "promise" that the world is a certain way.[11] In promising that the world is a certain way, a speaker is expressing a specific kind of commitment, what Brandom calls a "*doxastic* or *assertional* commitment" (1994, 157). Like the more straightforward kind of commitment expressed in a typical promise, assertional commitments involve commitments to future courses of behavior. Someone who claims that the herd of prey is to the north is thereby committing to act in ways compatible with this fact, and opening herself up to sanction if she does not. But an assertional commitment involves more than just a normative constraint on future behavior. It also opens the asserter to potential sanction if the world is not how she says it is. Just as a promise commits the promiser to behave a certain way on pain of sanction, an assertion commits the asserter to the world's being a certain way on pain of sanction. By committing to the world's being a certain way, e.g., to there being a herd of prey to the north, an asserter entitles her audience to verify her claim and sanction her if it is false, just as a promise entitles an audience to sanction the promiser if she fails to fulfill it.

Thus, it is relatively straightforward to see how an initially rudimentary practice of signaling commitment to play roles in cooperative endeavors may have gradually evolved into the more complex and sophisticated commitment-signaling practice that constitutes contemporary language. Initially, our prehistoric ancestors may have been able to express commitments only to very broad future courses of behavior, associated with specific roles in cooperative endeavors. For example, performance of a mating ritual may have expressed commitment to sexual exclusivity, or performance of a war ritual may have expressed commitment to stand with one's fellows in battle. As cooperative endeavors became more complex and multifarious, driven by group selection for groups capable of more sophisticated

[11] Even in contemporary English, it is possible to emphasize one's commitment to an assertion by saying "I promise, it's the truth!"

cooperation, the commitment signaling practices supporting them would have to become correspondingly more complex and multifarious. As commitment-signaling systems came to match the expressive power of contemporary language, signalers could express commitments as varied and open-ended as the distinctions encoded in contemporary languages. Signalers could express commitments to more than just some broad role in a war party, or sexual exclusivity. They could use the expressive power of language to commit to roles as fine-grained as an "asserter-that-p" where "p" is any proposition encodable in the language.

Such complex and sophisticated commitment signaling would, nevertheless, retain the basic properties of earlier commitment signaling. Its point would remain the same: signaling commitment to and reliability at performing roles in cooperative projects. And it would be supported by the same sorts of normative attitudes as more primitive commitment signaling: signalers who violated the norms governing their signals would be sanctioned thanks to the normative attitudes of typical interlocutors. At the very least such reneging on signaled commitments would result in costs to status. False assertions would jeopardize one's status as a reliable reporter of facts. Unfulfilled promises would jeopardize one's status as a reliable promise-maker. In populations that were highly dependent on smooth coordination on cooperative projects, such costs to status would be far from trivial. More active forms of sanctioning, including gossip, mockery, ostracism, and physical punishment might also be employed. The overall point is that, even as commitment signaling became more complex and sophisticated, and hence supportive of more complex, sophisticated, and varied forms of cooperation, it would continue to rely on the same basic cognitive mechanisms as the earliest forms of commitment signaling, especially the disposition to sanction those who renege on their commitments. Just as an early, prehistoric human would have suffered deep reductions in status

if she reneged on a commitment to sexual exclusivity expressed in a mating ritual, a contemporary human suffers deep reductions in status if she consistently reneges on assertional commitments, e.g., if she is a habitual liar.

Once sophisticated commitment-expressing communication is on the scene, it is possible to understand full-blown propositional attitudes in the way that Brandom does: beliefs that p can be understood as assertional commitments to the claim that p (Brandom 1994, p. 157). How might this perspective shed light on the puzzle of sophisticated mindreading: why our ancestors became concerned with the mental reality behind behavioral appearances? As Dennett has long argued, once members of a population are constantly using a complex, public signaling system to signal commitments of various kinds, this introduces a new method of interpretation (1978, 19, 303-309). Besides adopting the intentional stance and asking what goals and access to information best rationalize an interpretive target's behavior, interpreters can also attend to her explicit expressions of commitment, taking her at her word (Ibid). But, as Dennett notes, the outputs of these two interpretive strategies are not always compatible: the goals and information access that best rationalize and predict a person's behavior may conflict with the goals and claims to which she explicitly expresses commitment.[12] When interpreters are surrounded by interpretive targets that are constantly making discursive commitments of various kinds and, at the same time, engaging in behavior that may or may not be rationalizable in terms of those commitments, interpreters must inevitably grapple with the question: *what do they really think*? In a population that relies exclusively on applying the intentional stance to bouts of behavior, such a question should never arise. If an initial interpretation leads to a false prediction, then it is

[12] Dennett uses the distinction between interpretation based on adopting the intentional stance toward a target's behavior and interpretation based on taking her explicit avowals seriously to address a variety of thorny issues in the philosophy of mind and action, including changing one's mind and weakness of the will (1978, 303-309), and consciousness (1991). More recently, Frankish (2004) has applied this distinction to explaining how System 1 reasoning can come to implement System 2 reasoning via discursive commitment.

withdrawn as incorrect, and replaced with a better one.[13] But in a population of individuals constantly signaling discursive commitments of various kinds, conflicts are not so easily resolved. Individuals can persist in their discursive commitments even as their behavior does not live up to them. In such cases, interpreters are faced with the question of what their targets really think: a distinction between behavioral appearance and mental reality is on the scene.

Is either interpretive strategy – the intentional stance or accepting an interpretive target's explicit commitments as definitive – the ultimate guide to what interpretive targets really think? Dennett seems to favor the intentional stance in this regard: an agent's beliefs and desires are whatever states rationalize and predict her behavior; her explicit avowals reveal only her "opinions" (Dennett 1978, 303-306). However, Dennett's point here is that there is no more fact of the matter regarding what an agent believes and desires than whatever attribution best explains or predicts her *entire* behavioral biography, *including details that are apparent only through scientific investigation* (Ibid 307). When the choice of canonical interpretation must be made between applications of the intentional stance *that are feasible in quotidian contexts*, and the agent's own explicit commitments, then it is not clear which is a better guide to her true beliefs and desires. Interpreters typically do not have as much evidence to go on as their targets, who witness more of their own behavior. An interpretation from the intentional stance that best makes sense of behavior that the interpreter *happens* to have observed might not account for significant behavior that only the target has observed. On the other hand, there is no doubt that

---

[13] Many non-human primates form expectations about each other's behavior, which are sometimes disappointed. However, if Povinelli (?) is right, this has not led to the evolution of an appreciation for the appearance/reality distinction in non-human primate species. Hence, mistaken expectations, alone, are not sufficient for inducing an appreciation of the appearance/reality distinction. This makes sense. When behavioral expectations derived from adopting the intentional stance are disappointed, it is natural to conclude that one has simply made an error in applying the intentional stance: perhaps one has assumed an inaccurate parsing of a bout of behavior, mistaking a goal for a means, for example. But this is not the same as appreciating the appearance-reality distinction presupposed by full-blown propositional attitude attribution. This involves more than the realization that one can be mistaken about behavioral appearances; it requires appreciating the possibility that there are multiple, mutually incompatible, unobservable mental realities that are equally compatible with observed behavior.

people often deceive themselves, and make discursive commitments to which they cannot live up.[14]

In my view, neither third-person, *quotidian* interpretation of behavior from the intentional stance, nor explicit, first-person discursive commitments count as definitive means of discovering an agent's true beliefs and desires. But, as I suggested above, the distinction between behavioral appearance and hidden, mental reality can nonetheless serve an important, non-epistemic social function. Once there are two, potentially conflicting ways of interpreting people, anomalous behavior is inevitable, and the conceit that someone may really think one thing, despite behavioral evidence to the contrary, can play an important role in rehabilitating status in the wake of behavioral anomalies. As I suggested earlier, when an agent's behavior appears anomalous because, for example, it is at odds with explicit commitments she has made, her status as a reliable cooperation partner is at risk. However, if she can somehow excuse the behavior, by appeal to some non-obvious belief or desire that rationalizes the anomaly, she has a tool for rehabilitating status. I think this is the role that the distinction between behavioral appearances and underlying, mental reality played initially, before we had reliable methods to put it to epistemic uses, like uncovering "core cognition" (Carey 2009) and other deep psychological facts about the etiology of behavior. The possibility that a person's behavior might conceal her true thoughts supports the presumption that anomalous behavior can be explained away, e.g., that once we know precisely what a person thought, an apparent reneging

[14] It seems that the choice between trusting third-personal vs. first-personal interpretations as revealing a person's true thoughts is relevant to a deep fault line in the philosophical tradition. Some philosophers are inclined toward skepticism about first-person expressions of doxastic commitment, preferring third-person interpretive frameworks like the intentional stance, due, perhaps, to worries about self-deception. Other philosophers put more credence in first-person expressions of doxastic commitment, perhaps on the grounds that behavior interpretable from the third-person often masks a person's true thoughts.

of a public commitment can be excused. According to Jerome Bruner (1990), this is the primary function of narratives that allude to a person's intentional states, e.g., beliefs and desires:

> … when you encounter an exception to the ordinary, and ask somebody what is happening, the person you ask will virtually always tell a story that contains *reasons* (or some other specification of an intentional state) … All such stories seem to be designed to give the exceptional behavior meaning in a manner that implicates both an intentional state in the protagonist (a belief or desire) and some canonical element in the culture … *The function of the story is to find an intentional state that mitigates or at least makes comprehensible a deviation from a canonical cultural pattern*. (49-50, original emphasis)

Here then is a good candidate for a non-epistemic, social function of distinguishing between behavioral appearances and mental reality that could have paid pragmatic dividends in the socio-ecology that I have argued characterized late prehistoric human populations.

To sum up: In a population employing only the intentional stance to interpret bouts of behavior, it makes no sense to distinguish between behavioral appearance and mental reality. Anomalous behavior should trigger a reapplication of the intentional stance with tweaked parameters, e.g., the interpretive target has a different goal from what the interpreter thought, or has access to different information than the interpreter thought. Basically, there is no conflict between behavioral appearances and mental reality because interpretation is concerned only with behavioral appearances, and some interpretations make sense of these better than others. However, once the capacity for discursive commitment is on the scene, there are two potentially conflicting methods of interpretation available. With Brandom, interpreters can treat their targets' explicit, first-personal, assertional commitments as constitutive of their beliefs, or, with Dennett, interpreters can adopt the third-personal intentional stance toward their targets' overall

behavior – verbal and non-verbal – and attribute goals and information access that best rationalize it. These two approaches inevitably yield conflicting interpretations, since people act in ways that are incompatible with their explicit commitments even as they refuse to give these up. This naturally raises the question of what people *really think*. Even if there are no reliable methods for answering this question, attempts to answer it can nonetheless perform an important, non-epistemic, social function: rehabilitating status through the exculpation of anomalous behavior. When one of our prehistoric ancestors jeopardized her status by doing something counter-normative relative to her explicit commitments, she could rescue her status by coming up with a narrative, appealing to her "real" beliefs and desires, that justified or made sense of the apparently anomalous behavior.

As an example, consider a scout returning to her hunting party to report that there is a large herd of prey to the north. The hunting party proceeds north and finds not a trace of prey. The scout's reliability as a cooperation partner is now in serious jeopardy. However, she can go some way toward rehabilitating it by constructing a Brunerian narrative that appeals to certain non-obvious, intentional states. Perhaps she was travelling at night, got lost, and *believed* that she had been heading north, when she had actually been heading east. The conceit that behavioral appearances might mask an exculpatory mental reality can be used to help mitigate the fallout from apparent reneging on discursive commitments. Given the importance of living up to such commitments in the socio-ecology that I have argued characterized human prehistory, a way of repairing the damage to social status caused by failures to live up to them would have been a very useful, non-epistemic, social function for the concept of a mental reality behind behavioral appearances.

This social function can also explain why the holism of the propositional attitudes is a feature rather than a "bug". In Chapter Three, I argued that holism causes much mischief for propositional attitude attribution considered as a tool for behavioral prediction. The reason is that it jeopardizes the simple links between observable behavior and mental states on which accurate and timely prediction must rely. However, if the original function of full-blown propositional attitude attribution was primarily exculpatory or justificatory, then holism was a feature rather than a bug: if any mental state is compatible with any behavior given adequate adjustments to other mental states, behavior that seems at odds with cooperative intentions or explicit commitments can always be rationalized away by appeal to other, mitigating mental factors of which one's interactants may be unaware. Furthermore, such exculpatory functions require that these other mental factors be treated as having a *causal* influence on behavior: behavior can be excused only by mitigating mental states that actually caused it. In the socio-ecology of our late prehistoric ancestors, the idea that the link between interpretations and behavior is not straightforward, and requires attention to a whole system of reasons for behaving, would already have been in the air, given that individuals sometimes failed to conform to the expectations triggered by their publicly expressed discursive commitments; so, appeal to covert commitments in order to rationalize anomalous behavior would have been an obvious strategy to deflect sanctions. Thus, the notion that there is a mental reality behind behavioral appearances, consisting of states with tenuous, holistically constrained, causal influence on behavior, likely played a very important social function, even if it was hopeless as a tool for behavioral prediction.[15]

---

[15] Obviously, this is a version of Sellars' "myth of Jones" (1963). Discursive practice is used as a model of the unobservable causes of behavior. The difference is that, as David Beisecker once put it to me, Jones was not a scientist. That is, the goal of using discursive practice to model the unobservable causes of behavior is not to

This hypothesis can also explain why the attribution of full-blown propositional attitudes was likely, from the start, a System-2 capacity. Coming up with an excuse for counter-normative behavior that is consistent with all potentially relevant evidence, and likely to convince a skeptical audience is precisely the sort of context-sensitive, "isotropic" task likely to stymie fast and frugal cognition. In addition, since the task involves establishing consistency with verbally expressed commitments, by verbally expressing other, previously unacknowledged commitments, language is already involved, and, on influential theories of Sytem-2 thinking, the domain integration for which it calls is possible only with the help of a public language (Carruthers 2006, ?). As Mercier & Sperber (in press) argue, System-2 reasoning was probably selected to support argumentation with interlocutors, rather than individual reasoning. If the evolutionary scenario sketched in Chapter Four and Chapter Five is on the right rack, then such argumentation often involved determining normative statuses, and hence appropriate sanctioning (if any), on the basis of contested, negotiated interpretations of behavior.

It has long been assumed that full-blown propositional attitude attribution can play both predictive and justificatory roles. Knowing an interpretive target's beliefs and desires can help both predict her behavior and see why it is rationally justified. This is why the attribution of beliefs and desires plays an important role not just in predictive/explanatory projects like psychology, but also in justificatory projects like epistemology and theories of practical rationality. However, philosophers of psychology have tended to stress the former over the latter role (Fodor & Lepore 1993). Beliefs and desires are posited primarily to causally explain behavior, and it just so happens that, because of contingent regularities relating propositional attitudes to each other and behavior, they can sometimes double as rational justifications of

---

produce a true theory of human behavior, but rather, as Bruner (1990) suggests, to provide exculpatory justifications of apparently counter-normative behavior.

behavior. The scenario sketched above reverses this priority. Relative to the socio-ecology of our late prehistoric ancestors, full-blown propositional attitude attribution was more likely to earn its keep playing a justificatory rather than a predictive role. The distinction between behavioral appearances and hidden mental reality was initially used, in the manner described above, to exculpate anomalous behavior, thereby mitigating the threats to status that such behavior triggered.

4. The Role of Propositional Attitude Attribution in Sophisticated Mindshaping

According to the traditional understanding of the relationship between propositional attitude attribution and various forms of mindshaping, the former makes the following contribution to the latter. Before we can shape a mind, through imitation, pedagogy, or normative sanctions, for example, we must first determine the propositional attitudes that animate it. We cannot shape a mind without first knowing the state it is in, and this is the function of propositional attitude attribution. The foregoing has suggested a radically different role for propositional attitude attribution in our mindshaping practices. For example, when it comes to normative sanctions, propositional attitude attribution is part of a negotiated give-and-take aimed at determining the normative status of an interpretive target, and hence whether or which sanctions are appropriate. Apparently anomalous behavior, e.g., the apparent reneging on an explicit commitment, is assumed to be sanctionable, unless the agent can provide an exculpatory narrative, referring to previously unappreciated propositional attitudes. If this interpretation convinces the audience, sanctions can be avoided; however, such interpretations are contestable and hence negotiable (Bruner 1990, p. 47). The model here is something like plea bargaining in courts of law, rather than inferring mental causes in psychology labs.

This raises the worry that there is no fact of the matter regarding what agents really believe and desire – all that matters is the story that works to justify or rationalize or exculpate some behavior, relative to one's audience. But this does not follow. It is important to stress that the target of my description here is our *quotidian* interpretive practice. Everything I say is compatible with the claim that there are facts of the matter regarding what agents believe and desire that careful scientific study can establish. But, as I stressed above, the appearance/reality distinction in general, and the distinction between behavioral appearance and mental reality in particular, are unlikely to play important *epistemic* roles in quotidian contexts because the careful application of the scientific method to the task of uncovering true, unobservable causes is impractical. Hence, we must rely on incomplete data, compatible with multiple interpretations, to negotiate an interpretation that settles the normative status of some agent's anomalous behavior in the eyes of the relevant group. Interpreters and their targets must argue over the relevance and importance that ought to be assigned to a variety of available evidence, including observable behavior, recent, explicit commitments, and the excuses that targets offer. Due to the impossibility of a rigorous, scientific approach to such questions in most quotidian contexts, such arguments are unlikely to identify the true mental causes of an interpretive target's behavior. However, consensus interpretations can be negotiated, and determine whether or which sanctions are appropriate, to the satisfaction of a critical mass in the relevant group. [16]

[16] The distinction between the ontological question of whether or not there are facts of the matter regarding what agents believe or desire, and the pragmatic question regarding the raison d'être of *quotidian* interpretation, helps deflect another potential objection: how can non-human animals have beliefs and desires if they do not participate in the negotiation of interpretations aimed at determining normative status? I have assumed throughout the foregoing that non-human animals have beliefs and desires, whether or not they can attribute them. But my claim here is *not* that only participants in our discursive practices *have* beliefs and desires. Rather, the claim is that only participants in our discursive practices *attribute* beliefs and desires, primarily, in *quotidian* contexts, to help maintain status via justificatory narratives. This does not mean that the attribution of beliefs and desires cannot have other uses, e.g., the epistemic function of accurately representing the psychological causes of human and non-human behavior in scientific explanations. It is entirely possible (and, I claim, actually true) that the practice of attributing beliefs and desires started as a means of justifying behavior, and was later co-opted by scientific psychology as a means of

There is a good deal of evidence from social psychology that behavioral interpretation in terms of propositional attitudes plays such a rationalizing role. For example, this hypothesis best explains a number of asymmetries in spontaneous behavioral explanation by adults, depending on whether the explanations concern our own or others' behavior (Malle et al. 2007). The main focus is the distinction between what Malle et al. call "reason explanations" and "causal history explanations" (2007, p. 494). "Reason explanations" refer to propositional attitudes that, in the agent's mind, lead to and *rationalize* the behavior through a process of deliberation. For example, someone might explain why she wore a hat in terms of her desire to shield her face from the sun. "Causal history explanations" refer to situations, character traits or mental states that lead to behavior, but do not figure in the subject's deliberate decision-making. For example, one might explain the fact that a subject did not vote in an election by claiming that she is lazy. Or one might explain someone's drug or alcohol abuse by reference to a difficult childhood. Such factors do not figure in the deliberate reasoning by which the subjects arrive at their decisions. One does not choose not to vote on the grounds that one is lazy, or choose to abuse alcohol on the grounds that one had a harsh childhood. Malle et al. (2007) found that subjects are more likely to use reason explanations, appealing to propositional attitudes, to explain their own behavior, and more likely to use causal history explanations to explain others' behavior.

What explains such asymmetries? One possibility is that subjects simply know the reasons for which they act better than the reasons for which others act. It is very difficult to accurately gauge the reasons that figure in another person's deliberations, for the sorts of reasons I explored in Chapter Three. Perhaps causal history explanations are simply easier to formulate than reason explanations for subjects other than oneself. However, Malle et al. (2007) controlled

causally explaining both human and non-human behavior in terms of accurate representations of mental states causally responsible for it.

for this possibility. They ran experiments in which subjects were asked to explain the behavior of (1) persons with whom they were intimately acquainted, like friends or family, (2) strangers the behavior of which they personally witnessed, and (3) strangers the behavior of which they merely heard about. Obviously, subjects' knowledge of their explanatory targets varied between these three conditions; however, *the likelihood of offering reason rather than causal history explanations did not*. Subjects were just as likely to offer causal history explanations of the behavior of intimates, the behavior of strangers that they had witnessed, and the behavior of strangers about which they had heard. So, increasing subjects' access to evidence relevant to determining explanatory targets' reasons did not increase the likelihood of reason explanations, suggesting that something else explains the self-other asymmetry between reason and causal history explanations.

Malle et al. (2007) also consider a different hypothesis: reason explanations play a role in "impression management," i.e., they portray an agent's behavior in a favorable light by showing it to be rational. While subjects are typically motivated to portray their own behavior in a favorable light, they are not typically motivated to portray others' behavior in a favorable light. This hypothesis has obvious affinities with the Brunerian account of the function of propositional attitude attribution sketched above. On the view I defend, our ancestors first started attributing full-blown propositional attitudes, understood as the mental reality behind behavioral appearances, in order to rehabilitate status in the wake of apparently counter-normative behavior, especially apparent reneging on explicit commitments. According to Malle et al., behavior explanations serve more than a cognitive function:

> they are [also] a social activity to manage ongoing interactions ... Explanations can be used to clarify, justify, defend, attack, or flatter; they serve as tools to guide and influence

> one's audience's impressions, reactions, and actions … Such impression management can be used from both the actor perspective and the observer perspective, but actors will more often portray themselves in a positive light. Thus, actors' greater use of impression management may help explain at least some of the actor–observer asymmetries. (Malle et al. 2007, p. 504).

If this is true, then the self-other asymmetry in providing reason vs. causal history explanations should disappear if subjects are motivated to portray others' behavior in a positive light. And this is precisely what Malle et al. (2007) found. While manipulating subjects' knowledge of their explanatory targets did not make them more likely to explain their behavior in terms of propositional attitudes, manipulating their motivation to portray even unfamiliar targets' behavior in a favorable light did.

This result favors the mindshaping account of propositional attitude attribution over the mindreading account. If the function of propositional attitude attribution is primarily epistemic, i.e., accurately representing the propositional attitudes that lead to an interpretive target's behavior, then one would expect interpreters with more evidence of an interpretive target's propositional attitudes to be more likely to provide reason explanations. But this is not what Malle et al. (2007) found. Instead, they found that interpreters are more likely to provide reason explanations, *even of the behavior of complete strangers*, when they are motivated to portray them in a positive light. The fact that the availability of evidence relevant to propositional attitude attribution makes *no difference* to the likelihood of providing reason explanations, while the motivation to portray behavior in a good light does, strongly suggests that the purpose of such explanations is not epistemic, but social. In Malle et al.'s (2007) terms, it serves an impression management function; or, in my terms, it functions to preserve or rehabilitate status

in the wake of apparently counter-normative or otherwise puzzling behavior.[17]

Another set of recent, experimental results that are entirely unsurprising from the mindshaping perspective suggests that the attribution of intentional states is influenced by normative judgments (Pettit & Knobe 2009; Knobe 2003; 2006). In the classic study that first showed the so-called "Knobe effect," adult subjects were asked to judge whether or not an interpretive target intentionally allowed a morally significant side effect of a decision she made. For example, subjects read the following two variants of the same vignette and were asked whether or not the protagonist intentionally caused the side effect. In the "help" variant, the vice president of a company proposes a new program to the chairman of the board, saying that it will increase profits and, as a side effect, help the environment. The chairman of the board endorses the program, saying that she cares only about increased profits, and not about helping the environment. In the "harm" variant, the scenario is exactly the same, except the side effect involves harming the environment. Again, the chairman of the board endorses the program, due to the increased profits, and expresses indifference about the harmful effects on the environment. Despite the fact that, intuitively, the chairman of the board in these two scenarios is in exactly the same type of mental state, subjects differed in their assessments of whether or not the

---

[17] Malle et al. (2007) is an extremely rich discussion of a very interesting set of experimental results, and I have touched only the tip of the iceberg here. They also identify two other strong self-other asymmetries in behavior explanation: the tendency to refer to beliefs vs. desires in reason explanations, and the tendency to leave appeals to beliefs unmarked vs. explicitly marked in belief explanations. Their discussions of these asymmetries are also very interesting, and congenial to my project. For example, they explain the self-other asymmetry in belief vs. desire explanations by reference to the difficulty of determining others' beliefs given the idiosyncratic means people have of forming them (495). This fits very nicely with the arguments of Chapter Three. They explain the self-other asymmetry in unmarked vs. marked belief attributions in terms of subjects' desire to distance themselves from the beliefs of others: marking a belief-reason *as a belief* functions to highlight that the attributor does not necessarily endorse it (496). This has clear affinities with Brandom's theory of belief attribution, according to which self-attribution of belief, or, equivalently, doxastic/assertional commitment amounts to *undertaking* or endorsing a claim, while attribution to others does not (Brandom 1994, pp. 161-3). In addition, Malle et al. (2007) make a persuasive case against the traditional characterization of self-other attributional asymmetries in social psychology, according to which subjects explain their own behavior in terms of situational causes and others' behavior in terms of dispositional/person causes. Malle et al.'s (2007) arguments and the experiments on which they draw are a convincing refutation of this traditional framework, and a vindication of their own alternative: "the folk-conceptual theory of behavior explanations".

chairman intentionally helped/harmed the environment. Subjects were significantly more likely to judge that the chairman intentionally harmed the environment in the second scenario, than that she intentionally helped the environment in the first scenario.

There has been vigorous debate about how to interpret this result, especially about whether or not it shows a pervasive influence of normative considerations on our interpretive dispositions. Some argue that the effect is idiosyncratic to judgments of whether or not actions *are intentional* (?). However, drawing on evidence pertaining to other folk psychological judgments, including *having the intention* to help/harm, *intending* to help/harm, *having the desire* to help/harm, *advocating* help/harm, and *being in favor of* help/harm, Pettit & Knobe (2009) argue persuasively that the influence of normative considerations on folk psychological interpretation is pervasive. They also tentatively propose an explanation of this effect in terms of the structure of our representations of what they call "pro-attitudes", i.e., attitudes involving support for some outcome. Pettit & Knobe (2009) are beyond modest about their proposal: they hold out little hope that their explanation will survive future experimental evidence. So the broader, theoretical significance of the Knobe effect is still extremely controversial. However, one aspect of it is not controversial: it is very surprising. I submit that this is an artifact of the received, minreading view of propositional attitude attribution, i.e., that its function is primarily epistemic.

As I have made clear, on the received view, normative sanctions and other forms of mindshaping depend on prior, independent, and accurate mindreading via the attribution of propositional attitudes. We decide whether or not some behavior violates a norm based on a prior assessment of the propositional attitudes – the beliefs, desires, and intentions – that caused it. Given this assumption, it is indeed surprising that the normative assessment of a behavior

should affect our interpretation of it. We are supposed to determine whether or not, for example, the chairman of the board in the foregoing vignettes has violated some norm, partly on the basis of judgments about whether or not she allowed something to happen intentionally, had the desire that it happen, was in favor of it, etc. However, the evidence suggests that folk interpretation does not work this way. Prior judgments about the normative status of the outcomes of the chairman's decision affect our willingness to attribute certain intentions, desires and other pro-attitudes to her. If the mindshaping account is correct, and propositional attitude attribution functions mainly to manage impressions, then the Knobe effect is far less surprising. Of course, in the case of the Knobe effect, the attribution of propositional attitudes functions to justify a *negative* normative status, rather than to rehabilitate or maintain status by providing an exculpatory narrative as a justification of apparently counter-normative behavior. However, the broader point remains: propositional attitude attribution is not some kind of independent prelude to normative judgment. Rather, normative judgment influences propositional attitude attribution.

This conforms to the picture I sketched above: propositional attitude attributions are in the service of justifying prior determinations of normative status. If I am defending myself against an assault on my status triggered by some apparently counter-normative behavior, I will self-attribute propositional attitudes supportive of an exculpatory narrative. If I am enforcing a norm against negligent attitudes toward the environment, I will attribute propositional attitudes that justify my indignation. This is in contrast to the traditional picture, on which, prior to any normative judgment, I use behavioral evidence to determine the facts about the propositional attitudes of relevant persons. [18]

[18] I acknowledge that the story I defend is counter-intuitive. The attribution of propositional attitudes does not *seem* like an exercise in "on the fly" grasping at exculpatory justifications or condemnatory interpretations. It seems like an entirely epistemic activity: ascertaining the true reasons for an agent's behavior, i.e., the mental states that actually caused it. However, it is, by now, a commonplace that we are often very mistaken about the true functions

Again, as I noted above, there is no reason to infer from the picture I advocate that there is no fact of the matter about what people really believe or desire. The point is that the evidence available to *quotidian* interpreters often underdetermines such facts. So there is space for prior normative judgments to nudge interpretations one way or another. Furthermore, nothing that I say implies that we can just make up an interpretation to suit our normative prejudices out of whole cloth. Available evidence, e.g., the chairman of the board telling the vice president that she doesn't care about the effects on the environment, help constrain which interpretations are viable, i.e., which interpretations are likely to convince an audience. However, such evidence still leaves open a choice of interpretations. And, as the Knobe effect demonstrates, prior normative judgments can nudge interpreters toward one of these at the expense of others. The "help" and "harm" variants of the vignette provide exactly the same behavioral evidence of intention; however, given that different propositional attitude attributions are favored in these different circumstances, to the folk, this evidence underdetermines interpretation, and the options can be narrowed further by exclusively normative considerations.

The suggestion here is not that we have the capacity to learn the truth about each other's thoughts, or, perhaps, already know it on some level, and then cynically suppress this in favor of more self-serving interpretations. Rather, I claim only that the inevitably imperfect evidence to which we have access in quotidian contexts leaves interpretation massively underdetermined.

---

of our practices and behaviors. The Nineteenth Century theories of Marx, Darwin, and Freud all imply that the ideological glosses with which we comfort ourselves often conceal unpleasant truths about the raisons d'être of our belief systems. For example, persons may sincerely defend theories of the differences between races or genders on the grounds that their aims are purely epistemic, i.e., that they aim only to uncover the truth behind some phenomenon. Yet the persistence and persuasiveness of such theories may have nothing to do with truth and everything to do with maintaining some normative regime, like slavery or sexist political institutions. Twentieth Century social psychology provides an array of empirical results showing that we are masters at rationalizing self-deception. For example, we concoct rationalizations to explain our preferences for certain products, when objective evidence shows that such preferences are based on completely non-rational factors, like a bias for objects on the right (?). I think that the intuitive plausibility of the traditional view that the attribution of propositional attitudes aims to accurately represent mental causes of behavior is a similar self-deception. We like to think of ourselves as engaged in dispassionate inquiry into the truth about what we think. This is more comforting than the view that we are actually desperately fishing for interpretations that justify our normative intuitions.

When an interpreter looks at all the observable, behavioral evidence to which she has access regarding another's behavior or her own, this does not rule out incompatible interpretations. Instead, the interpreter is faced with something like a "Necker Cube" phenomenon, or Wittgenstein's "Duck-Rabbit". When the evidence is looked at in one way, and some components are foregrounded at the expense of others, one interpretation in terms of propositional attitudes "pops out". But the same evidence can be looked at differently, with different components foregrounded, and a different interpretation will pop out. Such underdetermination of interpretations by evidence is inevitable even under ideal conditions. As Kuhn (1977, 320-339) notes, it pervades theory selection in natural science. Two, equally rational scientists can look at all the same data and arrive at radically different theoretical interpretations because of the different values that they put on different aspects of the evidence, and on various scientific virtues, like simplicity, coherence with the rest of science, breadth of scope, accuracy, and suggestiveness of future research. Given the fact that behavioral data available in quotidian interpretive contexts is far less thorough or carefully gathered than data available in scientific contexts, such problems of indeterminacy are bound to be worse in quotidian interpretive contexts. Hence, it is entirely unsurprising and not even objectionable that, in quotidian contexts where determinate interpretations are required, "impression management" or status preservation will have a role in eliminating the interpretive indeterminacy of inevitably incomplete and imperfect data.

There is also another difference between scientific and quotidian contexts. Kuhn (1977) notes that scientists wedded to a certain paradigm may unconsciously and in good faith engage in selective interpretations of the data in order to support their preconceptions. However, although scientists inevitably emphasize some data at the expense of other data in order to promote their

favored theories, they do not and cannot, typically, make data disappear or manufacture data. When it comes to the quotidian interpretation of behavior, on the other hand, such data manipulation is likely to be routine. The reason is the intimate connection between human interpreters and the objects they interpret, e.g., themselves. If I am wedded to particular theory of myself, then I can do more than just emphasize behavioral data that confirms it over behavioral data that confutes it. I can directly alter my behavior, such that the latter kind of data is systematically extinguished and the former kind of data is systematically promoted. In other words, I can treat self-interpretations as *regulative* (McGeer 2007) rather than epistemic frameworks, thereby turning them into self-fulfilling prophecies.

Although interpreters' control of *others'* behavior is much less direct than this, it is still far more direct than the control that scientists typically have over their domains. Thus, the sorts of automatic mindshaping dispositions discussed in Chapter Two can easily give interpretations a role in the regulation of others' behavior. For example, as Mameli (2001) notes in his discussion of mindshaping, gender-biased interpretations of infant behavior can become self-fulfilling prophecies because children conform to social expectancies: if, for example, their cries are routinely interpreted as expressions of anger rather than sadness, then they will act in ways that conform to these expectations. The process of norm internalization described by Sripada and Stich (2006) is similar. Lower status individuals, like children, can internalize normative construals supplied by higher status members of their culture, and use these to regulate their own behavior. Such phenomena suggest that, in addition to the kind of selective interpretation of data that we find in science, quotidian interpretation allows for the routine, automatic and unconscious *manufacturing* of data that confirms interpretive preconceptions, as well as the routine, automatic and unconscious *extinguishing* of data that violates them.

Given the importance that human beings have always placed on impression management or status maintenance, we are likely to be obsessive rationalizers or concocters of justifications. In self-interpretation, when we find ourselves engaged in some behavior, we are constantly trying out different construals of the imperfect behavioral data to which we have access to arrive at interpretations capable of justifying the behavior in the eyes of our community. Our interpretations of others are also always constrained by judgments of normative status that we want to support. This, I claim, is the central role of narrative understanding in human social life. Cultures afford a relatively small selection of public, justificatory narratives in terms of which their members interpret incomplete, behavioral evidence. The stereotypes encoded in these narratives can then serve a regulative function.

For example, a person might observe that she has become disposed to blush or otherwise become agitated when someone's name is mentioned. Such behavioral evidence underdetermines interpretation in terms of propositional attitudes. But public narratives involving stereotyped characters afforded by her culture are immediately salient: maybe she interprets the behavior as indicating that she has a crush. This interpretation carries with it a whole set of regulative expectations about how the "crush narrative" might unfold. These then feedback into the person's motivational economy, amplifying an initially vague and indeterminate set of behavioral dispositions into an elaborate set of propositional attitudes that a "crush haver" is *supposed* to token, according to her culture. If the object of the crush is familiar with such narratives, and has complementary initial dispositions, interaction is facilitated: both know, roughly, how such narratives are supposed to play out, the kinds of propositional attitudes that are appropriate, and the roles that each is supposed to play. Attributions of propositional attitudes to self and other, in such circumstances, have more to do with what culturally available

narratives indicate characters are *supposed to* believe and desire, than with arriving at true representations of individuals' actual mental states on the basis of careful interpretation of behavioral evidence. Since interactants typically share such normative expectations because they share the same culture and hence similar histories of mindshaping, such collaborative stereotype-confirmation can facilitate coordination.

In Chapter Two, I referred to such sophisticated, language-involving mindshaping as the employment of a "self-constituting narrative". As Ross (2007) argues, such self-constituting narratives shrink the space of "games" or interactions possible for agents familiar with the same narratives. There are culturally sanctioned patterns of thought and behavior that constitute socially constructed roles like "parent", "teacher", "priest", "politician", "businessman", "adolescent", etc. These culturally sanctioned patterns of thought and behavior are an extremely small selection from the space of possible patterns of thought and behavior of which human beings are capable. When we interpret each other's behavior or our own, the input is meager behavioral evidence that is compatible with this larger space of possible interpretations. However, those made salient by mindshaping practices prevalent in our culture are primed, and automatically triggered by behavioral data. They then exert a regulative pressure: initially indeterminate behavioral data is shaped such that it develops into a culturally acceptable pattern. If most of one's potential interactants self-regulate in complementary directions, coordination is dramatically facilitated. In fact, this explains the ease with which propositional attitude attribution can be used to *predict* behavior, despite the holism problem. Because most interactants are shaped to interpret indeterminate behavioral evidence as implicating the same, or complementary, culturally salient, stereotypic narratives, and then use these narratives to regulate further thought and behavior, the attribution of propositional attitudes expected of characters in

such narratives can support reliable behavioral prediction.

As I also noted in Chapter Two, there is some strong behavioral and neural evidence supportive of this picture. Gazzaniga's work with split-brain patients shows that, in most patients, the left hemisphere functions as a kind of self-interpreter: meager data is rationalized in terms of some culturally acceptable narrative. In one experiment, a split-brain patient induced to stand by a command to which only her language-impoverished right hemisphere had access, supplied an on-the-spot, false rationalization when her left hemisphere was asked about her motive: she said that she wanted to retrieve a soda from a nearby house (Gazzaniga 1995, 1393). Furthermore, she then proceeded to act on this confabulated desire: she left the room to retrieve the soda. This illustrates the role that, on my view, sophisticated mindreading plays in sophisticated mindshaping. Whether or not our public self-interpretations are justified or true, we actively work to confirm them (Carruthers 2009a, 127). For the most part, such regulative interpretation is most effective when it is self-directed. In fact, as McGeer (1996) argues, we can understand the apparent *epistemic* authority of self-interpretation in terms of its direct implications for action regulation, rather than in terms of some kind of privileged, Cartesian access to our own motivations. However, third-person regulative interpretation is also possible. If our interpretive targets are lower status and hence defer to our interpretations of their behavior, then they will use such interpretations to regulate their own behavior. This is arguably what happens when children accept and conform to the implications of their parents' interpretations of their behavior (Mameli 2001).

Finally, it is relatively straightforward to see how such self-constituting narratives might come to support epistemic or cognitive self-regulation. Among the narratives afforded by cultures are epistemic narratives, constituting such roles as a "good scientist", or "cogent

reasoner". Given the salience of such narratives in a culture's mindshaping practices, individual members will interpret their own behavior and circumstances in ways that enable epistemic self-regulation. For example, finding oneself disposed to endorse some claim, one can then interpret oneself in terms of a culturally afforded "cogent reasoner" stereotype, and make sure that one also endorses all claims that are logically implied by it, and only claims that are logically compatible with it. Or, finding oneself disposed to explain some phenomenon in a certain way, one can then interpret oneself in terms of a culturally afforded "good scientist" stereotype, and devise ways of testing one's hypothesis. In other words, the same mechanisms that facilitate interaction among individuals familiar with complementary self-regulating narratives can implement so-called "System-2 reasoning" (?). As Frankish (2004) argues, such reasoning is a product of training in culturally and linguistically transmitted epistemic practices - like logic, probability theory, and the scientific method - that avoid the biases of fast, automatic, unconscious System-1 reasoning inherited from our prehistoric ancestors.

On Frankish's view, System-2 reasoning takes place in a "supermind" implemented on the "basic mind" that is a part of our innate, biological endowment. The supermind traffics in "superbeliefs" (Frankish 2004, Chapter Five), understood, roughly, as natural language sentences that we consciously, explicitly accept for use as premises in further reasoning. Frankish's superbeliefs are close kin of Brandom's doxastic commitments: they involve attitudes of commitment to claims formulated in natural language. Clearly, most belief attributions explicitly formulated in natural language are either undertakings (in the case of self-attributions) or attributions of doxastic commitments. If such belief attributions play the role in System-2 reasoning suggested by Frankish, then their mindshaping function is quite obvious: they function to regulate human minds such that they conform to rational and epistemic norms, formulated and

transmitted through cultural and linguistic mechanisms.

5. Final Words

The foregoing has been an examination of two competing stories of the phylogeny of human social cognition. According to the story assumed, I think, by most researchers in the field, human beings came by their distinctive socio-cognitive capacities in the following way. In general, evolution by natural selection rewards cognitive sophistication: individuals of any species who know more about how central features of their ecology work do better than individuals who know less. The most important features of most primate ecologies are social. Hence, in the human lineage, individuals who knew more about how their conspecifics worked did better. In particular, individuals who understood that the behavior of their fellows was caused by unobservable states of mind, and especially propositional attitudes, outcompeted those who did not. Such knowledge enabled them to out-manipulate their dimmer conspecifics. It also enabled better cooperation and coordination, thanks to sophisticated policing against free riders, and sophisticated language designed to reveal independently constituted propositional attitudes.

I have argued against this story, and in favor of the following alternative. Our ancestors' socio-cognitive capacities did not differ significantly from those of other primates; at their best, these capacities amount to highly intelligent behavior tracking, i.e., sophisticated versions of the intentional stance. However, ecological circumstances idiosyncratic to our lineage favored groups with cooperative members, more so than in any other primate species. At first, these groups were small and sparse, and focused on the hunting of megafauna. In such circumstances, cooperation was not the challenge that many take it to be; most behavior was public, and public sharing of information was rewarded. As megafaunal populations dwindled and demographics

and lifeways changed, cooperation and behavioral anticipation became more challenging, because individuals interacted with a greater variety of other individuals, and a division of labor balkanized the earlier information commons. Cooperation was maintained in such circumstances due to group selection: mindshaping mechanisms like conformism, imitation, pedagogy and norm enforcement maintained the intra-group homogeneity and inter-group variation necessary for group selection. Those groups with mindshaping practices that best promoted coordination on cooperative projects outcompeted other groups. Norms supporting plural subject formation evolved, allowing for smoother coordination. None of this was the result of better mindreading: groups with better mindshaping simply did better and produced more individual members. Through time however, the presence of mindshaping practices in such groups may have selected for improved socio-cognitive capacities, like more sophisticated versions of the intentional stance. However, there was no need for full-blown propositional attitude attribution, or for an understanding of the distinction between behavioral appearance and mental reality on which it depends.

As interaction with relative strangers increased due to increasing population density and diversity, complex, ritualistic communication systems evolved as means of signaling commitment to and competence in coordination on cooperative projects. Our capacity for structurally complex language descends from such ritualistic precursors. When integrated with a structure-less, purely lexical protolanguage that probably evolved for other reasons, these ritualistic precursors gave rise to contemporary language, and, in particular, the capacity to make assertional or doxastic commitments. This introduced two, sometimes competing means of interpreting the behavior of our fellows: rationalizing overall behavior from the intentional stance vs. accepting public expressions of doxastic commitment. The conflicting interpretations

that ensued first triggered an appreciation of the appearance/reality distinction applied to human behavior: suddenly it made sense to ask what interpretive targets really thought. This question was answered with explanations that appealed to full-blown propositional attitudes, but the function of such explanations was not to reveal the true causes of behavior; the only reliable method known for discovering true, unobservable causes – science – arrives very late in the history of the species, and is not easily applied in quotidian contexts. Rather, the function of propositional attitude explanations was impression management: the maintenance, diminution or rehabilitation of status, in the wake of apparent reneging on explicit commitments and other kinds of counter-normative behavior.

Obviously, one advantage of the received story over the alternative I defend is simplicity. I can express the former in one short paragraph, but it takes over a page to express the latter. But I think simplicity in matters of human cognition is often deceptive. The reason is that we have concocted simple self-conceptions that serve a mindshaping function. The received view is an example of this: we think of ourselves as perceptive natural psychologists that can somehow ascertain each other's mental states quickly and reliably. But this is more of a regulative ideal that puts pressure on human beings to make themselves more interpretable to each other. The true reality hidden behind such comforting illusions is far more complex. Yet it can better explain a variety of recent empirical results in the cognitive sciences.

Here is a list of empirical facts about human social cognition that I think are puzzling on the first story yet entirely unsurprising on the second story:

- The fact that, of all highly social, highly intelligent primates, only human beings attribute full-blown propositional attitudes.
- The fact that all and only human offspring compulsively overimitate.

- The fact that all and only human offspring interpret stereotyped, adult, communicative behavior as pedagogical.
- The fact that all and only human offspring display an early capacity to acquire norms and enforce them with the help of normative attitudes like resentment, indignation, shame and guilt.
- The fact that human beings tend to find the reliable attribution of higher than first-order propositional attitudes extremely difficult.
- The fact that human beings, in all societies, are willing to pay material costs to punish counter-normative behavior.
- The fact that human beings are more likely to provide reason explanations when motivated to manage impressions, yet not when provided with more evidence of interpretive targets' reasons.
- The "Knobe effect", according to which human beings use independent normative judgments to resolve interpretive indeterminacies.
- The chameleon effect.
- The fact that both human adults and human children are more likely to trust and cooperate with partners with whom they've engaged in varieties of rhythmic synchrony, including singing and marching.
- The fact that human interpreters employ different brain areas to interpret targets perceived as in-group (including themselves) vs. targets perceived as out-group.
- The fact that brain mechanisms that generate error signals responsible for individual, trial-and-error learning in response to failed predictions are also triggered by failures to conform to majority opinion.

This list simply highlights some of the more striking empirical evidence discussed above. It is the tip of the iceberg: as the foregoing discussions suggest, there is an impressive variety of evidence that is naturally explained by the mindshaping hypothesis, yet surprising on the received, mindreading hypothesis. In particular, for reasons reviewed in Chapter Three through Chapter Five, it is difficult to explain the phylogeny of the human socio-cognitive syndrome on the hypothesis that some ancestral hominid population, with socio-cognitive capacities roughly comparable to those of contemporary non-human apes, was invaded by mutants that were capable of full-blown propositional attitude attribution, prior to invasion by mutants capable of human-like mindshaping. As Chapter Three argued, full-blown propositional attitude attribution is unlikely to be reliable and timely enough to make a difference to behavioral anticipation in the absence of prior mindshaping. As Chapter Four argued, it is difficult to see how full-blown propositional attitude attribution could help solve the sorts of coordination problems success at which explains our species' distinctive, evolutionary trajectory; yet, it is easy to see how human-like mindshaping could help solve these sorts of problems.

Although such empirical evidence and arguments support the mindshaping hypothesis, they are, of course, not decisive. As I have stressed in this chapter, empirical evidence can always be reinterpreted in ways that confirm incompatible explanatory hypotheses. I am sure that proponents of the mindreading-first model of the phylogeny of the human socio-cognitive syndrome can interpret the evidence I have reviewed in ways that are compatible with their perspective. This is typical of science: an explanatory framework can often accommodate evidence that, according to other frameworks, definitively refutes it. In light of this inevitable dynamic, it is perhaps best to treat the foregoing as an initial formulation and defense of a heterodox view, aimed at showing that the received view is not the only game in town. Science

flourishes when there are competing, explanatory frameworks, and my principal goal in the foregoing has been to suggest that there is a viable and under-explored framework for explaining the phylogeny of the human socio-cognitive syndrome that is worth developing.

Nonetheless, in closing, let me suggest reasons to favor a somewhat less modest construal of the foregoing. The mindshaping hypothesis has a non-empirical advantage over the mindreading hypothesis that, I think, provides it with a slight edge. Besides explaining a variety of otherwise puzzling empirical facts, the mindshaping hypothesis promises to bring extremely diverse, and seemingly unconnected traditions of human inquiry under a single framework. It is not easy to see what recent, prominent results in comparative psychology, developmental psychology, social psychology, experimental economics, evolutionary game theory, plural subjects theory, hominid paleobiology, evolutionary linguistics and the neural basis of social cognition have in common. However, from the mindshaping perspective, it is possible to discern a common theme: human social cognition relies centrally on our capacities and dispositions to shape each other and ourselves to conform to normative expectations that prevail in groups of likely interactants. The explanatory unification suggested by mindshaping does not end there. According to an influential model of human cognition, our most impressive cognitive feats often involve altering the environments in which we act, to make them easier to negotiate using our limited, internal cognitive resources. Mindshaping is clearly an example of such "epistemic action" (Kirsch ?; Clark 1997) applied to the social domain. The intensity of so-called "tribal instincts" (Richerson & Boyd 2005, 229-30) in driving religiously and ethnically motivated conflict in the contemporary world is also easily explained on the mindshaping hypothesis: we care that our own normative frameworks prevail against others because our abilities to understand and flourish in the social world depend on the integrity of the normative frameworks

that shaped us and continue to structure our social worlds. Furthermore, the growing evidence that language and culture have profound effects on cognitive style (Gumperz & Levinson 1996; Nisbett 2003; Gentner & Goldin-Meadow 2003; Levinson 2003; Thierry et al. 2009) provides further confirmation of the mindshaping hypothesis.

Finally, the notion of mindshaping promises to reveal the mutual relevance of philosophical traditions that sometimes seem antithetical. In this chapter, I have noted how well Brandom's (1994) philosophical theory of discursive practice fits with the empirically motivated phylogenetic story defended in Chapter Four and Chapter Five, as well with recent, empirical results in social psychology. But Brandom's views are often seen as antithetical to the assumptions of mainstream analytic philosophy and, in particular, its focus on a scientifically supported understanding of human cognition (Fodor ?). Brandom himself is a critic of certain "naturalistic" projects in the philosophy of mind (1994, p. ?). From the perspective of the mindshaping hypothesis, this antagonism rests on a fundamental conflation: naturalism, broadly understood as an approach to philosophy constrained by empirical results from science, is conflated with specific, philosophical theories about the place of mind in nature. In particular, one prominent assumption of contemporary naturalism about the mind is that our quotidian interpretive practices are in the same business as the sciences of the mind, i.e., identifying the unobservable causes of behavior. However, it is possible to reject this assumption and remain a naturalist; the mindshaping hypothesis shows how. The best scientific explanation of the phylogeny and persistence of our interpretive practices might require that we understand their roles as primarily justificatory rather than causal/explanatory, and that we appreciate that they presuppose group-relative, normative regimes maintained by the normative attitudes of group members. In short, the mindshaping hypothesis shows how it is possible to be a naturalist about

the mind, while endorsing the kind of understanding of our quotidian, interpretive practices suggested by Brandom's theory of discursive practice.

Indeed, the "ecumenical" potential of the mindshaping hypothesis is even more dramatic than this. Arguably, one of the central differences between the so-called "analytic" and "continental" traditions of western philosophy concerns the status of explanatory frameworks applied to human behavior. Whereas the goals of such frameworks tend to be taken at face value in the analytic tradition, i.e., as aiming to represent mental reality, the continental tradition treats them in a more circumspect way, wary of the normative regimes that they can sometimes unconsciously enforce. For example, Foucault (1961; 1975) is often read as arguing that allegedly universal, scientific truths about human nature "discovered" by science "are, in fact, often mere expressions of ethical and political commitments of a particular society … the outcome of historically contingent forces" (Stanford Encyclopedia of Philosophy 2008). Such claims are obviously in tension with the naturalistic commitments of contemporary analytic philosophy. For example, Foucault would certainly reject the claim that contemporary concepts of the propositional attitudes constitute an innate, biological endowment, as many naturalistic, analytic philosophers assume. The mindshaping hypothesis offers the promise of reconciling these seeming antithetical perspectives. With analytical philosophical naturalism, it accepts that there can be a science of human nature that determines truths about the human mind, like the phylogenetic roots of the human socio-cognitive syndrome and the neural bases of human social cognition. However, on the mindshaping hypothesis, Foucault's understanding of the socio-normative functions of explanatory frameworks applied to human behavior is on the right track. Our quotidian interpretive practices aim to shape our social environments, through a variety of mindshaping mechanisms, including normative attitudes and social institutions based on them,

rather than represent them accurately. For this reason, Foucault is right that we must be constantly vigilant against normative assumptions snuck in as scientifically established, universal truths of human nature.

It is possible to dispute that such "ecumenical" potential is a virtue. Nevertheless, any concept with the potential to uncover an underlying unity among apparently antithetical intellectual traditions, persuasive to thousands of intellectually honest and accomplished scholars, deserves exploration. The notion that our socio-cognitive feats depend on a variety of sophisticated and pervasive mindshaping practices holds such promise.