

folk explanations of behavior:

finding meaning and managing interactions

bertram f. malle

university of oregon

chapter 2

explanations within a folk theory of mind and behavior

Behavior explanations are a fascinating human activity. In fact, they seem to be *two* fascinating human activities. For one thing, they are models people form in their heads to find order and meaning in puzzling social actions or psychological states. For another, they are themselves social acts that can have a number of functions, such as helping others understand the meaning of a behavior, assigning blame or praise, and presenting a certain image of the agent.

But what holds these two distinct aspects of behavior explanations together? How can one and the same phenomenon be a cognitive model and a social act with so many different functions? Part of the answer lies in the conceptual framework within which behavior explanations are embedded—the network of fundamental assumptions people make about human agency, about its relation to the mind, and about its place in the physical world. I refer to this framework as the *folk theory of mind and behavior*, and this chapter describes its major components, functions, and possible origins. I then close with a sketch of how this folk theory shapes behavior explanations.

Briefly, I should remind the reader of the standard attribution position in this matter. According to textbooks and the majority of research publications, people treat human behavior much like any other event: as an effect that is brought about by causes. When explaining behavior, people allegedly classify these causes into two major categories: person and causes. Thus, whenever social perceivers try to explain a behavior, they figure out whether it was primarily caused by the person or by the situation. The way they figure this out, so the standard theory goes, is by searching for covariation patterns—information about the co-occurrence of the behavior in question with (a) the given agent vs. other agents and (b) the given stimulus or context vs. other stimuli or contexts.

Though this standard theory may be valuable in specific domains and under specific conditions (which we will identify in chapters 4 and 5), it does not tell the whole story of behavior explanation. To begin with, standard theory greatly simplifies the conceptual framework in which explanations are embedded. The present approach is committed to an empirical study of ordinary people's assumptions about human mind and action, which guide their explanations of behavior. It will quickly become clear that these assumptions go far beyond standard attribution concepts of cause-effect and person-situation and represent a sophisticated folk model of mind and behavior.

2.1 what a folk theory is (and is not)

2.1.1 labels

People make a number of fundamental assumptions about human behavior and its relation to the mind. These assumptions are interrelated and form a network that is variably referred to as *common-sense psychology*, *naïve theory of action*, *theory of mind*, or *folk psychology* (Churchland, 1991; Heider, 1958; Premack & Woodruff, 1978; Wellman, 1990). One might expect that these different labels refer to different slices of the phenomenon (Whiten, 1994), but currently there is little consensus on what those fine-grained distinctions might be. I will use the term *folk theory of mind and behavior* (and sometimes the short form *theory of mind*) to designate the *conceptual framework that guides people's cognition of behavior and the mind*.

2.1.2 conceptual framework

When I characterize the folk theory of mind as a *conceptual framework* I am referring to a network of concepts (such as *agent*, *intention*, and *reason*) that stand in semantic relation to each other and form a model of the interrelated phenomena in question. These concepts serve as filters and categorization devices in that they selectively respond to certain perceptual input and classify that input as, say, an agent, action, or intention. These classifications activate (or in some cases inhibit) other concepts and then trigger or serve certain psychological processes: prediction, explanation, and evaluation, among others.

Explaining an intentional action, for example, relies on the perceptual classification of movement patterns as *action* performed by an *agent*. This classification normally triggers an immediate search for other stimulus information that may reveal the agent's *intention* (Dittrich & Lea, 1994; Premack & Premack, 1995). It also triggers a search for stored knowledge or reasonable assumptions about the agent's *beliefs*, *desires*, and other mental states pertinent to the action and context. When these presumed mental states can be arranged in a reasoning chain (simulating the agent's own reasoning process), the action is explained by its reasons (Malle, 1999).

In addition to emphasizing what the folk theory of mind and behavior is (a conceptual framework) I should also point out what it is not. First off, it is not just a set of *beliefs* about how the mind works. Beliefs, in the ordinary sense of the word, are acquired through experience,

including indirect experience such as hearsay. But humans do not learn merely through experience that other people have mental states or that there is a basic difference between intentional and unintentional behavior (Fodor, 1992). It even sounds odd to claim that people “believe” that their fellow humans have mental states, for they really couldn’t imagine otherwise. Experience may very well be necessary to practice and refine the application of key concepts of mind and behavior, but experience is not what teaches humans about the concepts themselves, at least not in the way that experience teaches humans a vast store of facts about plants, animals, weather, and the like.

Second, the folk theory of mind is not a set of cultural maxims. Such maxims are rules or obligations that can be broken, but there really isn’t any obvious way in which social perceivers could “break” their assumptions about the mind. Autistic children, who lack certain concepts about the mind, cannot be quite said to break rules either, because telling them that they better adhere to these rules does squarely little to their difficulties in dealing with other minds.

Strictly speaking, theory of mind is not an ability either. The relevant ability is that of inferring mental states, and this ability is made possible by a number of factors, among them perceptual sensitivities, inferential processes, and—a sophisticated conceptual framework. So theory of mind is a requisite component of the mental state inference ability, but it is not identical to it

Finally, theory of mind is not a collection of platitudes. In philosophical discussion, “folk psychology” is sometimes exemplified by platitudes such as “If a person wants O and believes that action A leads to O, she will intend to A” (see Lewis, 1972; Ravenscroft, 1997). But such platitudes presuppose the meaning and distinguishability of concepts such as *want*, *believe* and *intend*. Without these concepts, the platitude is useless (and impossible to acquire).¹ The theory of mind and behavior, and especially the relations among its concepts, might *generate* platitudes (or beliefs, or maxims), but this “theory” is really a conceptual framework.

I should also emphasize that the folk theory of mind and behavior is a part of human social cognition but is certainly not synonymous with it. Sometimes the two terms, *theory of mind* and *social cognition*, are used synonymously (along with *social intelligence*, *Machiavellian intelligence*, and the like), but such an equation would be a mistake (Haslam & Fiske, 2002). Theory of mind as a conceptual framework *influences and supports* a variety of social-cognitive processes. But these processes are phenomena in their own right, and together with a theory of mind they make up the complex web of social cognition. For example, social cognition includes conceptions of relationships, sensitivity to power, formation of categories for groups of people, stereotypes based on easily classifiable features, and an implicit theory of personality. None of these conceptions and processes would be what they are without the framework of mind and behavior, but they are certainly not reducible to that framework.

2.1.3 correspondence to reality

The postulate of a folk theory of mind and its foundational role in social cognition and interaction does not come with a guarantee that this theory always leads to accurate representations of what is “out there”—the objective behavior and mental states of other human beings—or even of what is “in there”—one’s own mental states. It seems highly unlikely, however, that *Homo sapiens* would have evolved this sophisticated conceptual framework without there being a sufficient correspondence between its concepts and the social reality humans try to understand and adapt to.

The emphasis here must be on *sufficient* correspondence between concepts and social reality. We should not, for example, expect correspondence at the level of brain structure, as the folk theory of mind does not imply any claims about neurological architecture (Egan, 1995; Margolis, 1991). Rather, the theory makes “functionalist” claims: It characterizes the phenomena in question by their regular antecedents and consequences as well as by their relations to other phenomena in the same domain. At this functional level, the folk theory of mind corresponds sufficiently well to the reality of human mind and behavior so as to be successful in explanation, prediction, and control of interpersonal behavior. But it has nothing to say about the “constituent nature”—neurological, ontological, or other-logical—of mental states, intentionality, and the like. Nor does it have to. To interpret other people’s behavior or to coordinate one’s preferences and plans with those of others, the folk theory of mind is just at the right level of analysis. In this sense, theory of mind is perhaps like Newtonian physics—highly useful in macro reality, but incomplete when applied to microscopic (neuronal) or large-scale (sociological) processes.

Whatever its precise accuracy as a model of reality, the folk theory of mind has a tremendous impact on social behavior, for without such a conceptual framework people would not grasp the complexity of human action and experience. This impact is something cognitive and social scientists must not ignore, whether or not they imagine beliefs and desires to truly exist in the “objective” mind.

2.1.4 a theory?

There has been some debate over the question whether people’s model of mind and behavior truly warrants the *theory* designation (for a review see Davies & Stone, 1995). As it often happens with such debates, most observers will find that neither of the extreme positions is particularly compelling. It seems difficult to deny that there exists *some* similarity between the folk theory of mind and a scientific theory. Both relate concepts to each other, include general assumptions, postulate unobservables, and serve explanatory and predictive functions. However, equally hard to deny is the fact that there are important differences between folk theories and scientific theories. One of the critical differences is that people can operate in any given domain without a scientific theory, but they could not successfully operate in the domain of human

affairs without a folk theory of mind and behavior. Thus, this folk theory resembles a set of Kantian² categories of social cognition—i.e., the fundamental concepts by which people grasp social reality. Unlike the concepts of a scientific theory, these folk concepts are not formalized in any way and are implicit—that is, people don't normally apply them consciously (Forguson, 1989).

2.1.5 abstract laws or simulation?

Conceiving of the folk theory of mind and behavior as a conceptual framework may help resolve another debate—this one over the specific capacity that underlies people's ascription of mental states (see Carruthers & Smith, 1996; Davies & Stone, 1995).

On one side of the debate we find scholars who have characterized theory of mind as a set of abstract principles or law-like knowledge structure—rather like a scientific theory (Gopnik & Wellman, 1992, 1994; Gopnik & Meltzoff, 1997; see also Ravencroft, 1997). This position, called the “theory theory,” emphasizes that the child's inferences, explanations, and predictions of mental states rely on such principles or laws as “the actions of ourselves and others are linked to internal states” (Gopnik & Meltzoff, 1997, p. 134), “people act to satisfy their desires” (Mitchell, 1997, p. 5), or “If an agent desires x, and sees that x exists, he will do things to get x” (Gopnik & Wellman, 1994, p. 265). When social perceivers make inferences about their own and other people's minds, they apply these and other abstract laws to the specific situation. Importantly, there is no difference between ascriptions of mental states to others and to oneself—both are theoretical inferences grounded in an abstract knowledge structure (Gopnik, 1993).

On the other side of the debate we find scholars who argue that explanations and predictions of others' behaviors rely on a process of “simulation.” That is, social perceivers use their own faculties of perceiving, feeling, and reasoning as models that deliver predictions or explanations about another person's behavior or mental states (Gordon, 1986, 1992; Goldman, 1989, 2001). In the simplest case this process is something like *projection*. The social perceiver assumes that “other = self” with respect to, say, perceptions, beliefs, or preferences, and given that the perceiver has access to his own perceptions, beliefs, etc., ascribing it to others is a straightforward matter. More sophisticated is the attempt to literally simulate the other person's situation and mind states if they differ from one's own. Here too, however, perceivers use their own faculty of deliberation, reasoning, and decision making to deliver, say, an action explanation or prediction, corrected for whatever differences they consider between the other and themselves. A key claim of simulation theory is that mental-state ascription isn't based on theoretical inference, either in the first-person case or the third-person case. People don't infer their own mental states, because they are simply available to them; and they don't infer other people's mental states using abstract laws, because they can more easily project or simulate those states.

Each position has its supporting evidence as well as its specific problems, but what they have in common is that they focus on the psychological mechanism of mental state ascription

more so than on the conceptual framework that underpins it.³ In fact, this conceptual framework is typically presupposed while researchers debate how social perceivers *use* this framework—either to make inferences on the basis of abstract laws or to run simulations on the basis of first-person data. But neither inferences nor simulations are possible without the fundamental concepts that organize perception, reasoning, simulation, and inference. No abstract principle can be acquired or grasped without concepts acting as filters and groupings of perceptual input; and no introspective or simulating process can get off the ground without the prior classification of (one’s own) mental states into such central categories as belief, desire, and emotion.

If we consider theory of mind fundamentally to be a conceptual framework, we are free to allow a variety of psychological processes to do the job of mental state attribution—inference from knowledge structures, projection, conscious or unconscious simulation, introspection, and perhaps several more. Indeed, the research literature suggests that all these processes play a role in the social cognition of mind and behavior (Blakemore & Decety, 2001; Krueger & Clement, 1997; Nickerson, 1999; Ross, Greene, & House, 1977), and often a mixture of them is necessary to solve any given problem. For example, conducting a simulation of a particular person’s mental states in a particular context requires a wealth of cultural, situational, and person-specific knowledge, which includes at least some abstract rules and laws (Wilkerson, 2001). Similarly, abstract principles such as the desire-belief-intention inference rule must be “filled in” with the other person’s presently occurring *contents* of mental states, and this filling-in process may very well rely on projection and simulation (Heal, 1996).

In sum, the debate over the nature of mental state inference loses its edge when we heed the distinction between a conceptual framework of mind and behavior on the one hand and the various cognitive processes and structures that make use of this framework on the other. The implication for a model of folk behavior explanations is clear. When social perceivers offer behavior explanations, they rely (a) on a network of concepts that filter, classify, and organize perceptual input and existing knowledge and (b) on a number of subsequent processes, such as inference and simulation, that deliver an explanatory proposition. We will see later that folk behavior explanations can be grouped into four different modes, and these modes differ both in the concepts that define them and in the kinds of processes that are recruited to produce them. Indeed, the multi-faceted nature of folk behavior explanations may be one of the best arguments for a pluralistic interpretation of theory of mind, embedding both simulation and abstract inference within a mentalistic conceptual framework (Malle, 2001).

2.2 function and dysfunction of theory of mind

I suggested earlier that people would not successfully operate in the domain of human affairs without a folk theory of mind and behavior. Evidence to support this claim is not quite as direct as one would like, because no known people completely lack a theory of mind. But we do have both anecdotal and systematic evidence suggesting that a folk theory of mind frames and

enables complex perception and cognition of human behavior in a way that is just about indispensable. I begin by describing a few cognitive and interpersonal achievements that would not be possible without concepts of mind, and then I examine what happens when at least part of a theory of mind is missing.

2.2.1 achievements of a theory of mind

Consider first the case of a perceiver who notices another person pull out her wallet in front of a cashier. Without a conceptual framework of mind and behavior the perceiver would not understand what the large moving object's encounter with the smaller object means. He would also be rather ineffective at predicting the other large moving object's likely response. With a framework of mind and behavior, however, perceivers can parse this complex scene into fundamental categories of reaching, grasping and exchanging (Baird & Baldwin, 2001; Woodward, Sommerville, & Guajardo, 2001), and after acquiring the pertinent cultural knowledge, they elaborate their interpretation into the script of paying (Schank & Abelson, 1977). People's theory of mind thus frames and interprets perceptions of human behavior in a particular way—as perceptions of agents who can act intentionally and who have feelings, desires, and beliefs that guide their actions (Wellman, 1990; Perner, 1991).

Suppose now you are in the market for an office chair and actually found one you approve of. However, you aren't quite convinced that you will still like it after sitting on it for a whole day, typing away at your computer. So you ask the salesperson whether you could take the chair home with you to try it out for a day or two. The salesperson agrees but asks to take an imprint of your credit card. Why would you agree to that? You reason that he wants some kind of security because he fears there is a chance you might not return the chair without it. You further realize that he thinks just having your address wouldn't suffice (it might be a fake, suing you would be a hassle, etc.), but he assumes a credit card imprint would do because if you don't return the chair, he can charge the purchase price to your card. So the transaction makes sense in light of the salesperson's goals. You also realize that it fulfills your own goals, because you get to try out the chair without giving up any cash (which, you notice, you don't have on you) and you still have the option of not buying the chair. Furthermore, even though the imprint is blank right now, you can be sure that the salesperson won't go mad and charge \$10,000 to your account with the imprint, because he must know he would lose his job and could get sued for fraud, and even if he did go mad, you know that the credit card contract wouldn't require you to pay the \$10,000. Finally, you know that the salesperson knows all that, and presumably knows, too, that you are aware of it. So you jointly realize that this is a fair transaction and go ahead with it.

No doubt, without a theory of mind you would be quite lost in this case. In fact, it is not entirely clear whether, without a theory of mind, there would even be such things as office chairs and credit cards. But granted there are, nobody would agree to this transaction (and many others like it) absent a theory of mind. Neither you nor the salesperson could rely on conditioning from

past history with the other person (because very often there is no past experience), nor could you rely on reciprocal altruism, because there is no guarantee for a future transaction. Plain and simple, you need to understand minds (others' as well as your own) to engage in social transactions and exchanges.

Even more obvious, but no less powerful, is the role of theory of mind in communicative action. Take a speech act such as *promising*, which would be impossible to accomplish without significant considerations of one's own commitment to action, expressed as a public announcement perceived by the other person as that commitment for future action (Astington, 1990, Searle, 1969). And even such seemingly innocuous communicative behaviors as initiating a conversation or taking turns require an appreciation of the other person's attention and intentions at that moment (Clark & Brennan, 1991; Schober). More generally, linguistic behavior is infused with speakers' subtle adjustments to what they assume the listener already knows, doesn't want to hear, or tries to find out, and these adjustments are found at all levels of language—phonetic, morphemic, syntactic, semantic, and pragmatic (Clark, 1997; Schober & Brennan, 2002; Krauss & Fussell, 1991; Sperber & Wilson, 1996).

Finally, and most fundamentally, to communicate something to another person (an "addressee") is an intention to bring about, with the things one says, a certain mind state in the addressee that involves her recognition of that intention.⁴ This may sound complicated, yet we do it all the time. For every utterance spoken, the addressee must make multiple inferences—the intended audience of the speaker's utterance, the referents and meaning of the speaker's words, and the type of social act intended (assertion, question, advice, teasing, etc.). The process of inferring what the speaker "has in mind" is so automated that we don't have to track it consciously—unless it begins to break down. When we ask, for instance, "What do you mean by that?", we signal that we heard the speaker's words but did not recognize the intention behind them, did not recognize which mind state the speaker wanted us to be in upon hearing those words.

2.2.2 theory of mind deficits

In addition to showing some of the achievements made possible by a theory of mind, we can also look at the striking cases in which some parts of that framework are missing. Most widely known in this respect are autistic individuals who have enormous difficulty dealing with other people's mental states (Baron-Cohen, 1995; Frith, 2000; Leslie, 1992). Autistic people are not completely unaware of other minds, but their conceptual understanding of the mental world is severely limited, and as a result they are baffled by the complexity of mind-behavior connections. Often they respond merely to surface behaviors or are not responsive at all, because many of their interaction partners' intentions, thoughts, and sentiments elude them.

The problem, however, does not seem to be one of *perception* of relevant inputs, but primarily one of lacking an *interpretive frame*, resulting in social perception that is strangely raw and mechanical. One autistic person reports⁵:

I know people's faces down to the acne scars on the left corners of their chins and what their eyes do when they speak, and how the hairs of their eyebrows curl, and how their hairlines curve around the tops of their foreheads. [...] The best I can do is start picking up bits of data during my encounter with them because there's not much else I can do. It's pretty tiring, though, and explains something of why social situations are so draining for me. [...] That said, I'm not sure what kind of information about them I'm attempting to process. (Blackburn, Gottschewski, George, & L—., 2000)

What seems to be missing, as another autistic person remarks, is an “automatic processing of ‘people information’.” The data come in, but they cannot be interpreted using concepts of agency and mind. Temple Grandin remarks in one of her illuminating books about living with autism: “I do not read subtle emotional cues. I have had to learn by trial and error what certain gestures and facial expressions meant” (Grandin, 1995, p. 135).

How can one survive in social interactions if emotional cues and social meaning are so elusive? As one discussant put it, “autistic people who are very intelligent may learn to model other people in a more analytical way.” Temple Grandin states that “it was years before I realized that other people are guided by their emotions during most social interactions. For me, the proper behavior during all social interactions had to be learned by intellect” (Grandin, 1995, p. 87). This mechanical, analytical mode of processing, however, is very tiresome and slow: “Given time I may be able to analyze someone in various ways, and seem to get good results, but may not pick up on certain aspects of an interaction until I am obsessing over it hours or days later” (Blackburn et al., 2000). Temple Grandin again:

I had to think about every social interaction. When other students swooned over the Beatles, I called their reaction an ISP—interesting sociological phenomenon. I was a scientist trying to figure out the ways of the natives. (Grandin, 1995, p. 132.)

Thus, many autistic persons in principle seem able to take in the available social information (facial expressions, body movements, etc.), but they lack the *network of concepts* that would allow them to interpret with ease and swiftness the meaning of this information.⁶ As a result, faces, looks, and gestures are merely physical events for autistic persons, and the distinction between persons and objects is largely overlooked. To illustrate, Simon Baron-Cohen (1992) describes the case of Jane, who at one point sat at a lunch table with several people and suddenly climbed onto the table, using other bodies as support, stepping into and knocking over people's food, all in pursuit of grabbing a piece of cake at the other end of the table. “The idea that she could have used words, or gestures, or even eye-contact, to request a piece of cake did not seem to have even entered her mind,” writes Baron-Cohen (1992, p. 13).

Neither before nor after the table incident did Jane engage in any kind of impression management (e.g., “Excuse me,” “Oops,” “I am sorry”) or any behavior explanations (e.g., “I just could not resist this gorgeous piece of cake...”). Because she is oblivious to other people's thoughts and feelings, the need for managing people's impressions and reactions would never occur to her. This obliviousness to others' (unfavorable) impressions not only spoils ongoing

interactions but also stands in the way of generally grasping social conventions such as politeness, etiquette, and other rules of conduct. These conventions protect or rather manage the impressions and reactions of other people, and if one doesn't understand what is there to be protected, conventions make little sense. The only way to fit in with conventional social life is to learn rules by heart and create, as Ms. Grandin does, a library of *if-then* scenarios, all the while remaining hopelessly confused when even small details in an interaction pattern are novel and unlike any pattern previously encountered, learned, and catalogued. In a sense, learning social interaction for autistic persons is like syntax without semantics—learning the grammar of a language without understanding the meaning of its words.

A simple example of obliviousness to social conventions can be found in Oliver Sacks's (1995) description of meeting Temple Grandin for the first time. He arrived at her office, hungry, thirsty, and exhausted after a long day of travel, and was hoping that Ms. Grandin would offer him coffee or something like it. Not aware of any of the bodily and mental states of her visitor, Ms. Grandin immediately started talking about her work and, "with a certain unstoppable impetus and fixity," continued on for a long time, until Sacks finally broke convention (among strangers) and asked directly for a cup of coffee. Lacking the ability to infer other people's thoughts and feelings in context, Temple Grandin can act appropriately only if she can recall a rule from her "video library" of how people behave in different circumstances. "She would play these [videos] over and over again and learn, by degrees, to correlate what she saw, so that she could then predict how people in similar circumstances might act" (Sacks, 1995, p. 260).

But of course it is impossible to memorize predictive rules for every possible situation. Robert Gordon (1992) argued that the impossibility of having comprehensive rules of this sort is significant evidence for the (at least partial) involvement of "simulation processes" in mental state inference. Had Temple Grandin even briefly simulated what it might be like for Oliver Sacks in that situation, she may have offered him some water, coffee, or a snack. Lacking this spontaneous ability to simulate the other mind, she must rely on the catalogue of rules she has acquired from numerous past interactions. But all too often she will lack a rule, or the appropriate rule is simply not triggered by the slightly novel interaction in which she finds herself.

The deficit of theory of mind in autism is striking, but equally striking is the circumscribed nature of that deficit. Mental concepts appear to be the only concepts that are reliably missing among autistic people. Their deficit can therefore not be attributed to some sort of general concept acquisition problem. What then might prevent autistic children from acquiring mental-state concepts?

There are at least three routes toward answering this weighty question. First, researchers are trying to match the brain areas deficient in autistic people with the brain areas possibly involved in mental-state inference (e.g., Fletcher et al., 1995; Gallagher et al. 2000; Happé et al., 1996). Although some evidence points to the involvement of a localized region in the left medial prefrontal cortex, it is too early to draw any strong conclusions, because few laboratories and few

methodologies are currently devoted to this investigation. In addition, there is a potential problem with this search for a particular brain region (harboring the theory of mind “module”), because it assumes that the regions that accomplish mental-state reasoning in the adult brain are also the ones that are disturbed in the autistic infant brain. There is reason to doubt this assumption, because the adult mental-state reasoning system is in all likelihood the result of a complex chain of neuro-cognitive developments, and some of these developments — precursors to a genuine theory of mind — may take place *outside* any adult theory of mind region. (Exactly this situation holds for the development of language in the brain, where regions of infant language development and regions of adult function are nonidentical; Mills, Coffey–Corina, & Neville, 1994.) Thus, the autistic deficit may emerge in an area quite distinct from any (purported) adult region, halting theory of mind development already at a precursor stage.

The second route toward clarifying the theory of mind deficit in autism takes us directly to critical precursors of mental state reasoning. Significant progress has been made in documenting *joint attention* as one such precursor (Dawson et al., 2002; Mundy & Sigman, 1989; Mundy & Neal, 2002). Situations of joint attention involve two people jointly attending to some object or event, with the jointness coming from the mutual realization that the other is attending to the same event. Autistic children have trouble entering such situations of joint attention (Dawson et al., 2002; Mundy, Sigman, & Kasari, 1990), and without the practice that comes from these situations, social referencing, empathy, and coordination of desires become extremely difficult. Furthermore, because joint attention is a requirement for cultural learning in general and language learning in particular (Carpenter & Tomasello, 2000), autistic children’s language deficits may be caused in part by their failure to engage in joint attention. Reduced communicative interaction then also reduces opportunities to learn about others’ minds and coordinate one’s own mind with theirs.

Joint attention is not the only critical precursor to theory of mind development. Closely related are autistic individuals’ lower rates of imitation (Hobson & Lee, 1999; Rogers, 1999) and their reduced responsiveness to faces (e.g., Klin, Jones, Schultz, Volkmar, & Cohen, 2002; see Dawson et al., 2002, for a review). The three capacities — face processing, joint attention, and imitation — are related, as joint attention requires monitoring of gaze and facial expression, and imitation requires joint attention.

Deficits in all three capacities would make the development of a theory of mind extremely difficult. For one thing, the processing of facial information normally reveals valuable clues to mental states (e.g., Baron-Cohen, 1995), such as attention, perception, and interest. Without proper interpretation of faces, those clues are going to be missed.

Furthermore, joint-attention situations permit the monitoring of how self and other differentially respond to the same object, thus potentially launching, or at least practicing, the ability to take another person’s perspective. Impoverished experiences of joint attention would, in turn, hinder the development of perspective taking.

Finally, imitation appears to couple behavioral information with mental information in both self and other. While infants, toddlers, and young children imitate the adult's actions, they may begin to associate their own experience of so acting with the representation of another person acting, thus opening the path for analogical inferences of others' mental states on the basis of their observable behavior and one's own associated mental state (Meltzoff & Brooks, 2001; Goldman, 2001). This form of analogical inference may be facilitated by brain networks that are active both when observing others' actions of a certain type and when planning to or in fact performing this type of action oneself (Blakemore & Decety, 2001; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996; Williams, Whiten, Suddendorf, & Perrett, 2001).

The third path toward understanding the origins of the theory of mind deficit in autism concerns the capability, just mentioned, of registering one's own mental states when they accompany certain actions or expressions and making them available as a model of others' corresponding states when *they* perform those actions or display those expressions. This is a form of introspection but need not be thought of as propositional or even fully conscious (Goldman, 1989; Gordon, 1986). At a minimum, it requires associative linkages of self-experience with self-behavior, of self-behavior with other-behavior, and of self-experience with other-behavior.

Few studies have investigated potential introspective deficits in autistic children, though Frith and Happé (1999) and Raffman (1999) provided arguments for the hypothesis that autism comes with a reduced capacity to either experience or conceptually grasp one's own mental states. Along the same lines, Temple Grandin writes:

My emotions are simpler than those of most people. I don't know what complex emotion in a human relationship is. I only understand simple emotions, such as fear, anger, happiness, and sadness" (p. 89).

Such impoverished self-categorization may set limits on the conceptual categories that are applied to other people. However, the process might also go in the opposite direction (Gopnik, 1993). Because autistic children do not learn mental state concepts in social perception (in instances of face monitoring, joint attention, and imitation), they have an impoverished category system available for their own mental states. Even though the data are not yet available that would distinguish between these two positions, I dare venture a guess that the learning process occurs simultaneously in the introspective and the social-cognitive realms—with any progress in one providing opportunities to refine one's dealings with the other (e.g., Summerville, 2002).

As already indicated, research is beginning to integrate these three routes of investigation, with studies of brain dysfunction being more focused on precursors or components of theory of mind and studies of social cognition incorporating cognition of the self. Another important step would be to delineate the role of affect, and especially the affective interaction context in which

much infant learning occurs, in its relation to theory of mind development in normal and autistic children (Bornstein & Bruner, 1989; Stern, 1985; Kasari, Sigman, Mundy, & Yirmiya, 1990).

2.3 the developing framework of mind and behavior

If the folk theory of mind is a conceptual framework, what concepts does it include and how are they networked together? Work by Tom Shultz and his colleagues (1980, 1988; Shultz & Wells, 1985) highlighted the importance of a concept of intentionality in developing conceptions of behavior. Roy D'Andrade (1987) outlined a folk model of the mind that comprises a small number of mental state types (perceptions, beliefs, feelings, desires, intentions) that differ in their causal origin (outside or inside the mind), their controllability, and in their typical relations to each other (e.g., feelings explain desires explain intentions but not the other way around). Developmental models by Josef Perner (1991) and Henry Wellman (1990) provided a detailed look at fundamental concepts of mental states, especially the representational concepts of belief and desire and their possible origins in the grasp of perception and emotion. Alan Leslie (1994) and Simon Baron-Cohen (1995) discussed what they consider core modules of a developing theory of mind but which one could also interpret as core concepts. The following sketch of the folk theory of mind and behavior draws on this previous work (and the expanding theory of mind literature in general) but also tries to integrate the concepts of *mind* with the concepts of *behavior*. The sketch is set up as a number of conceptual distinctions and their interconnections.

The first distinction is that between *agents* and all other entities in the physical world. Soon after birth, infants show a capacity to imitate human facial movements (Meltzoff & Moore, 1977, 1989), and by three months infants can distinguish human motion from random, nonbiological motion (Bertenthal, 1993). By 9 months of age we see first evidence of children's perceptual sensitivity to self-propelled movement and goal-directed action (Gergely, Nádasdy, Csibra, & Bíró, 1995; Premack, 1990; Wellman & Philips, 2001; Woodward et al., 2001). By one year of age, infants begin to view the goals of others in a more abstract manner, dissociating the individual actions within a sequence from the ultimate goal (Summerville, 2002). Though this goal concept is not mentalistic (that is, no understanding of the mental world is necessary to have the concept), the appreciation of goals to which actions are directed is an important first step toward distinguishing intentional from unintentional behavior.

Another important step is the recognition of agents' gaze (Corkum & Moore, 1998; Johnson, Slaughter, & Carey, 1998; Phillips, Wellman, & Spelke, 2002), which both offers a reliable clue for goal direction and supports the infant's capacity to engage in joint attention — something she cannot do with objects, only *with people toward* objects. By focusing their attention on significant objects and coordinating this attention with another person's attention, children learn to coordinate joint action and interaction. Moreover, the objects of joint attention are anchors in the world akin to meeting places, so we might say that a “meeting of the minds” occurs first in the external world. These objects or anchors are also reference points from which

important deviations can be registered (“now we are both attending, now we are not”), and these deviations can help explain behavior (“when we are not jointly attending, our responses are different”). In fact, it seems possible that in its earliest stages, joint attention is an expansion of the individual mind (“we attend to X”). Only later, discrepancies in the data base are noticed—e.g., memory of attending alone is different from memory of attending *with* another person and different from memory of attending *to* that person. As a result, children begin to differentiate reliably between their own contribution and the other person’s contribution to joint attention and joint action.

With their focused interest in agents, infants get a lot of practice in perceiving intentional action. They may at first have a concept of intentionality that is confounded with the concept of agency (intentional actions are just what agents do). But already at 9 months infants not only differentiate between human hand movements and mechanical movements, but they seem to differentiate between human hand movements that are goal-directed and those that are not (Woodward, 1998, 1999). Thus, goal-directedness is not only a feature that discriminates agents from nonagents but also goes some way toward distinguishing particular agent behaviors (intentional ones) from others (unintentional ones). What helps in this task of distinguishing intentional from unintentional behavior is infants’ ability to detect a meaningful structure in continuous behavior streams—a structure that corresponds to the pattern of intentions the agent executes. Baldwin, Baird, Saylor, and Clark (2001) showed that infants as young as 10-11 months old are more sensitive to interruptions of a behavior stream in the middle of an executed intention than to interruptions at the end of an executed intention, suggesting that they take the complete arc of an intentional action as a natural unit.

Early in the second year of life, children show compelling evidence of classifying intentional and unintentional behaviors into distinct categories (Carpenter, Akhtar, & Tomasello, 1998; Phillips et al., 2002). This distinction relies at first on several perceptual and functional features. Accidental behaviors are those that look uncoordinated, are not directed at significant objects, and are associated with adults’ characteristic expressions of negative affect in face and voice (e.g., “uh-oh”). Intentional behaviors are those that look coordinated, are directed at significant objects, and are associated with adults’ expressions of positive affect in face and voice (e.g., “there!”). By 18 months, the processing of action has become so sophisticated that children can imitate and complete action sequences even if another person performed those actions incompletely or unsuccessfully (Meltzoff, 1995). Thus, children at this age seem to infer the goal or intention inherent in an initiated action and execute this intention themselves. What may underlie this inference process is a sort of self-other matching mechanism that links representations of other-performed actions with plans for self-performed actions (Grèzes & Decety, 2001; Meltzoff & Brooks, 2001). Thus, when observing the adult’s action pattern, the child’s representation of planning that same action becomes activated and is then available for execution.

Both the agency concept and the intentionality concept are initially based on the infant's sensitivity to characteristic features of human behavior, but because these features are in reality associated with certain classes of mental events (e.g., direction of action goals; boundaries between actions intentions; facial expressions emotions), children learn to deal with minds while they are processing behavior. During the preschool years, genuinely mentalistic concepts finally emerge, beginning with the understanding of desires (by age 2) and beliefs (by age 3), concepts that are also reliably used in children's talk about the mind (Brown, 1973) and even explanations of behavior (Bartsch & Wellman, 1989).

At the age of 3, however, children still have difficulty realizing that different people can have different beliefs, and especially that another person might believe something that they themselves do not believe. Believing, at that age, is understood more as copying reality rather than representing reality, and perspective taking — or the awareness of others' mental states as different from one's own — is still very difficult. This all changes at the watershed of theory of mind development around the age of 4 (Perner, 1991; Wellman, 1990), when children acquire a full-fledged belief concept that makes them see beliefs as representations (not copies) of reality, permitting a distinction between what the child believes and what other people believe (cf. Wimmer & Perner, 1983). Aided perhaps by numerous clashes between their own and other people's desires, beliefs, and intentions, children in the preschool years thus learn that different people represent the world in different ways and therefore frequently want, see, and know different things.

With the arrival of genuinely mentalistic concepts, the distinction between intentional and unintentional behaviors is understood in a novel and refined way, namely, as based on characteristic mental states, primarily intention, desire, and belief. At first, the concepts of desire and intention seem to be confounded (Astington, 2001; Moses, 2001), but with the solidification of the belief concept, children recognize that intentions rely on desires (and beliefs) and that desires can stand on their own without necessarily leading to action.

When fully developed (perhaps not before puberty; Kugelmass & Breznitz, 1968), the intentionality concept consists of five components that must all be seen as present for an action to be considered intentional: The agent had a *desire* for an outcome, a *belief* that the action would lead to that outcome, an *intention* to perform the action, the *skill* to perform the action, and *awareness* of fulfilling the intention while performing the action (Malle & Knobe, 1997a). I should emphasize that the actual cognitive process of assessing intentionality often relies on heuristics (e.g., assuming intentionality unless counter-evidence is available) rather than on a five-step decision process. Moreover, even after infancy, perceptual discrimination based on behavioral indicators (e.g., facial expressions, motion pattern) features prominently in judgments of intentionality. The full five-component concept, however, sets the boundaries for any judgment of intentionality and provides the conditions that settle disputes about an action's intentionality.

With the intentionality concept becoming “mentalized,” the idea of self-propelledness turns into a more refined notion of *choice* (Kalish, 1998)—a conceptual advance that is probably aided by the child’s refined capacity for self-regulation (Metcalf & Mischel, 1999; Russell, 1996). Choice captures both the process of forming an intention (of deciding) and its behavioral implementation in an act of trying. In the adult folk theory, choice is seen as the key force in human behavior. Choice is not normally assumed to be present in other objects and events (an assumption from which the philosophical tension derives between freedom of the will and determinism), though choice is sometimes projected onto other events, as in the anthropomorphizing of natural and technological phenomena (e.g., Herbsleb, 1999; Nass & Moon, 2000).

In contrast to the generative power of choice and intentional action, unintentional events are perceived as mere results of other events—in the physical world (e.g., the gale sends the sailor overboard), in the social world (e.g., someone trips over a rock), or in the psychological world (e.g., fatigue impairing my concentration).

Besides the intentional-unintentional distinction, people also make a mind-body or *observable-unobservable* distinction, whose origin lies in the first mental concepts emerging during the preschool years. Perhaps facilitated by increasing self-awareness and introspective abilities, children begin to recognize the correspondence but also discrimination between publicly observable signs and unobservable mental states.⁷ These mental states (which are frequently unintentional; see chapter 3) can have important influences on the agent’s actions and interactions with others, so increasingly fine distinctions are made among such event classes as bodily states, sensations, emotions, and thought processes.

The developing social perceiver does not stop at recognizing certain behaviors as performed intentionally. He also wants to know what the agent’s specific intention or ultimate goal is, and what specific emotions she feels when failing or succeeding in her action. Such contentful mental state inferences require a grasp of the complex interplay between behavioral and situational cues, cultural norms, and the agent’s idiosyncratic attributes (such as preferences and attitudes). From this database, and (sometimes, at least) from the perceiver’s own simulation of what he would feel, think, or do in the given circumstances, specific ascriptions can emerge of the agent’s beliefs, desires, and emotions in the given context. A full appreciation of these contentful mental states also involves an understanding of equifinality (Heider, 1958)—the idea that intentional agents can fulfill a goal in multiple ways, and if they fail one way, they will reason about it and try to pursue it another way.

In addition to the core distinctions of agency, intentionality, observability, and their subsumed concepts of beliefs, desire, intention, emotion, and so on, the conceptual network of mind and behavior has various extensions. Many of them are concepts that are abstracted from single events, such as *attitude* and *value*, presumably derived from more occurrent desires and preferences, and the notion of *personality traits*, presumably derived from patterns of behavior assumed to be caused by characteristic mental states (Ames et al., 2001). These concepts are

used to grasp more temporally stable individual differences among agents and form the basis of person schemas (Kelly, 1955) as well as stereotypes.

2.4 evolutionary origins

In contrast to the increasingly detailed picture of the developing theory of mind in ontogeny, we obviously know much less about the evolutionary path of this powerful social-cognitive tool. A sketch of some reasonable answers, however, is possible, and I will organize it around three main questions:

- I. What were the selective advantages that favored an emerging capacity to represent the mind?
- II. Out of what did the capacity emerge? That is, what were the cognitive precursors of a fully fledged theory of mind?
- III. When did the capacity emerge?

2.4.1 selective advantages

The representation of mental states influences both self-regulation (when one represents one's own mind) and social interaction (when one represents other people's minds), so both of these sets of functions should in principle be candidates for selective advantages. However, the literature has overwhelmingly focused on social functions as the evolutionary advantage of theory of mind. Perhaps this is for good reason, but I delay examining this question until after I have sketched the two main models that delineate the social functions that have favored the emergence of theory of mind: the first is to provide an edge in *social competition* and the second is to aid in *social coordination*.

The first model is perhaps the more radical of the two, subscribing to a *Machiavellian hypothesis* — that mental state inferences support the manipulation of others for selfish gain, rely mainly on deception and counter-deception, and are part of an arsenal of manipulative tactics in a competitive social game (Cummins, 1998; Humphrey, 1976; Krebs & Dawkins, 1984; Whiten & Byrne, 1988).

Cummins (1998), for example, places the emergence of a theory of mind into the context of a dominance hierarchy. Such a hierarchy provides fertile ground for the development of deception, Cummins argues, as lower-ranking individuals would benefit from a capacity to deceive higher-ranking individuals in order to access a greater share of resources (such as food or mating partners). Higher-ranking individuals, on their part, would benefit from a capacity to detect deception and cheating.

The struggle for survival in chimpanzee societies is best characterized as a struggle between dominance and the outwitting of dominance, between recognizing your opponent's intentions and hiding your own. *The evolution of mind emerges from this scene as a strategic arms race in which the weaponry is ever-increasing mental capacity to represent and manipulate internal representations of the minds of others.* (p. 37; italics in the original)

As a modern example of such a struggle, Cummins cites sibling rivalry, in which younger siblings are apt to develop faster than normal their potential to represent the older sibling's mental states. And indeed, studies have shown that children with older siblings pass the false-belief test (the litmus test for theory of mind) at an earlier age on average than do children without older siblings (Jenkins & Astington, 1996; Perner et al., 1994; Ruffman, Perner, Naito, Parkin, & Clements, 1998). However, on closer inspection we find that the data do not strongly support the rivalry hypothesis. Some findings show no relationship of theory of mind indicators with the number of older siblings (Arranz & Olabarrieta, 2002); some show a relationship with number of *any* siblings, even younger ones (Peterson, 2001); and some show a relationship with number of adult kin or peers generally available for interaction (Lewis, Freeman, Kyriakidou, Maridaki-Kassotaki, 1996). In addition, explanations offered for these correlations involve factors that don't fit the dominance theory very well, such as pretend play (Watson, 2000), mental state language (Ruffman et al., 1998), parental encouragement to reflect on other people's feelings (Ruffman, Perner, & Parkin, 1999), and general interaction opportunities (Peterson, 2001). Thus, the overall results suggest a general benefit of "interaction practice" for the development of mental state inference, quite independent of dominance lines.⁸

There are other, more general problems with Machiavellian models. The "arms race" argument implies that modern humans should be highly adept at both deceiving and detecting deception; but the evidence on the detection side clearly does not speak for human excellence in deception (CITE De Paulo). Humans of different cultures are apt to spot and shun "social cheaters" (those who broke or are about to break a social contract; Sugiyama, Tooby, Cosmides, 2002), but such detection does not necessarily rely on a theory of mind and the cheater responsiveness actually suggests that selfish and exploitative social behavior is not productive in the long run, especially not as a community-wide pattern. Recent mathematical simulations by John Orbell and his colleagues confirm this point, suggesting that cooperation and not selfish competition wins out in social communities that initially consist of a mix of "collaborative" and "selfish" individuals (CITE).

A further problem with Machiavellian models is that they don't account for some of the most powerful tools of human evolution: imitation, teaching, communication, and the growth of organized social groups—all phenomena that rely on trust and cooperation, not on selfishness and competition (Givón & Young, 1994). As a corollary, Machiavellian models overlook the fact that theory of mind develops in the first four years of life, a time when children certainly do not primarily compete with their parents or refine their talents of deceiving and detecting deception. Children of this age show enormous trust, attachment, and vulnerability vis-à-vis their parents and other adults, who are their teachers, protectors, and partners in growing up.

In contrast to the relatively dark Machiavellian portrait of humans as engaged in grim social competition, *coordination models* paint the functional story of theory of mind in brighter colors. In these models, humans track other people's thoughts, goals, and emotions to coordinate ongoing interaction (Goody, 1997; Malle, in press; Strum, Forster, & Hutchins, 1997). As Asch

(1952) put it, “We interact with each other . . . via emotions and thoughts that are capable of taking into account the emotions and thoughts of others” (p. 142).

We can break down the coordinative benefits for interaction into at least three elements. First, mental state inferences serve the completion of communal actions, such as group hunting, building shelter, or migrating into new territories. In such joint actions of two or more individuals, sophisticated prior planning, division of labor, and the dynamic updating of one another’s mental states during execution are critical for success.

Second, the ability to empathize with others’ emotions or to correctly guess their desires and beliefs is especially important for hominid child rearing, because human newborns (at least since *homo erectus* 3 million years ago) are far less developed and therefore need more care, protection, and teaching than any other primates. Without a theory of mind, then, human ancestors would not been able to raise adaptively fit offspring.

Finally, mental state inference is a key ingredient in the most powerful cooperative cultural processes: teaching and communication (e.g., Mameli, 2001; Origg & Sperber, 2000; Sperber, 2000). Whenever social learning, linguistic communication, and direct instruction arose during human evolution, the capacity to represent and adjust to others’ mental states must have been in place at that time.

Until recently, coordination models have not been promoted as strongly as their Machiavellian alternatives. Work by Mike Tomasello, Terrance Deacon, and others is beginning to document, however, the many strengths of this hypothesis. Theory of mind, then, appears to be adaptive primarily for its power to facilitate and refine social coordination in communal action, child rearing, and cultural processes.

2.4.2 precursors to an evolving theory of mind

Most scholars assume that the phylogenetic emergence of a theory of mind was a gradual process. The goal then becomes to identify precursors of the full-fledged capacity. A precursor must be primitive enough to operate without mental state concepts or inferences, but it must be sophisticated enough to provide a true launching pad for such concepts and inferences. Without attempting to be exhaustive I focus on five candidate precursors (for background reading see Baron-Cohen, 1995; Gopnik & Meltzoff, 1997; Leslie, 1994; Premack, 1990; others?).

The first is the capacity for imitation, which involves the linking of a representation of another’s behavior to the organism’s own motor program for that same behavior (Blakemore & Decety, 2001). What makes imitation an important precursor of mental state inference is that the linkage of others’ behavior with one’s own behavior can be expanded to a linkage of mental states accompanying own behavior with (thus inferred) mental events accompanying others’ behavior (Meltzoff & Brooks, 2001). This expanded linkage can rely on two different mechanisms.

One mechanism is a noninferential form of empathy, in which the other person’s affect-expressing behaviors (e.g., crying) are imitatively mirrored in the perceiver’s behavior, which in

turn triggers (in reverse direction) an affective state in the perceiver similar to that in the other person (Levenson & Ruef, 1997). To the extent that perceivers can, in this way, “reconstruct” in themselves the mental states that the other person is in, behavioral imitation becomes mental imitation. We might say that the perceiver “resonates” with the other’s affect (Gallese & Goldman, 1998)—a capacity that may be especially important in child rearing.

The expanded linkage between one’s own and other people’s mental states may, alternatively, rely on a primitive form of introspection (far from full-blown self-consciousness). Here, the perceiver registers his own mental states that accompany certain behaviors and replicates or simulates these mental states when observing another person perform the respective behaviors (Goldman, 2001; Gordon, 1992).

In either case, to move from behavioral imitation to mental imitation requires some ability to distinctly register and/or reconstruct one’s own mental states.⁹ This primitive introspective ability should therefore be considered a second precursor of mental state inference.

Third, of critical importance is also the grasp of a person’s *directedness* to an object. An organism that understands directedness observes the reliable orientation of certain body parts (e.g., eyes, hands) toward certain objects and, from these observations, makes predictions about subsequent behavior. The directedness concept thus precedes the more sophisticated mental concepts of attention and goal (Wellman & Phillips, 2001).

The understanding of directedness relies on the prior ability to appropriately parse the behavior stream into intention-relevant units (Baldwin & Baird, 2001), which is the fourth precursor. These units may at first be derived from a spatial frequency analysis of movement (e.g., fast vs. slow, small vs. large, start vs. stop), with no understanding of the units’ meaning. With increasing appreciation of person-object directedness (and aided by repeat viewing), certain movement patterns, such as approach+grasp or look+turn, will become distinct and—with the help of imitative and introspective capacities—meaningful beyond observable pattern recognition.

A final precursor is the capacity for joint attention, which is the recognition that self and another person are both directed at the same object. This recognition relies in all likelihood on eye detection and gaze following (Baron-Cohen, 1995; Butterworth, 1991), and it requires, similar to imitation, a sort of matching between self and other (“I am directed at O and she is directed at O”). The behavior of declarative pointing (“Look, a butterfly”) becomes a powerful means to instigate joint attention, and emotions simultaneously triggered by the object of attention become shared emotions, furthering the practice for empathy.

The emergence of these precursors may well have taken a few million years (MacWhinney, 2002). But once several of them are in place, they build on each other and enable new capacities to evolve. I have already pointed to the supportive role that introspection plays in refining imitation and to the necessary role of grasping directedness for the development of joint attention. Furthermore, joint attention and introspection help differentiate the concept of directedness into subclasses that may launch distinct mental state concepts: pre-action approach

(→ goals), pre-action avoidance (→ fear), post-action success expression (→ joy), and post-action failure expression (→ anger). This differentiated grasp of directedness in turn refines imitation, because the perceiver now becomes sensitive to (and can imitate) more abstract agent-action-object relations, rather than mere physical movement patterns. An appreciation of joint attention and imitation finally facilitates simple forms of teaching, sharing, and other socially coordinated actions.

2.4.3 when did theory of mind emerge?

One approach to narrow down the time frame for the emergence of a theory of mind is to look at successively older milestones in human evolution and ask whether they could possibly be accomplished without a theory of mind. The first such milestone is *Homo sapiens*'s painting of imaginary figures in the caves of Lascaux, about 20,000 years old. To paint such figures, one needs to represent one's own action of painting, one's representation of a (never before seen) creature, and most likely also the responses of community members. Lewis Binford (1981) also credits *Homo* of this age as fully mastering organized hunting, which certainly requires multiple representations of one's own and others' plans, perceptions, thoughts, and intentions, all embedded in a joint (or group) goal.

The next milestone is the great migration period, starting at least 100,000 years ago, when *homo sapiens* left Africa and expanded into Asia, Europe, Australia, and finally the Americas (Cavalli-Sforza, 2000). The level of coordination, planning, mutual trust, and joint action in preparing for and executing an extended migration is inconceivable without a theory of mind. Moreover, the migration to Australia, and probably to southeast Asia, required the use of boats, whose construction, use, and maintenance required social planning and understanding as well as teaching and learning of technology and maritime navigation principles. Cavalli-Sforza (2000) also suggests that human language acquired its modern complexity around 100,000 years ago, which would have been a tremendous tool to use during migration over many generations. What exactly "modern complexity" means is a bit unclear, but it can perhaps be defined as the simultaneous presence of syntactic power to represent complex facts and communicative power to socially transmit these representations to offspring and community members, leading to a hitherto unknown efficiency of teaching and learning. Clearly, these characteristics of language, and the conditions of acquiring it, presuppose an advanced theory of mind.

According to the archeological record there was an even earlier migration period, when *Homo erectus* expanded from Africa into Asia about 1-2 million years ago (Cavalli-Sforza, 2000). If this is correct, theory of mind must be several million years old, as is the social organization in which mentalizing was learned, practiced, and put to cooperative use. One critical archeological finding strengthens this contention (Leahey, 1994).

Skull measurements of the Turkana boy (a young *Homo erectus* from about 1.5 million years ago) suggest that *erectus* infants were born with brains about a third of the adult size because, given the constraints on the diameter of the female pelvis opening that would still

support flexible locomotion, baby brains could not be any larger. This low maturity and helplessness (unparalleled among primates) required more care and social protection, but it also opened the door for more social learning. The infant's brain growth occurred in the context of social and emotional interactions and in exposure to sophisticated behavior patterns, the learning of which literally became part of the child's anatomy. We might also speculate (with Mameli, 2000) that adults may have treated infants as more capable than they really were, expressing expectations that pushed learning forward in each successive generation. This is a purely cultural evolution process, but one with enormous power. Just consider similar phenomena today, when parents expect their children to learn and do and become so much more than genetically indistinguishable children 100-200 years ago. Education is progressive in that each generation, which was expected to learn more than the previous one, expects the next generation to learn even more.

Homo erectus still had much to learn. There is no archeological record older than 1.4 million years of organized tool manufacturing using template replications; no record of rituals, personal decoration, or burying; and little evidence of sustained and sophisticated social organization (Mithen, 1997). Over the next million years, the *Homo* species experienced a last increase in brain volume (from about 900cc to 1350cc), but it may have been the anatomical integration and cultural exploitation of the isolated abilities already available to *homo erectus* that made the difference (Mithen, 1997). Similarly, the reasonably advanced theory of mind in *erectus* may have displayed some intrinsic improvement over this time period, but it was arguably the coalescence of theory of mind, language, and cognitive simulation — all in the context of growing and complexifying social organization — that spawned modern intelligence, social and otherwise (Carruthers & Smith, 1996; Devlin, 2000; Dunbar, 1993; Malle, 2002).

The current archeological record is sparse for the time between 2 and 6 million years ago, so it is difficult to say when the precursors of *Homo erectus*'s theory of mind emerged. Initial research on theory of mind was sparked by the hypothesis that the great apes¹⁰ share with humans the capacity to represent mental states (Premack & Woodruff, 1978; see also Byrne & White, 1988). If so, theory of mind would be at least 6 million years old, which is the time when humans split off from the evolutionary line shared with apes. Increasingly over the last decade, however, theories and evidence have shifted toward the position that genuine theory of mind capacities can only be found in humans and must therefore have evolved some time after the hominid split-off 6 million years ago (Baron-Cohen, 1996; Malle, 2001b; Povinelli, 1996, 2001; Tomasello, 1998). The evidence currently available is incomplete and thus makes any position on this issue tentative. However, the data appear to favor the claim that apes, smart as they are in many respects, do not have genuine mindreading capacities.

To begin with, the type of evidence supporting the claim that apes are able to make mental state inferences consists primarily of field observations, anecdotes, and single-case studies (White & Byrne, 1988; Gomez; Premack & Woodruff, 1978; Savage-Rumbaugh). By

contrast, the evidence against mindreading capacities in apes consists primarily of controlled experiments and some field studies (Povinelli & Eddy, 1996; Tomasello?).

When comparing these findings, one might argue that in the wild apes seem to show more evidence of mental state inference capacities than in the laboratory. But that would not be correct. It isn't the case that the same tests are run in the wild and in the lab and that apes pass them in one context but not in the other. It also isn't the case that the laboratory somehow inhibits intelligent behavior, for some of the most remarkable achievements in ape symbolic communication, imitation, and attention have occurred in laboratory contexts (e.g., Povinelli et al., 1990; Savage-Rumbaugh; Tomasello). Rather, the social behaviors that seem to suggest mindreading capacities in the wild are subject to a number of alternative explanations not involving mental state inference. Under laboratory conditions, by comparison, proper controls can be put in place that isolate genuine mindreading, and in those contexts apes do not show compelling evidence of grasping mental states. They display, instead, two critical capacities that make them socially intelligent without having to employ mental state inferences: refined behavior reading and intelligent learning. Behavior reading is the ability to monitor other organisms' movements, orientations, gazes, and action directions without having to consider mental states. Intelligent learners rely on associative and operant learning but are sensitive to complex stimulus configurations and, with enough trials, can detect statistical relationships between certain behaviors and certain outcomes.

Until recently, the predominant belief was that apes recognize and manipulate mental states whereas monkeys are merely excellent behavior readers (e.g., Cheney & Seyfarth, 1990; Mitchell, 1997; Mithen, 1997; Whiten, 1997). However, Povinelli (2001) examined the primates literature on theory of mind and concluded that apes, too, derive their social intelligence from a refined behavior reading system. Apes' behavior reading is more sophisticated and flexible than the monkey system, but it is still limited in that it lacks the recognition that the mind is the underlying source of observed behavior.

As one example of this limitation, Povinelli and Eddy (1995a) found that, despite their refined practice of locking onto eyes and following gaze, chimpanzees showed no grasp of the mental nature of seeing. In study after study, the apes failed to appreciate the fact that eyes that are covered (e.g., by a bucket or blindfolds) cannot process visual information. Eyes, as well as body posture, are carefully processed as indicators of subsequent events and can therefore become discriminative stimuli; but eyes are not understood as an entrance to the mind

When we apply this model of behavior reading + learning to anecdotes that are highly suggestive of primate theory of mind (e.g., Byrne & Whiten, 1988), we see that a postulate of mindreading is not necessary to account for these findings. For example, chimp **A** holds a banana behind her back until competitor **B** is out of sight and then eats her banana. Or consider chimp **E**, who tries to mate with a female in the vicinity of a higher-status male and covers his erect penis in a way that prevents the higher-status male from seeing it. Behaviors like these are

often interpreted as demonstrating deception and *thereby* demonstrate representations of the other's seeing, wanting, and believing.

There is little doubt that these behaviors are functionally deceptive; but what cognitive mechanism underlies them is far less clear (Hauser, 1997). It isn't difficult to see that the behaviors may well be enabled by behavior reading capacities and intelligent learning. Chimp **A** can accomplish the banana deception by (a) monitoring a conspecific **B**'s body-orientation and field of gaze vis-à-vis **A**, (b) being sensitive to a class of **B**'s orientations that in the past have led to loss of resources, and (c) learning that positioning the banana behind her back is met with the reward of keeping it. Similarly, chimp **E**'s deception requires orientation monitoring and learning that positioning his hand over his penis a certain way leads to positive outcomes.

Just as we don't ascribe mentalizing capacities to animals that play dead, feign injury, or change their appearance upon sighting a predator, we should be cautious in ascribing mentalizing capacities to the deceptive behaviors we see in wild apes. Sophisticated as they are, these behaviors can be accomplished by good behavior-reading and learning skills.

The same caution applies to laboratory findings. For example, Sarah, the chimp tested by Premack and Woodruff (1978), needed 50 learning trials to reliably accomplish a deceptive pointing gesture (misleading a trainer into turning over a cup that did not contain food so that Sarah could keep the food under the other cup to herself). This large number of trials suggests both that imperative gestures (pointing to the cup that has the food that Sarah wants) are deeply ingrained in apes and that extended learning, not mental understanding, may explain Sarah's behavior.

Similarly, Povinelli and colleagues (1992) trained chimps to take one of two roles in a mutually dependent interaction sequence with a human. In the sequence, the "informant" pointed to a tray that visibly contained food. Then the "operator," who could not see the contents of the trays, followed the pointing and made the selected tray available. Finally, both informant and operator shared the culinary reward. Three out of four chimps were easily able to switch roles—that is, they were able to be a proper operator after having been trained as informant or a proper informant after having been trained as an operator. One might be inclined to interpret this finding as demonstrating chimpanzees' ability to read their partner's intentions and then replicate that intention after switching to his role (Mitchell, 1999). However, an alternative account emphasizes the chimpanzees' ability to parse and represent action sequences and to be sensitive to the mutual dependence of these actions in gaining a reward. It appears that the chimps translated an other-action into a self-action, reminiscent of the findings of mirror neurons and suggestive of at least a simple form of imitation. By itself, however, this accomplishment does not provide strong evidence for mental-state inferences (Povinelli 2001).

When we turn away from theory of mind capacities proper and examine some of the precursors of theory of mind, it is obvious that great apes can parse action into intention-relevant units (e.g., when responding to communicative actions; Savage-Rumbaugh et al., 1993), and they are capable of understanding the directedness of actions to objects and individuals (e.g., when

interacting in mutual dependence; Povinelli et al., 1992). But we also find significant limitations of apes' precursor abilities. Uncultured chimpanzees show no reliable skills of joint attention and social referencing (Tomasello, 1996, 2000). That is, they don't point to or show objects to each other (Premack, 1988), and they don't use others' faces as indicators for how they should feel about a new object. Apes' imitative learning abilities are also limited (Hauser, 1996; Smith, 1996; Whiten, 1999), though this assessment is difficult to make given the many different subforms of imitation (Russon, 1997; Whiten & Ham, 1992). Apes' action programs can be "primed" by others' actions, which increases the likelihood that they will perform a similar action, but spontaneous copying of others' behaviors is rare. Also, it seems quite clear that active adult-to-child teaching is virtually nonexistent (Boesch, 1991, 1993), though once again, it depends on what we expect from genuine teaching. At last some apes have put their young into opportunities that facilitate individual learning; but they hardly ever *demonstrate* to their young a sequence of actions (Russon, 1997). We might say that apes show simple forms of imitation and teaching, which can be pushed somewhat further by enculturation (Tomasello). However, we certainly don't see a complete set of theory of mind precursors in our closest primate relatives.

Perhaps the future will bring empirical evidence that could convince skeptics of genuine mental state inference abilities — or at least precursors of theory of mind — in the great apes. For now, I conclude that theory of mind evolved some time between 6 million and 2 million years ago. This is, unfortunately, a large time window and one for which we currently have the sparsest archeological evidence (Leakey, 1994). In consolation, we can be fairly confident that future research in archeology as well as primatology will teach us much more about the evolutionary history of theory of mind.

2.5 behavior explanations within a theory of mind

I have gone into some detail in describing the components, functions, and origins of the folk theory of mind, and I did so for two reasons. First, the conceptual framework of mind is an essential element of human social cognition but has been repeatedly overlooked in social psychological treatments of social cognition (e.g., Augoustinos?; Fiske & Taylor, 1991; Kunda, 1998). For that reason alone it deserves serious attention. Second, the folk theory of mind powerfully directs all thinking about human behavior, and so behavior explanations, too, are under the direction of this folk theory. What, then, uniquely characterizes behavior explanations that are part of a theory of mind?

2.5.1 unique explanations

One possible characteristic of behavior explanations within a theory of mind is that they make reference to *mental causes* (whereas explanations of physical events do not) . This is the position taken by several developmental researchers, who have traced the origin and advancement of behavior explanations throughout the preschool years. According to this position, children as young as 3 systematically use "psychological" or mental-state explanations for human behavior (e.g., Wellman, Hickling, & Schult, 1997). Such psychological explanations

comprise statements that refer to the agent's desires and beliefs but also to moods and lack of knowledge (Schult & Wellman, 1997; Bartsch & Wellman, 1995, chap. 6).

The mental-cause position is certainly correct. Humans not only recognize, just as the great apes (Povinelli, 2000?), causality among physical events, but they also recognize causality among mental and behavioral events; and they can depict these causal relations in sophisticated language. Mental causes set psychological explanations apart from physical explanations, which is one achievement of a theory of mind (Wellman et al., 1997; Schult and Wellman, 1997). That is not all, however, that a theory of mind confers on explanations.

A pure mental-cause model confounds two types of "psychological" causation that are distinct in people's folk theory of mind (Buss, 1978; Heider, 1958; Malle, 1999; Searle, 1983). The first is a version of straightforward "mechanical" causation—one that explains unintentional events by referring to a variety of causes from physical to behavioral to mental. The second is *intentional causation*, which refers to representational mental states (such as beliefs and desires) as *reasons* of an agent's intentional action. Within mechanical causation, mental causes (such as moods, emotions, or wants) explain unintentional behavior in the same mechanical way that physical causes explain physical events. Within intentional causation, however, the agent's reasons explain intentional action in a unique way that presumes reasoning, rationality, and choice on the part of the agent.

As a result of conflating the two mechanisms of causation, current developmental studies on explanation do not tell us whether 3-year-old children who give "psychological explanations" appreciate the difference between mental states as reasons (e.g., "She bought milk because she wanted to make a cake") and mental states as causes (e.g., "She was nervous because she really wanted to win the game"). This is a particularly intriguing question since 3-year olds do seem to distinguish between intentional and unintentional behavior, and we must now wonder whether they have a distinct concept of intentional causation.

The class of behavior explanations that is uniquely guided by the human theory of mind is thus defined not by the type of cause (e.g., mental vs. physical) but by the presumed mechanism of causation (i.e., intentional vs. mechanical). The uniqueness of explanations for intentional behavior is anchored in the folk concept of intentionality. This concept not only identifies certain mental causes that characteristically bring about intentional behaviors (i.e., beliefs, desires, intentions), but it tells us more about the causal mechanism that is uniquely involved in producing intentional action—that of reasoning and choice. Thus, to describe the behavior explanations that inherit unique attributes from a theory of mind we need to take a close look at folk explanations of intentional action.

2.5.2 explanations of intentional action

The multi-faceted folk concept of intentionality provides the frame within which people explain intentional behavior—or, more precisely, the frame for three modes of explanation that clarify distinct aspects of intentional behavior (Malle, 2001). In Chapter 4 I will discuss these

explanation modes in detail; here I only introduce them and position them within the concept of intentionality and, hence, within the folk theory of mind (see Fig. X).

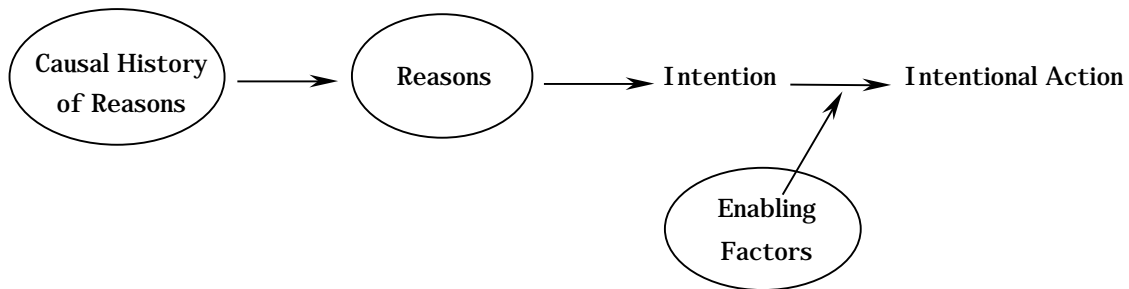


Figure X Three modes of explanation (circled) within the folk concept of intentionality

Working backward from the completed action, we find one explanation mode as concerning the intention-action link, clarifying what factors *enabled* the intention to become a successfully performed action (Malle, et al., 2000; McClure & Hilton, 1997, 1998). Hence, this mode is called *enabling factor explanations* (Malle, 1999; Malle et al., 2000). The second explanation mode concerns the origin of the agent’s intention in her *reasons* — the beliefs, desires, and valuings she considered and deemed to be grounds for adopting the intention in question (Malle, 1999). This mode, then, is called *reason explanations*. The third explanation mode concerns the origin and *causal history* of the agent’s reasons — such as her habits, personality, or unconscious mental processes, and the surrounding context, culture, or biology (Malle, et al., 2000; O’Laughlin & Malle, 2002).

These three explanation modes are tied together by the critical concept of *intention*, the unique mental state of having a reasoned commitment to performing an action (Malle & Knobe, 2001). By contrast, explanations of unintentional behavior are far simpler, as they look for causes that “mechanically” and more or less directly brought about the behavior in question. These causes can be mental or physical, inside or outside the person; but none of them presupposes an intention or a reasoned decision to act (Kalish, 1998).

In sum, classifications of explanations according to broad domains (e.g., psychological vs. physical) provide useful approximations but are insufficient for a detailed account of behavior explanations within the richest of explanatory domains—that of human behavior. Behavior explanations that uniquely derive from the human theory of mind are characterized not so much by their (surface) reference to mental causes but by the conceptual assumptions on which they rely. The folk concept of intentionality codifies these assumptions and provides the conceptual map for asking explanatory questions about human behavior and for offering a variety of distinct explanation modes (reasons, causal histories of reasons, and enabling factors).

In addition, the intentionality concept demarcates as “unintentional” those behaviors that are not relying on the assumptions of reasoning and choice.

In subsequent chapters I will argue that the conceptual framework of mind and behavior is one of two essential determinants of the natural phenomenon of behavior explanations (the other determinant being the set of social functions that explanations serve in communication, impression management, and interpersonal manipulation). My most immediate task is now to outline the fully mature system of behavior explanations among adults. I begin with some basic questions. Why do people form behavior explanations, under what conditions, and which kinds of behaviors do they explain?

endnotes

¹ David Lewis’ (1972) approach was to show formally that if everyday talk about the mind were written as a conjunction of (interrelated) platitudes using mental state terms, each of these terms could then be defined by means of those platitudes, suggesting that everyday talk about the mind is a theory that implicitly defines its key terms. Such an approach is more interesting logically than psychologically because it does not clarify how this kind of theory is represented in social perceivers nor how children come to acquire concepts of mental states, especially at a pre-verbal and thus pre-platitudinal age.

² Kant (1787) postulated a number of categories that the human mind applies to the perception of objects (among them space, time, causality, and substance). These categories, Kant argued, are not just arbitrary frames but the very conditions of the possibility of perception. By analogy, the concepts of a theory of mind would then be the conditions of the possibility of social cognition. But this should not be taken as a logical claim (i.e., that to posit social cognition without a theory of mind would be a formal contradiction); rather, we may say that this framework provides the concepts in terms of which social cognition and interpretation has proven most effective for dealing with other human beings.

³ Wellman (1990, chap. 4) is somewhat of an exception in that he sketches out a network of interconnected concepts that operate like filters in the cognition of human behavior. As a committed theory theorist, however, Wellman insists that the network develops like a scientific theory or “research programme” (Wellman, 1993, p. 18) and that people use this network as a set of laws and abstract principles that aid in action explanation and mental state ascription.

⁴ The addressee’s mind state must involve the recognition of the speaker’s communicative intention or else it is not a communication but merely an act causing *some* mind state in the addressee (e.g., intending to frighten or confuse another person). Communication, according to Grice (1957), Sperber and Wilson (1986), Gibbs (2000), and many others, requires that there is some sort of mutual recognition of the act *as* a communicative act.

⁵ This report and subsequent quotes are extracted from a fascinating discussion of autistic adults who have read the theory of mind literature and try to make sense of their own limitations.

⁶ This is not to say that there aren’t autistic persons with more severe deficits even at the level of information input or processing that go beyond theory of mind deficits (CITE). But my point here is that, even when information input is largely intact, the relevant information cannot be interpreted (cf. Baron-Cohen, 1992).

⁷ This distinction has been expanded in philosophy and all of science to a general dichotomy between measurable, observable appearance and unobservable, underlying reality (e.g., of mathematical

relationships, subatomic forces, and the like)—a dichotomy that we take for granted as characterizing good science (Moravcsik, 1998).

⁸ To reinstate confidence in Cummins' dominance theory we would need to demonstrate that the degree of rivalry between siblings, and especially the degree of dominance exerted by the older sibling, differentially predict precocious theory of mind performance.

⁹ Some pet owners may be certain that their dogs, horses, or other companions empathize with them, as they clearly seem capable of taking on the mood or emotion of their owner. However, we should not underestimate the power of learning by association and reward. Dogs can learn that, under special circumstances—e.g., owner walks slowly, sits with head down, doesn't make sounds—low energy displays prompt more attention and rewards (e.g., petting, hugging) than high energy displays. Neither imitation nor introspection is necessary for this pattern, only a remarkable reading of the owner's behavior to detect the discriminative stimulus.

¹⁰ Apes include orangutans, gorillas, bonobos, and chimpanzees. Together with humans, apes form the class of primates.