Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious.* New York: Oxford University Press.

Chapter 9

Folk Theory of Mind:

Conceptual Foundations of Human Social Cognition

Bertram F. Malle

The ability to represent, conceptualize, and reason about mental states is one of the greatest achievements of human evolution. Having an appreciation for the workings of the mind is considered a prerequisite for natural language acquisition (Baldwin & Tomasello, 1998), strategic social interaction (McCabe, Smith, & LePore, 2000), reflexive thought (Bogdan, 2000), and moral development (Hoffman, 1993). Initial research on representations of mental states was sparked by the hypothesis that apes, too, have such a theory of mind (Premack & Woodruff, 1978), but more recent theories and evidence suggest that the evolutionary emergence of a genuine theory of mind occurred after the hominid split off and may thus be uniquely human (Baron-Cohen, 1999; Malle, 2002; Povinelli, 1996, 2001; Tomasello, 1998).

The ability to reason about mental states has been called a theory of mind because it shares some features with scientific theories (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Wellman, 1990): It postulates unobservables, predicts them from observables, and uses them to explain other observables. This model of theoretical inference is often contrasted with a model according to which people deal with other minds by simulating, in their own minds, what is going on in the other person (Goldman, 1989, 2001; Gordon, 1986; see also Blakemore & Decety, 2001). However, the two approaches are compatible if one regards simulation as one of several processes involved in attributing mental states (another being inference) and if one recognizes that both processes rely crucially on a conceptual framework of mental states and their relation to behavior. I will thus refer with the convenient phrase "theory of mind" to this conceptual framework of mind and behavior, allowing a variety of cognitive processes, such as simulation or inference, to make use of the framework (see Malle, 2001).

In social psychology, considerations of others' mental states have often been treated as a special case of dispositional inference, akin to imputing traits or merely as a precursor to imputing traits (Jones & Davis, 1965; Shaver, 1975). Mental states are comparable to traits in that they are unobservable constructs, but they have a number of unique features. First, mental states are conceptualized in folk psychology as events that actually occur in a distinct domain—that of "minds" or subjective experience; by contrast, the location and nature of traits are left fairly unspecified and abstract. Second, perceivers expect mental states of other agents to be roughly of the same nature as their own mental states and therefore use their own minds to simulate others' mental states, whereas they ~~cannot~~ do not use their own personality to simulate others' traits. Third, and most important, reasoning about mental states is part of a unique and sophisticated conceptual framework that relates different mental states to each other and links them up to behavior (D'Andrade, 1987; Malle & Knobe, 1997a). The nature and elements of this framework of mind and its central functions for social cognition are the topic of this chapter.

Theory of Mind as a Conceptual Framework

A conceptual framework can be regarded as a cognitive capacity that operates prior to any particular conscious or unconscious cognition and provides (by means of classification and process initiation) the framing or interpretation of that cognition. This framing process is unconscious in an interesting way.

Most unconscious processes perform roughly the same functions as do corresponding conscious processes; they just do it more efficiently. (Therein lies the appeal of much modern research on the unconscious, which shows that plenty goes on below the awareness threshold that nevertheless is quite similar to what goes on above the threshold.) But a conceptual framework performs a function that no specific conscious or unconscious process can perform; rather, it is presupposed by these processes.

Take the case of a perceiver who notices another person pull out his wallet in front of a cashier. Without a conceptual framework of mind and behavior, the perceiver would not understand what the larger object's interaction with the smaller object means. She would also be rather ineffective at predicting the other large 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 (Gopnik & Meltzoff, 1997; Perner, 1991; Wellman, 1990).

When this framing and interpretation are lacking, as in the case of autism (Baron-Cohen, 1995; Frith, 2000; Leslie, 1992), the resulting social perception is strangely mechanical and raw. One autistic person (in a fascinating e-mail discussion about theory of mind) reports:

> 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 discussant remarks, is an "automatic processing of 'people information.'" The data come in, but they cannot be interpreted in a parsimonious way using concepts of agency and mind. "Instead, it is all processing abstract concepts and systems—much like computer programs or physical forces" (Blackburn et al., 2000). Or, as one discussant put it, "autistic people who are very intelligent may learn to model other people in a more analytical way." 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).

How is it possible that some people interpret social information so effortlessly while others struggle to find meaning in it? It has been known for some time that human cognition relies heavily on associative structures such as schemas and scripts that simplify encounters with complex stimuli (e.g., Fiske & Taylor, 1991; Schank & Abelson, 1977). But these structures are characterized as a form or process of representation that is so generally applicable that it does not constrain (or code for) the content that it represents. On the level of cognitive organization, then, the schema of a social action such as paying looks just like the schema of a rainstorm brewing.

What is then social about social cognition? The answer usually points to the type and complexity of objects that are at stake—social cognition, in short, is cognition of social objects such as people, relations, groups, and the self (Fiske & Taylor, 1991; Schneider, 1991). But the category of a social object is precisely what general cognitive structures, content-free as they are, cannot easily identify or distinguish from nonsocial objects. How does a general cognitive process "know," as it were, that it deals with another person rather than a lifeless object? (One can easily see the adaptive importance of such a discrimination.) To perform this discrimination fast and efficiently, the human mind appears to rely on a conceptual framework that classifies certain stimuli into basic social categories. Details aside, objects that are self-propelled are classified into the category of agent (Premack, 1990), the coordinated movements of an agent into the category of action (Wellman & Phillips, 2001), and so forth. This category system

develops early in childhood, presumably aided by an innate sensitivity to certain stimulus configurations in streams of behavior (Baldwin & Baird, 1999; Dittrich & Lea, 1994; Gergely, Nádasdy, Csibra, & Bíró, 1995; Woodward et al., 2001; see chapter 10, this volume). Once in place and well practiced, the category system can be activated very easily, as Heider and Simmel (1944) have shown with stimuli as simple as triangles that move around in space, and it can be applied to complex objects such as machines or computers (Dennett, 1987; Nass & Moon, 2000).1

A theory of mind is thus a framework through which certain perceptual input is interpreted or conceptualized as an agent, an intentional action, or a belief; moreover, it frames and directs further processing that is promptly performed on this input (e.g., an inference of the agent's motive for the action). People with a deficient theory of mind, by contrast, might take in all the information that is available (facial features, 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 (see Baron-Cohen, 1992).

If the conceptual framework of mind and behavior, once developed to maturity, is presupposed by any specific conscious or unconscious cognition of human behavior, then this framework resembles Kantian categories of (social) perception—that is, the fundamental concepts by which people grasp social reality. Let me explore this parallel a bit further. Kant (1787/1998) 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 have proven effective for dealing with other human beings.

This view also allows for cases in which these concepts are missing (as in autism, but also in certain forms of frontal brain damage, perhaps in schizophrenia, and in other animals) and for cases in which the concepts have not yet developed (as in very young children). Both types of cases are highly instructive as evidence for the claim that theory of mind is a domain-specific structure or module (e.g., Baron-Cohen, 1995; Hirschfeld & Gelman, 1994; Leslie, 1995; Wellman, 1990). For example, even though autistic children have enormous difficulties with reasoning about mental states, they show average or above-average capabilities in causal reasoning about physical events (Baron-Cohen, Leslie, & Frith, 1985; for reviews see Baron-Cohen, 2000; Leslie, 1992). However, theory of mind is not an isolated module either. Executive control appears to play a role in mental-state reasoning (Carlson, Moses, & Hicks, 1998; Hughes, 1998), and introspection may be involved in this reasoning as well (Goldman, 2001). Moreover, and the capacity for language is linked to theory of mind in both development and evolution (Malle, 2002). For example, a rudimentary appreciation of others' attention focus and communicative intentions is involved in early word learning (Baldwin, 1993), but mastery of certain syntactic structures may be a prerequisite for the realization that beliefs are subjective representations of reality (De Villiers, 2000).

Unfortunately, research has focused primarily on cases in which theory of mind is either missing or not yet fully developed. It appears that the capacities to simulate and reason about mental states are taken for granted among adult social perceivers, and only the absence of this capacity attracts attention among ordinary folk or psychologists. In particular, research on the fundamental assumption that others are agents who act on the basis of mental states is not a central concern of current social psychology, even though several pioneers of the field emphasized its importance. Asch (1952), for example, argued that people "interact with each other … via emotions and thoughts that are capable of taking into account the emotions and thoughts of others" (p. 142). Similarly, Heider (1958) emphasized that "persons have abilities, wishes and sentiments; they can act purposefully, and can perceive or watch us" (p. 21). And Tagiuri, in the foreword to the seminal volume by Tagiuri and Petrullo (1958), proposed to use "the term person perception whenever the perceiver regards the object as having the potential of representation and intentionality" (p. x). Besides work on empathy and perspective taking (e.g., Davis, Conklin, Smith, & Luce, 1996; Ickes, 1997; Krauss, Fussell, & Chen, 1995), contemporary social psychology includes few

investigations into the social perception of mental states. But perhaps this trend is reversing with the growing recognition that mental-state inference is one of the most fundamental tools of social cognition (Ames, in press; Baldwin & Tomasello, 1998; Bogdan, 2000; Graesser, Singer, & Trabasso, 1994; Malle, Moses, & Baldwin, 2001; McCabe et al., 2000; Trabasso & Stein, 1994).

Mind and Behavior

The social-cognitive function of a theory of mind is not just to paint a picture of the mental landscape but to support the understanding of and coordination with other people's behavior, which is achieved by linking behavior to mind (chapter 10). Taking into account the mental states of others helps people make sense of past behavior, permits influence on present behavior, and allows prediction of future behavior. At the same time, reasoning about the mind is grounded in behavioral evidence to maintain reliability and to permit intersubjective discourse about mental states (Bartsch & Wellman, 1995). Without this discourse, mental-state inference would be a private and haphazard endeavor, opening up radical actor-observer asymmetries instead of facilitating human coordination (Wittgenstein, 1953).

The specific connections between mental states and behavior are usually of two forms: mental states that find their expression in behavior (such anger shown in the face) and mental states that guide or influence behavior (such as an intention to act). Significantly, behavior that is connected to mental states breaks down into two fundamentally different types (Heider, 1958): intentional action, which is caused by the agent's intention and decision; and unintentional behavior, which can be caused by internal or external events without the intervention of the agent's decision. This distinction is one of the most influential and illuminating concepts of the folk theory of mind (Malle et al., 2001).

Intentionality

Intentionality is a complex folk concept that specifies under what conditions people judge a behavior as intentional (or done on purpose). This judgment relies on (at least) five conditions (Malle & Knobe, 1997a; Mele, 2001): An action is considered intentional when the agent had (1) a desire for an outcome, (2) a belief that the action would lead to that outcome, (3) an intention to perform the action, (4) the skill to perform the action, and (5) awareness of fulfilling the intention while performing the action. Of course, the cognitive process of assessing intentionality often relies on cues and heuristics rather than on a five-step decision process (e.g., Knobe, 2003). However, the folk concept sets the boundaries for any judgment of intentionality and provides the conditions that settle disputes about an action's intentionality.

Some of the individual components of the intentionality concept are themselves powerful tools for making sense of behavior. For example, people differentiate between two motivational states, desire and intention, when explaining, predicting, and influencing behavior. The two states differ in at least three respects (Malle & Knobe, 2001). First, intentions represent the intender's own action ("I intend to A," where A is an action), whereas desires can represent anything ("I want O," where O can be an object or state of affairs, including another person's actions or experiences). Second, intentions are based on a certain amount of reasoning, whereas desires are typically the input to such reasoning ("I intend to A because I want O"). Third, intentions come with a characteristic commitment to perform the intended action whereas desires do not. This distinction has clear consequences for self-regulation, interpersonal perception, and social coordination (including its breakdown in the case of misunderstandings), and future research on these relations would be highly desirable.

Another important folk distinction revealed by the intentionality concept is that between desires and beliefs. Desires are strongly embedded in a culture's shared knowledge base (Bruner, 1990) and are considered the primary motives of action (Searle, 1984, chapter 4). This is because desires represent the end toward which the agent strives, whereas beliefs represent the various aspects of the path toward that end (Dretske, 1988). Desires also seem more primitive and easier to infer for children, who learn to attribute desires before they learn to attribute beliefs (e.g., Nelson-LeGall, 1985; Wellman & Woolley, 1990; Yuill & Perner, 1988). Relatedly, most autistic children lack the ability to ascribe beliefs to other people but have less difficulty ascribing desires to them (Baron-Cohen, 1995). Among adults, too, beliefs

and desires have distinct informational and impression-management functions in explanations of action (Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000).

The full concept of intentionality plays an important role in a number of social-cognitive phenomena. Frequently mentioned is its impact on the assignment of responsibility and blame for actions (e.g., Shaver, 1985): Agents are more likely to be held responsible or to be blamed when they performed the action in question intentionally. But even for unintentional behaviors and outcomes, the concept of intentionality is at work. Responsibility is still assigned when the outcome is considered to have been preventable (aka controllable; Weiner, 1995) by the agent and when it was his or her duty to do so (Hamilton, 1978). Both preventability and duty entail intentionality, because assigning duties to a person presumes that the person can intentionally fulfill them, and preventability presumes that the agent could have intentionally prevented the outcome.

Perhaps the most important function of the intentionality concept is to divide all behavioral events into two different domains that are subsequently manipulated in distinct ways by various cognitive tools (e.g., attention, explanation, prediction, blame). Heider (1958) was the first social psychologist to emphasize that people not only distinguish between intentional and unintentional behavior but also assume two different models of causality for them: Intentional behavior relies on agentic ("personal") causality, in which actions are based on the agent's reasons, deliberation, and formation of an intention; unintentional behavior relies on mechanical ("impersonal") causality, in which no reason or intention is involved.[2]

Observability

Another folk distinction leads to different cognitive manipulations: that between publicly observable and publicly unobservable events (Funder & Dobroth, 1987; John & Robins, 1993; Malle & Knobe, 1997b), which is really the distinction between mind and behavior. Considered jointly, the concepts of intentionality and observability generate a map of behavioral events that are relevant to social cognition—that is, events that people attend to, try to explain, predict, and evaluate (Malle & Knobe, 1997b; Malle & Pearce, 2001).

Attention to and Explanation of Behavioral Events

For convenience, we (Malle & Knobe, 1997b) adopted the following labels for the four regions of the behavioral events map: actions (observable and intentional), mere behaviors (observable and unintentional), intentional thoughts (intentional and unobservable), and experiences (unintentional and unobservable; see table 9.1). The labels themselves are of little significance, but the conceptual definitions of event types as combinations of intentionality and observability are. That is because the features of intentionality and observability allow us to predict, using a few simple postulates, the patterns of attention to and explanation of these behavioral events under various conditions (e.g., from the actor and the observer role and in communication or in private thought).

Table 1. *Postulated Folk Classification of Behavioral Events*

	Intentional	**Unintentional**
Observable	actions	mere behaviors
Unobservable	intentional thoughts	experiences

Which Behaviors People Attend To

To predict the allocation of attention to the four behavioral events in social interaction, we identified two factors that are known to govern attention allocation in general (e.g., Fiske & Taylor, 1991; Posner, 1980) and that are important to social interaction as well: epistemic access and motivational relevance. First, to turn one's attention to a particular behavioral event, one needs to have access to it—that is, become in some way aware of it taking place (through introspection, perception, or at least inference). Second, attention to an event increases if it is relevant (i.e., helpful) for the perceiver's processing or coordinating the current interaction (e.g., Jones & Thibaut, 1958; Wyer, Srull, Gordon, & Hartwick, 1982).

For actors, epistemic access is greater to their own unobservable events than to their own observable events, because they are constantly presented with their stream of consciousness but cannot easily monitor their own facial expressions, gestures, or posture (Bull, 1987; DePaulo, 1992; Gilovich, Savitsky, & Medvec, 1998). For observers, access is greater to other people's observable events than to their unobservable (mental) events. We therefore predicted that social interactants attend as observers to more observable events than as actors, whereas they attend as actors to more unobservable events than as observers (Hypothesis 1).

In addition, for observers the perceived relevance of intentional events is greater than that of unintentional events. That is because intentional events define the main business of an encounter (Goffman, 1974), because they are directed at the other and thereby demand a response, and because they have powerful effects on the other's emotions and moral evaluations (Shaver, 1985). By contrast, for actors the perceived relevance of unintentional events is greater than that of intentional events, because unintentional events were not controlled and therefore must be monitored and understood, whereas the execution of intentional events frequently relies on automatic programs (Norman & Shallice, 1986). We therefore predicted, second, that social interactants attend as observers to more intentional events than as actors, and they attend as actors to more unintentional events than as observers (Hypothesis 2).

We tested these predictions using an experimental paradigm in which pairs of participants had a conversation and, immediately afterward, were asked to report in writing everything "that was going on" with their partner (on one page) and with themselves (on another page), in counterbalanced order. The reports were then coded for references to behavioral events (verb phrases that referred to actions, mere behaviors, intentional thoughts, or experiences) and classified according to their intentionality and observability (for details of the coding, see http://darkwing.uoregon.edu/~interact/bevd.html).

Results across three studies confirmed both hypotheses (Malle & Pearce, 2001). In conversations among strangers, people reported overall 8 to 10 behavioral events per page (i.e., per perspective), but supporting Hypothesis 1, actors reported 2.3 more unobservable events than did observers, and observers reported 2.3 more observable events than did actors (η^2 = 50–60%).3 In addition, supporting Hypothesis 2, actors reported 1.1 more unintentional events than did observers, and observers reported 1.1 more intentional events than did actors (η^2 = 14–19%). These results suggest that attention in social interaction is allocated in ways that reveal the powerful impact of epistemic access and relevance on the behavioral event classification according to intentionality and observability.

Which Behaviors People Wonder About and Explain

Given this effect of intentionality and observability on the events people attend to, we should expect parallel asymmetries in the events people wonder about and try to explain. Moreover, the principles that guided the predictions in the domain of attention should be similar to those in the domain of wondering why and explaining, because the latter two processes imply a focused form of attention, guided by specific goals (Malle & Knobe, 1997b). Thus, for an event to elicit a wondering (and, under most circumstances, an explanation), three conditions must be met: there must be access (people must be aware of the event to wonder about it), nonunderstanding (people must not already have an explanation for the event), and relevance (people must find it useful and important to generate an explanation for the event).

From these conditions, we derived two predictions about patterns of wonderings: Because of differential access, actors should wonder more often about unobservable than observable events, while observers should wonder more often about observable than unobservable events. In addition, because of differential nonunderstanding, actors should wonder more often about unintentional than intentional events, and because of relevance observers should wonder more often about intentional than unintentional events (for details, see Malle & Knobe, 1997b, pp. 289-290).

We tested these predictions in two studies, collecting wonderings from memory protocols and twentieth-century novels and applying a coding scheme for the behavioral events that were explained (http://darkwing.uoregon.edu/~interact/bev.html). Confirming our predictions, actors wondered about more unobservable events (67%) than observable events (33%), whereas observers wondered about more observable events (74%) than unobservable events (26%). In addition, actors wondered about more unintentional events (63%) than intentional events (27%), whereas observers wondered about more intentional events (67%) than unintentional events (74%).

When deriving predictions about patterns of explanations (which are answers to wonderings), we drew a distinction between explanations that are directed to oneself (in private thought) and explanations that are directed to a partner (in communication). Explanations to oneself answer one's own wonderings, so they should show the same actor-observer asymmetries as wonderings, and data collected from memory protocols and diaries strongly confirmed this prediction. Explanations to others in communication, however, answer the others' wonderings, which come from the observer perspective, and so actors should explain behavioral events about which observers wonder, namely, intentional and observable ones. Observers, meanwhile, still explain the events that they wonder about (also intentional and observable ones), so in communication we should find no actor-observer asymmetries in the kinds of behavioral events people explain, and that was what we found (Malle & Knobe, 1997b).

The studies on both attention and explanation of behavioral events suggest that one function of the folk theory of mind and behavior is to divide the diversity of human behavioral and psychological stimuli into broad classes, such as action, experience, and so on, guided by the concepts of intentionality and observability. These event classes can be more easily managed cognitively by social perceivers, and they are tied to certain assumptions, such as about epistemic accessibility and relevance. Once again, these categorizations into broad event classes and their attendant assumptions in subsequent processing are not a matter of conscious decision ("I classify this as an action"; "I will pay more attention to her actions than to my own actions"). Some of the subsequent processes can certainly be put under conscious control, such as when an empathy instruction leads social perceivers to increase their attention to the other person's thoughts and feelings (Davis et al., 1996; Klein & Hodges, 2001; Malle & Pearce, 2001). But the classificatory function of the framework of mind and behavior precedes any particular processing.

Because the conceptual presorting that is achieved by a theory of mind guides and frames subsequent processes such as attention and explanation, variations in the conceptual framework itself will have direct effects on people's attention and explanations. For example, the degree of refinement in a theory of mind will influence the balance of attention allocation to all four behavioral event types. Consider the following remark by an autistic person: "It seems impossible to try to focus on my own thoughts or feelings and consider different thoughts or feelings in another person or persons at the same time, especially if I am talking or actively listening to the other person talk" (Blackburn et al., 2000). If the process of conceptual classification comes with ease and little ambiguity along the category boundaries, then attention regulation can more easily operate on it, because a directive such as "Attend more to the other's experiences" can be readily implemented. By contrast, if the conceptual classification is onerous, unreliable, and full of vagueness, then attentional regulation will have a difficult time holding on to the correct events and letting attention, explanation, or other processes operate on it.

Folk Explanations of Behavior

After this discussion of early categorization of behavioral events, I now move to the question of how and for what purpose people explain behavior. The folk theory of mind, and especially the intentionality

concept, plays a vital role in behavior explanations. Indeed, explaining behavior has sometimes been characterized as the hallmark of folk psychology or theory of mind, even though other processes such as prediction, control, and evaluation are of equal importance. Explanations, however, often come in verbal form and are therefore more amenable to investigation, especially if we want to learn about both their conceptual underpinnings and their role in social interaction.

Explanations and Theory of Mind

A first issue to address is the functional relation between behavior explanations and theory of mind. Is the function of a folk theory of mind to explain behavior, as most scholars assume, or is the function of explanations to rehearse and advance the theory of mind, as Gopnik (1998) suggests? Gopnik argues that explanations are like orgasm, which does not itself fulfill an evolutionary function but makes procreation, the important end, more desirable. However, the analogy becomes questionable when we consider that explanations, unlike orgasms, have many important uses beside making another end (theory advance) more desirable. That is, in addition to advancing theory of mind, explanations help in making sense of concurrent behavior, coordinating joint action, offering clarification, managing impressions, and so on. Furthermore, a theory of mind—even a very advanced one—is not really good for anything unless it improves or expands social performance and hence adaptive fitness either of the individual or the group. Behavior explanations constitute one such performance domain that is improved (or made possible) by a theory of mind, with others being prediction and influence. So the function of a theory of mind is not merely to explain behavior but to facilitate—by means of explanations and other tools—successful social cognition and social coordination (Malle, 2002). At the same time, the function of explanations is not merely to advance a theory of mind but to take on select social tasks, such as understanding, coordination, and impression management.

Now I can tackle in more detail the connection between the conceptual framework of mind and the social activity of offering behavior explanations. One possible position is that explanations within a theory of mind make behaviors understandable by identifying their mental causes. This is the position taken by many developmental researchers, who have traced the origin and advancement of explanations throughout the preschool years, demonstrating that children as young as 3 years systematically use "psychological" (mental state) explanations for human behavior (Wellman, Hickling, & Schult, 1997). However, these researchers group under psychological explanations statements that refer to the agent's desires and beliefs but also statements that refer to moods and lack of knowledge (Bartsch & Wellman, 1995, chapter 6; Schult & Wellman, 1997).

This global classification is problematic because it loses sight of two types of causation that people distinguish (Buss, 1978; Heider, 1958; Malle, 1999; Searle, 1983): The first type, intentional causation, refers to mental states as reasons of an agent's intentional action; the second may be called involuntary or "mechanical" causation, which refers to a variety of factors (including mental states) as causes of an agent's unintentional behavior. Current developmental studies leave open the question whether 3-year-old children who give mental-state explanations differentiate between mental states as reasons (for intentional behavior; e.g., "She bought milk because she wanted to make a cake") and mental states as mere causes (for unintentional behavior; e.g., "She was nervous because she really wanted to win the game"). Perhaps children first apply mental-state explanations broadly to human behavior and learn to distinguish between reasons and other (mental) causes only after acquiring the full-fledged concept of intentionality, around the age of 5 (Shultz & Wells, 1985). Command over this concept involves the differentiation of action-relevant mental states into the triad of belief, desire, and intention, which are partially confounded at an earlier age (chapter 10, this volume; Lyon, 1993; Moses, 2001). Competence over this triad of concepts implies the understanding that beliefs and desires are combined in a reasoning process to give rise to intentions, which themselves direct action (Malle & Knobe, 1997a), and this understanding amounts to an appreciation of the scope and limits of choice, also acquired around the age of 5 (Kalish, 1998).

My goal now is to outline the fully mature system of behavior explanations among adults and its grounding in a theory of mind. This grounding entails that behavior explanations can be constructed only

within the conceptual space of the folk theory of mind, and this space is broadly defined by the major distinction between intentional and unintentional behavior and by the specific concepts of reason and intention that underlie the folk notion of intentional action (Malle, 1999, 2001). To begin, I introduce a model of folk explanation that features four modes of explanation differentiated by the kinds of behaviors they explain (intentional vs. unintentional) and by the specific aspects of intentional behavior they target. Then I discuss conditions of use for each explanation mode. I close with an emphasis on both the cognitive and interpersonal functions of explanations, which also illuminate the cognitive and interpersonal functions of the folk theory of mind.

Four Modes of Behavior Explanation

When explaining behavior, people distinguish sharply between intentional and unintentional events (Heider, 1958; Malle, 1999; White, 1991), relying on the folk concept of intentionality (discussed in an earlier section). Unintentional events are explained by referring to mechanical causal factors (e.g., mental states, traits, others' behaviors, physical events), and we may label them cause explanations (top of figure 9.1). Traditional attribution models apply fairly well to these cause explanations, because people presume no other link between explanation and behavior besides causality (i.e., no components of intentionality such as awareness or intention).

Where traditional attribution theory fails is in its account of how people explain intentional behavioral events. These events are far more complex in that they are defined, according to the folk concept of intentionality, by awareness that accompanies the behavior, an intention that precedes the behavior, and beliefs and desires that precede and rationally support the intention (Malle & Knobe, 1997a). As a result of this complex definition, explanations of intentional behavior break down into three modes, which correspond to three domains that people find worth explaining (figure 9.1): reasons, causal history of reasons, and enabling factors.

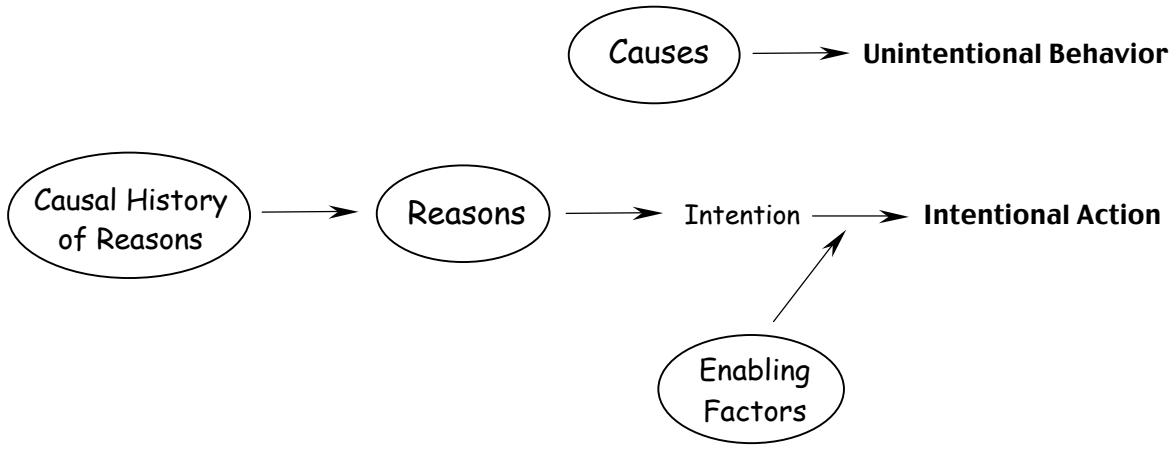

Figure 2. Four modes of explanation for unintentional and intentional behavior

Reason Explanations

The first and most important domain of behavior explanation comprises the agent's reasons for acting (Audi, 1993; Buss, 1978; Davidson, 1963; Locke & Pennington, 1982; Malle, 1999). Reasons are seen as representational states (i.e., mental states that represent an object or proposition) that the agent combines in a process of reasoning that leads to an intention and, if all goes well, to the intended action. The concept of intentionality specifies two paradigmatic types of reasons that precede the formation of an

intention: the agent's desire for an outcome and a belief that the intended action leads to that outcome. For example, a student explained why she chose psychology as her major by saying, "I want to go to graduate school in counseling psychology [desire]; I think psychology is the right major to have as background for counseling psychology [belief]." In many naturally occurring explanations, other reasons are mentioned in addition to or instead of the paradigmatic reasons, such as desires for avoiding alternative outcomes, beliefs about the context, beliefs about consequences, and valuings of the action itself. In all these cases, when an agent forms an intention in light of certain beliefs, desires, or valuings, these mental states constitute the reasons for which the agent forms the intention.

Reasons have two defining features, which can be labeled subjectivity and rationality (Malle, 1999, 2001; Malle et al., 2000). Subjectivity refers to the fact that reason explanations cite what explainers consider the agent's own subjective reasons for acting. That is, explainers try to reconstruct the deliberations (minimal as they may be) that the agent underwent when forming an intention and thus take the agent's subjective viewpoint when explaining the action. For example, the explanation "My father puts pressure on me because he wants many doors to be open to me" cites a desire in light of which (the explainer assumes) the agent decided to put pressure on her. In another example, "Why did she rush off?—She thought she was late for her class," we see even more clearly the subjectivity assumption, because the explainer subtly distances himself from the agent's belief and implies that, in reality, she probably was not late. But it was that subjective belief (and not objective reality) that guided the agent's action and thus explains it.

Rationality, the second defining feature of reason explanations, refers to the fact that the contents of beliefs, desires, and valuings that are cited as reasons have to hang together so as to offer rational support for the reasonableness of the intention and action that they brought about. Philosophers often speak of a "practical reasoning argument" that has reasons as its premises and the intention to act as its conclusion (e.g., Harman, 1976; Snare, 1991). The folk concept of rational support is probably not as strict; it demands merely that the intended action is a reasonable thing to do in light of this agent's desires and beliefs about fulfilling those desires. In the example above, the agent's action of rushing off was rationally supported by her belief that she was late for class (and it would not have been rationally supported if the agent had thought there was plenty of time left or if she had had no desire to be on time). To complete the practical argument in the first, the rational case, we would need to add (at least) her desire to be on time and her belief that rushing off may help her get to class on time. But one of the fascinating aspects of reason explanations is that the conceptual constraints that the folk theory of mind puts on reasons (especially the assumptions of subjectivity and rationality) allow explainers to mention a single reason and to trust the audience to fill in the remaining reasons and comprehend why the agent decided to act (Malle, 1999; Slugoski, Lalljee, Lamb, & Ginsburg, 1993).

Causal History of Reason Explanations

The second domain of explanation refers to factors that lie in the causal history of reasons (CHR) and thus clarify what led up to these reasons in the first place (Hirschberg, 1978; Locke & Pennington, 1982; Malle, 1994). For example, the statement "Anne invited Ben for dinner because she is friendly" attempts to explain Anne's action, but the content of the explanation ("she is friendly") refers to a factor in the causal history of her reasons, not to a reason itself. The explainer would not claim that Anne deliberated, "I am friendly—I should invite Ben for dinner"; rather, the explainer cites Anne's friendly disposition as a relevant causal history or background to whatever specific reasons Anne had for inviting Ben. Causal history explanations do not just cite traits (in fact, only about 20% of them refer to traits) but also include childhood experiences, culture, past behavior, current physiological states, and situational cues that trigger a particular belief or desire (Malle, 1999).

Even though CHR explanations help clarify intentional actions, they do not function like reasons and therefore are not subject to the constraints of subjectivity and rationality. That is, the agent need not have considered or been aware of the causal history factors cited in the explanation (Malle et al., 2000), nor do CHR factors provide rational support for an explained action. In fact, the causality type assumed for causal history explanations is identical to that of cause explanations—both describe a mechanical,

involuntary generation of events. However, CHR explanations apply to intentional behavior, whereas cause explanations apply to unintentional behavior.

Enabling Factor Explanations

The third domain of explaining intentional action refers to factors that enabled the action to come about as intended (Malle, 1999). Such enabling factor explanations refer to the agent's skill, effort, opportunities, or facilitating circumstances (see McClure & Hilton, 1997; Turnbull, 1986). These explanations take it for granted that the agent had an intention (and reasons) to perform the behavior and clarify how it was possible that the action was in fact performed. For example, "She hit her free throws because she had practiced them all week." There is no mention of the agent's reasons (or any causal history of those reasons); rather, the explanation clarifies how it was possible that the agent acted as she had intended.

In sum, the concept of intentionality spans four domains of explanation and their corresponding modes. When intentionality is not ascribed, people offer cause explanations. When intentionality is ascribed, people offer either reason explanations, causal history of reason explanations, or enabling factor explanations. These four explanation modes have different conceptual assumptions and linguistic features (Malle, 1999; Malle et al., 2000); they can be reliably distinguished when coding naturally occurring explanations (Malle, 1998); and together they comprise a model of folk explanation that has clear advantages over classic attribution theory (Malle, 1999, 2001; Malle et al., 2000; O'Laughlin & Malle, 2002).

Social-Cognitive Conditions of Explanation Modes

I now examine the conditions that determine when and for what purposes these distinct modes of explanation are used in social interaction. This exploration illustrates two tight interconnections: between conscious and unconscious processes of explanation choice and between cognitive and interpersonal functions of behavior explanations.

The conditions that distinguish between the use of cause explanations and all other explanation modes are straightforward. The primary one is conceptual: the perceived intentionality of the explained behavior. Malle (1999) showed that the rated intentionality of 20 behaviors predicted the choice between cause and reason explanations at $r \geq .90$. This choice is determined by features of the conceptual framework itself and therefore largely unconscious. The intentionality judgment itself may be difficult, requiring conscious deliberation and a search for further information; but once the judgment is "unintentional," the decision to offer a cause explanation is conceptually bound.

The second condition that invites cause explanations is motivational in nature: the regulation of blame for socially undesirable behaviors. When an agent performs a socially undesirable behavior that could be seen as either intentional or unintentional (e.g., hitting one's opponent during racquetball), the agent will tend to offer cause explanations (e.g., "I didn't see you"), because they portray the behavior as unintentional (Malle, 1999), thereby limiting the amount of assigned blame. This decision process can be conscious (when the explainer effortfully creates a favorable impression) or unconscious (when the explainer deceives himself or herself into believing that the behavior was in fact unintentional).

More complex is the set of conditions that determine whether, given that a performed behavior is perceived as intentional, explainers offer a reason explanation (the default option for about 80% of explanations), a CHR explanation, or an enabling factor explanation. Research documents both cognitive and motivational conditions for this selection among explanation modes (see table 9.2).

Table 2. *Conditions Determining People's Mode of Explanation for Intentional Actions*

Conditions		Explanation Modes
Cognitive	*Kind of wondering*:	
	• What for? Why?	→ Reasons, CHRs
	• How was it possible? (difficult/obvious behaviors)	→ Enabling factors
	Information available:	
	• Specific	→ Reasons
	• General	→ CHRs
Motivational	*Impression management:*	
	• Appear rational	→ Reasons
	• Minimize blame	→ CHRs
	• Minimize moral implication	→ Enabling factors
	Audience Design:	
	• Listener wonders "Why?"	→ Reasons, CHRs
	"How was this possible?"	→ Enabling factors
	• Conversational maxims	→ e.g., CHRs for parsimony

Cognitive Conditions

A first cognitive condition is the type of wondering the explainer experiences when searching for an explanation. When the explainer tries to find out what motivated or instigated the behavior at hand, we may call this a "What-for?" wondering (best answered by offering reason explanations) or more generally a "Why?" wondering (best answered by offering reason or CHR explanations). By contrast, when the explainer tries to find out what made a particular intentional action possible, we may call this a "How-possible?" wondering, and it is best answered by offering enabling factor explanations.

Research shows that "How-possible?" wonderings are triggered by difficult or extreme behaviors (e.g., artistic or athletic feats) and by behaviors whose motives are obvious (in the given context). In Malle et al. (2000, Study 3), for example, difficult/obvious behaviors elicited enabling factor explanations in 40% of cases, whereas easy/nonobvious behaviors elicited enabling factors in only 6% of cases. Similar results, though cast in a different terminology, were reported by McClure and Hilton (1997).

An even more powerful cognitive condition of selecting explanation modes is the type of information the explainer has available about the agent and the action (Malle, 2001; O'Laughlin & Malle, 2002). Why-questions about intentional actions typically focus on a specific agent–action unit—for example, "Why did Phil [agent] wash the dishes after the party [action]?" Reason explanations, such as, "He wanted the kitchen clean in the morning," are the default response to such questions (Malle, 1999; Malle et al., 2000). Reasons are specific to the agent (they are the presumed subjective mental states that the agent considered when forming the intention to act), and they are specific to the action (they rationally support this particular action). When explainers do not have such specific information about why the particular agent performed the particular action, they try to recruit general information that is available about the type of agent or the type of action performed. General information—for example, about the agent's traits or

group memberships, the situational context, or the historical background of the action—is expressed in CHR explanations. For example, Phil's washing the dishes may be explained by saying, "He is a neurotic cleaner." Or, in a conversation between two teenagers, the question "Why didn't she speak to him?" was explained by the reply "The dynamics of their relationship have always been peculiar." In such cases, explainers apparently do not know the agent's specific reasons for performing the action in question. But they have general information available about the type of agent or the type of action performed, and they use this general information to construct a CHR explanation.

In support of this role of information availability, we found that people consistently use more CHR explanations when explaining other people's behavior than when explaining their own (Malle, Knobe, & Nelson, 2004), presumably because people rarely have access to others' specific reasons. In addition, people use more CHR explanations when explaining group actions than when explaining individual actions, because people tend to have more general than specific information available about groups (O'Laughlin & Malle, 2002).

Both of these cognitive conditions—type of wondering and type of information available—are likely to engage both conscious (effortful) and unconscious (automatic) processes. On the one hand, an explainer may consciously assess the specific context of the behavior in question, the information demanded by this context, and the availability of this information. On the other hand, these assessments are automatically fed into the conceptual framework of explanations, guiding the choice between the distinct modes of reason, CHR, and enabling factor explanations. As often pointed out, such routine aspects are well executed by unconscious processes, whereas the situationally specific assessments require some amount of effortful attention.

Motivational Conditions

The predominant motivational condition of selecting explanation modes is impression management. By crafting certain types of explanation, people are able to manage both their self-presentations and their portrayals of others. Self-presentation concerns have an obvious impact on explanations (Scott & Lyman, 1968; Tedeschi & Reiss, 1981), but this impact is not limited to a choice between "person causes" and "situation causes," as attribution researchers have suggested (e.g., Bradley, 1978; Miller & Ross, 1975).4

When explaining intentional behavior, people increase their use of reasons, especially belief reasons, when they want the agent to appear rational (Malle et al., 2000), and they prefer CHR explanations to dampen the appearance of the agent's deliberation and responsibility (Nelson & Malle, 2000; Wilson, 1997). When explaining group actions, people offer reason explanations to portray a group as jointly acting (O'Laughlin & Malle, 2002) and thus perhaps as more dangerous (Abelson, Dasgupta, Park, & Banaji, 1998). Finally, a number of philosophers have suggested that reasons mark an action's moral worth, whereas enabling factors, such as intelligence or skill, do not (Foot, 1978; Kant, 1785/1998). We would therefore expect that explainers who want to portray an agent as morally worthy will offer reasons, whereas explainers who want to portray the agent as capable will offer enabling factors. For example, a professor's behavior of giving especially clear lectures might either be explained with a reason (e.g., "because she wants students to really understand") and elicit moral praise, or it might be explained with an enabling factor (e.g., "because she is very intelligent"), eliciting a positive but probably not moral impression.

None of these decisions entails a conscious thought of the sort "I should offer a reason rather than a CHR factor." People do not have an explicit conception of these different explanation modes, even though they reliably distinguish between them implicitly (Malle, 1999). What people are conscious of is, again, the situationally specific demands and certain goals of dealing with them (e.g., "I should appease them"). The routinized framework of behavior explanations then provides the conceptual and linguistic tools that implement these demands and goals by means of particular modes of explanation.

Because behavior explanations are often embedded in conversation (Hilton, 1990; Kidd & Amabile, 1981; Malle & Knobe, 1997b), another important motivational condition of choosing among explanation

modes is audience design—the adjustment of an explainer's communication to the interest, knowledge, or expectation of the audience (Clark & Carlson, 1982; Fussell & Krauss, 1992; Higgins, McCann, & Fondacaro, 1982; Zajonc, 1960). To begin, listeners can experience different types of wondering, and explainers have to choose an explanation mode that answers the listener's specific wondering. These wonderings are most clearly expressed in explicit question formulations: "Why?" "For what reason?" "How was this possible?" (Malle et al., 2000; McClure & Hilton, 1998). In one study, participants offered 95% enabling factor explanations in response to the question "How was this possible?" but only 10% in response to the question "For what reason?"5 Once again, conversational demands are often effortfully processed if they are situationally specific and if they require fine-tuning of the message, but the implementation of the explanation in terms of particular modes and linguistic formulations will be largely automatic.

Audience design entails conformity to general conversational maxims (Grice, 1975), which are so well practiced that they are heeded automatically. When asked a why-question, people are expected to avoid giving obvious explanations, too many explanations, uninformative explanations, or no answer at all. Obeying these maxims is likely to have direct consequences for the modes of explanation people choose. For example, when they explain intentional actions of aggregate groups—whose members act independently and probably for very different reasons—explainers aim at parsimony. That is, they prefer to offer CHR explanations, citing one or two factors that preceded and brought about the abundance of individual reasons among members of the group (O'Laughlin & Malle, 2002).

Discussion

Three general points concerning the choice among explanation modes are worth discussing. First, if folk explanations of behavior rely on key conceptual components of theory of mind (e.g., the concept of intentionality, the distinction between beliefs and desires) and if a person lacks these concepts, then the person's choice of explanation should be reduced to one, a simple mechanical explanation mode. The following self-description of an autistic adult lends support to this hypothesis: "I assumed that everything is predetermined and that adults were taking care of us according to some sort of program, without their own decision making" (Blackburn et al., 2000; see also Baron-Cohen, 1992). Of course, systematic research on autistic children's behavior explanations is needed to test this hypothesis.

A second point concerns the microstructure of choosing between explanation modes, in which conscious representations (e.g., of the audience and its demands, of one's own curiosity) blend in gracefully with unconscious processes (e.g., reliance on conceptual assumptions and automatic choice of words when constructing the explanation). The division of labor between conscious and unconscious processes might appear roughly as follows: The unconscious apparatus of folk explanation is a toolbox (of conceptual assumptions, cognitive routines) whose tools are automatically assembled (e.g., put into words) before use. This toolbox represents a stable, reliable part of social cognition. By contrast, conscious representations track the moment-to-moment fluctuations in the situation (and in oneself) and repeatedly converge on macro choices (e.g., to offer an explanation) that are then translated into the microelements of appropriate conceptual structure, wording, and so on. These translations are much like buttons or switches on a stereo amplifier, each of which has a broader meaning (e.g., increasing volume, selecting a source) and translates that meaning into a complex, low-level operation that reliably gets the job done.

The third point is that the conditions of choosing explanation modes depict explanations both as a cognitive tool (to answer one's own wondering) and as a social tool (to manage impressions and adapt to an audience). This duality of functions also exists at other levels of analysis (Malle, in press). For example, reason explanations have several specific features, among them the type of reason cited (referring either to a belief state or a desire state) and the linguistic marking of that state with a mental-state verb ("I thought," "she wanted"). Knowing the agent's specific belief or desire reasons, a social perceiver can more easily understand and predict the agent's behavior, thus using explanations as a cognitive tool. But agents who explain their own behavior also use the different types of reasons for managing the audience's perception of their rationality and blameworthiness (Malle et al., 2000; Nelson & Malle, 2000). Similarly, when people

explain others' behavior, they use mental-state verbs to emphasize that these are the agent's (and not some commonly accepted) reasons, thus distancing themselves from the particular reason (e.g., "Why is she not eating any dessert?""She thinks she's been gaining weight"; Malle et al., 2000).

The fundamental duality of cognitive and social function characterizes not only modes and features of folk explanations but also the folk theory of mind as a whole, which is a conceptual apparatus that helps solve cognitive as well as social tasks. I have pointed to several cognitive tasks, including classification of behaviors as intentional or unintentional, regulation of attention to behavioral events, and explanation as well as prediction. Among social tasks, I mentioned interpersonal influence and persuasion, impression management (of self and others), and communicative design. It should not be surprising that this diversity of tasks and functions requires far more than a system of causal reasoning or trait/situation attribution; it requires an interwoven framework of folk concepts that tie behavior to mind and thus make behavior intelligible, predictable, and socially defensible.

Conclusion

The theoretical perspective of social cognition as theory of mind has been underrepresented in recent social-psychological thinking, despite its affinity with Heider's (1958) groundbreaking investigations. Perhaps its representation will increase once sufficient data are amassed that favor, for example, a folk-theoretical model of behavior explanations over the traditional trait and causal attribution models (e.g., Malle, 1999, in press; Malle et al., 2000; O'Laughlin & Malle, 2002). But the theory of mind perspective is more than a replacement of attribution theory. Rather, it directs the study of social cognition to the fundamental concepts by which people organize the social world, concepts that guide all other (conscious and unconscious) processing of human behavior and experience. The theory of mind perspective also makes clear what is uniquely social about social cognition: a mentalistic conceptual framework of human behavior that can evolve and develop only within a social environment (Dunn, 1999; Whiten, 1999), whose primary function is to improve social coordination (Humphrey, 1976; Malle, 2002), and whose most reliable trigger is ongoing social interaction (Ickes, 2002). While illuminating the uniqueness of human social cognition, the theory of mind perspective also links social psychology to other disciplines that are concerned with human cognition of mind and behavior, such as developmental psychology, primatology, anthropology, linguistics, and philosophy (Carruthers & Smith, 1996; Greenwood, 1991; Hurford, Studdert-Kennedy, & Knight, 1998; Malle et al., 2001; Rosen, 1995). From this perspective, then, a full understanding of the "new unconscious" includes the folk-conceptual unconscious as an essential part of social cognition, which itself ranges from the most fundamental conceptual assumptions about mind and behavior to the most sophisticated assessments of ongoing social interaction.

Preparation of this chapter was supported by NSF CAREER award SBR-9703315. I am grateful to Dan Ames, John Bargh, and Jim Uleman for their comments on a previous version.

Notes

1 Heider and Simmel's (1944) findings are probably more indicative of people's sensitivity to the experimenters' intentions (to display geometric figures that move like agents) than of a deep application of theory of mind to circles and triangles (Malle & Ickes, 2000). More interesting are extensions to natural phenomena (including deities) and machinery. Some scholars regard these extended uses as violating the domain-specificity and modularity of a theory of mind (e.g., chapter 11). A defender of modularity might argue, however, that evolutionary and developmental primacy is critical for domain-specificity, and the data do at least not speak against this primacy. Extended applications such as those to natural phenomena and machines are surely not ruled out by a domain-specific framework but may show its powerful capacity to reorganize thinking and reasoning about the world.

2 Heider's distinction between the two modes of causality—personal (intentional) and impersonal (unintentional)—is typically misrepresented as one between person causes and situation causes for any kind of behavior (Malle, in press; Malle & Ickes, 2000). This misunderstanding is perhaps the fundamental flaw of standard attribution theory, to which I will return in more detail later.

3 The reported differences represent the actual interaction effect, computed after removing main effects (Rosnow & Rosenthal, 1989).

4 When people offer cause explanations of unintentional behaviors or outcomes (such as failure, damage, accidents, etc.), they choose between causes of various types—person vs. situation, stable vs. unstable, global vs. specific, controllable vs. uncontrollable. Theories that model these choices (e.g., Weiner, 1986; Fincham, Beach, & Nelson, 1987) are clearly of psychological significance, but they leave out the conceptually more complex choices between modes of explanation for intentional behavior.

5 The reported numbers are for difficult behaviors. For easy behaviors, the corresponding numbers were 22% in response to "How was this possible?" and 0% in response to "For what reason?", attesting to the strong influence of the cognitive condition discussed earlier.

References

Abelson, R. P., Dasgupta, N., Park, J., & Banaji, M. R. (1998). Perceptions of the collective other. *Personality and Social Psychology Review, 2*, 243-250.

Ames, D. A. (in press). Mental state inference in person perception: Everyday solutions to the problem of other minds. *Journal of Personality and Social Psychology.*

Asch, S. E. (1952). *Social psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Audi, R. (1993). *Action, intention, and reason*. Ithaca, NY: Cornell University Press.

Baird, J. A., & Baldwin, D. A. (2001). Making sense of human behavior: Action parsing and intentional inference. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 193-206). Cambridge, MA: MIT Press.

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*, 832-843.

Baldwin, D. A., & Baird, J. A. (1999). Action analysis: A gateway to intentional inference. In P. Rochat (Ed.), *Early social cognition: Understanding others in the first months of life* (pp. 215-240). Mahwah, NJ: Erlbaum.

Baldwin, D. A., & Tomasello, M. (1998). Word learning: A window on early pragmatic understanding. In E. V. Clark (Ed.), *The proceedings of the twenty-ninth annual child language research forum* (pp. 3-23). Stanford, CA: Center for the Study of Language and Information.

Baron-Cohen, S. (1992). The girl who shouted in the church. In R. Campbell (Ed.), *Mental lives: Case studies in cognition* (pp. 11-23). Oxford: Blackwell.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S. (1999). The evolution of a theory of mind. In M. C. Corballis & S. E. G. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 261-277). New York: Oxford University Press.

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (pp. 3-20). New York: Oxford University Press.

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

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press.

Blackburn, J., Gottschewski, K., George, E., & L——, N. (2000, May). A discussion about theory of mind: From an autistic perspective. *Proceedings of Autism Europe's 6th International Congress, Glasgow*. Scottish Society for Autism. Downloaded from http://www.autistics.org/library/AE2000-ToM.html on May 1, 2001.

Blakemore, S., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*, 561-567.

Bogdan, R. J. (2000). *Minding minds: Evolving a reflexive mind by interpreting others*. Cambridge, MA: MIT Press.

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology, 36*, 56-71.

Brentano, F. C. (1973). *Psychology from an empirical standpoint* (A. C. Rancurello, D. B. Terrell, & L. L. McAlister, Trans.) New York: Humanities Press. (Original work published 1874)

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

Bull, P. E. (1987). *Posture and gesture*. Oxford, UK: Pergamon Press.

Buss, A. R. (1978). Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology, 36*, 1311-1321.

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development, 69*, 672-691.

Carruthers, P. & Smith, P. K. (Eds.). (1996). *Theories of theories of mind*. New York: Cambridge University Press.

Clark, H. H., & Carlson, T. B. (1982). Speech acts and hearers' beliefs. In N. V. Smith (Ed.), *Mutual knowledge* (pp. 1-36). New York: Academic Press.

D'Andrade, R. (1987). A folk model of the mind. In D. Holand & N. Quinn (Eds.), *Cultural models in language and thought* (pp. 112-148). New York: Cambridge University Press.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy, 60*, 685-700.

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology, 70*, 713-726.

De Villiers, J. (2000). Language and theory of mind: What are the developmental relationships? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (2nd ed., pp. 83-123). New York: Oxford University Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological Bulletin, 111*, 203-243.

Dittrich, W. J., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception, 23*, 253-268.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dunn, J. (1999). Mindreading and social relationships. In M. Bennett (Ed.), *Developmental psychology: Achievements and prospects* (pp. 55-71). Philadelphia: Psychology Press.

Fincham, F. D., Beach, S. R., & Nelson, G. (1987). Attribution processes in distressed and nondistressed couples: III. Causal and responsibility attributions for spouse behavior. *Cognitive Therapy and Research, 11*, 71-86.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.

Foot, P. (1978). Virtues and vices. In P. Foot, *Virtues and vices and other essays in moral philosophy* (pp. 1-18). Berkeley: University of California Press.

Frith, U. (2000). Cognitive explanations of autism. In K. Lee (Ed.), *Childhood cognitive development: The essential readings* (pp. 324-337). Malden, MA: Blackwell.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*, 409-418.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*, 378-391.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165-193.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75*, 332-346.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA: Harvard University Press.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language, 4*, 161-185.

Goldman, A. I. (2001). Desire, intention, and the simulation theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 207-225). Cambridge, MA: MIT Press.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*, 101-118.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge: MIT Press.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257-293). New York: Cambridge University Press.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language, 1*, 158-171.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Greenwood, J. D. (Ed.). (1991). *The future of folk psychology: Intentionality and cognitive science*. Cambridge, UK: Cambridge University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). New York: Academic Press.

Hamilton, V. L. (1978). Who is responsible? Towards a social psychology of responsibility attribution. *Social Psychology, 41*, 316-328.

Harman, G. (1976). Practical reasoning. *Review of Metaphysics, 29*, 431-463.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243-259.

Higgins, E. T., McCann, C. D., & Fondacaro, R. (1982). The "communication game": Goal-directed encoding and cognitive consequences. *Social Cognition, 1*, 21-37.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*, 65-81.

Hirschberg, N. (1978). A correct treatment of traits. In H. London (Ed.), *Personality: A new look at metatheories* (pp. 45-68). New York: Wiley.

Hirschfeld, L. A., & Gelman, S. A. (Eds.). (1994). *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.

Hoffman, M. L. (1993). Empathy, social cognition, and moral education. In A. Garrod (Ed.), *Approaches to moral development: New research and emerging themes* (pp. 157-179). New York: Teachers College Press.

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16*, 233-253.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. Hinde (Eds.), *Growing points in ethology* (pp. 303-317). New York: Cambridge University Press.

Hurford, J. R., Studdert-Kennedy, M., & Knight, C. (Eds.). (1998). *Approaches to the evolution of language: Social and cognitive bases*. New York: Cambridge University Press.

Ickes, W. (2002). Subjective and intersubjective paradigms for the study of social cognition. In J. P. Forgas & K. D. Williams (Eds.), *The social self: Cognitive, interpersonal, and intergroup perspectives* (pp. 205-218). Philadelphia: Psychology Press.

Ickes, W. (Ed.). (1997). *Empathic accuracy*. New York: Guilford.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*, 521-551.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219-266). New York: Academic Press.

Jones, E. E., & Thibaut, J. W. (1958). Interaction goals as bases of inference in interpersonal perception. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 151-178). Stanford, CA: Stanford University Press.

Kalish, C. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development, 69*, 706-720.

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Ed. & Trans.). New York: Cambridge University Press. (Original work published 1787)

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.). New York: Cambridge University Press. (Original work published 1785)

Kidd, R. F., & Amabile, T. M. (1981). Causal explanations in social interaction: Some dialogues on dialogue. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 307-328). Hillsdale, NJ: Erlbaum.

Klein, K. J. K., & Hodges, S. D. (2001). Gender differences and motivation in empathic accuracy: When it pays to care. *Personality and Social Psychology Bulletin, 27*, 720-730.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309-324.

Krauss, R. M., Fussell, S. R., & Chen, Y. (1995). Coordination of perspective in dialogue: Intrapersonal and interpersonal processes. In I. Markova, C. F. Graumann, & K. Scherer (Eds.), *Mutualities in dialogue* (pp. 124-145). Cambridge, UK: Cambridge University Press.

Leslie, A. M. (1992). Autism and the "theory of mind" module. *Current Directions in Psychological Science, 1*, 18-21.

Leslie, A. M. (1995). A theory of agency. In A. J. Premack, D. Premack, & D. Sperber (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 131-149). Oxford: Clarendon Press.

Locke, D., & Pennington, D. (1982). Reasons and other causes: Their role in attribution processes. *Journal of Personality and Social Psychology, 42*, 212-223.

Lyon, T. D. (1993). *Young children's understanding of desire and knowledge*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Malle, B. F. (in press). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Cambridge, MA: MIT Press.

Malle, B. F. (1994). *Intentionality and explanation: A study in the folk theory of behavior*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Malle, B. F. (1998). *F.Ex: Coding scheme for people's folk explanations of behavior*. University of Oregon. Downloaded from http://darkwing.uoregon.edu/~bfmalle/fex.html on April 1, 2004.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*, 23-48.

Malle, B. F. (2001). Folk explanations of intentional action. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 265-286). Cambridge, MA: MIT Press.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givon & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 265-284). Amsterdam: Benjamins.

Malle, B. F., & Ickes, W. (2000). Fritz Heider: Philosopher and psychologist. In G. A. Kimble & M. Wertheimer (Eds.), *Portraits of pioneers in psychology* (Vol. 4, pp. 193-214). Washington, DC and Mahwah, NJ: American Psychological Association and Erlbaum.

Malle, B. F., & Knobe, J. (1997a). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101-121.

Malle, B. F., & Knobe, J. (1997b). Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology, 72*, 288-304.

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 45-67). Cambridge, MA: MIT Press.

Malle, B. F., & Pearce, G. E. (2001). Attention to behavioral events during social interaction: Two actor-observer gaps and three attempts to close them. *Journal of Personality and Social Psychology, 81*, 278-294.

Malle, B. F., Knobe, J., & Nelson. S. E. (2004). *Actor-observer asymmetries in behavior explanation: New answers to an old question*. Manuscript under revision.

Malle, B. F., Knobe, J., O'Laughlin, M., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology, 79*, 309-326.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.

McCabe, K. A., Smith, V. L., & LePore, M. (2000). Intentionality detection and "mindreading": Why does game form matter? *Proceedings of the National Academy of Sciences, 97*, 4404-4409.

McClure, J., & Hilton, D. (1997). For you can't always get what you want: When preconditions are better explanations than goals. *British Journal of Social Psychology, 36*, 223-240.

McClure, J., & Hilton, D. (1998). Are goals or preconditions better explanations? It depends on the question. *European Journal of Social Psychology, 28*, 897-911.

Mele, A. R. (2001). Acting intentionally: Probing folk notions. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 27-44). Cambridge, MA: MIT Press.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*, 213-225.

Moses, L. J. (2001). Some thoughts on ascribing complex intentional concepts to young children. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 69-83). Cambridge, MA: MIT Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81-103.

Nelson, S. E., & Malle, B. F. (2000, April). *Explaining intentional actions in the context of social perception and judgment*. Poster presented at the annual meeting of the Western Psychological Association, Portland, OR.

Nelson-LeGall, S. A. (1985). Motive-outcome matching and outcome foreseeability: Effects on attribution of intentionality and moral judgments. *Developmental Psychology, 21*, 323-337.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 1-17). New York: Plenum Press.

O'Laughlin, M., & Malle, B. F. (2002). How people explain actions performed by groups and individuals. *Journal of Personality and Social Psychology, 82*, 33-48.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*, 3-25.

Povinelli, D. J. (1996). Chimpanzee theory of mind: The long road to strong inference. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 243-329). Cambridge, UK: Cambridge University Press.

Povinelli, D. M. (2001). On the possibilities of detecting intentions prior to understanding them. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 225-248). Cambridge, MA: MIT Press.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition, 36*, 1-16.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*, 515-526.

Rosen, L. (Ed.). (1995). *Other intentions: Cultural contexts and the attribution of inner states*. Santa Fe, NM: School of American Research Press.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105*, 143-146.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Schneider, D. J. (1991). Social cognition. *Annual Review of Psychology, 42*, 527-561.

Schult, C. A., & Wellman, H. M. (1997). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition, 62*, 291-324.

Scott, M. B., & Lyman, S. M. (1968). Accounts. *American Sociological Review, 33*, 46-62.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge, UK: Cambridge University Press.

Searle, J. R. (1984). *Minds, brains, and science*. Cambridge, MA: Harvard University Press.

Shaver, K. G. (1975). *An introduction to attribution processes*. Cambridge, MA: Winthrop.

Shaver, K. G. (1985). *The attribution of blame*. New York: Springer-Verlag.

Shultz, T. R., & Wells, D. (1985). Judging the intentionality of action-outcomes. *Developmental Psychology, 21*, 83-89.

Slugoski, B. R., Lalljee, M., Lamb, R., & Ginsburg, G. P. (1993). Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology, 23*, 219-238.

Snare, F. (1991). *Morals, motivation and convention: Hume's influential doctrines*. Cambridge, UK: Cambridge University Press.

Tagiuri, R., & Petrullo, L. (1958). *Person perception and interpersonal behavior*. Stanford, CA: Stanford University Press.

Tedeschi, J. T., & Reiss, M. (1981). Verbal strategies as impression management. In C. Antaki (Ed.), *The psychology of ordinary social behaviour* (pp. 271-309). London: Academic Press.

Tomasello, M. (1998). Uniquely primate, uniquely human. *Developmental Science, 1*, 1-16.

Trabasso, T., & Stein, N. L. (1994). Using goal-plan knowledge to merge the past with the present and the future in narrating events on line. In M. M. Haith & J. B. Benson (Eds.), *The development of future-oriented processes: The John D. and Catherine T. MacArthur Foundation series on mental health and development* (pp. 323-349). Chicago: University of Chicago Press.

Turnbull, W. (1986). Everyday explanation: The pragmatics of puzzle resolution. *Journal for the Theory of Social Behavior, 16*, 141-160.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: Guilford.

Wellman, H. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition, 35*, 245-275.

Wellman, H. M., Hickling, A. K., & Schult, C. A. (1997). Young children's psychological, physical, and biological explanations. In H. W. Wellman & K. Inagaki (Eds.), *The emergence of core domains of thought: Children's reasoning about physical, psychological, and biological phenomena* (pp. 7-25). San Francisco, CA: Jossey-Bass.

Wellman, H. W., & Phillips, A. T. (2001). Developing intentional understandings. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 125-148). Cambridge, MA: MIT Press.

White, P. A. (1991). Ambiguity in the internal/external distinction in causal attribution. *Journal of Experimental Social Psychology, 27*, 259-270.

Whiten, A. (1999). The evolution of deep social mind in humans. In M. C. Corballis & S. E. G. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 173-193). New York: Oxford University Press.

Wilson, J. Q. (1997). *Moral judgment: Does the abuse excuse threaten our legal system?* New York: HarperCollins.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Malden, MA: Blackwell.

Woodward, A. L., Sommerville, J. A., & Guajardo, J. J. (2001). How infants make sense of intentional action. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 149-170). Cambridge, MA: MIT Press.

Wyer, R. S., Srull, T. K., Gordon, S. E., & Hartwick, J. (1982). Effects of processing objectives on the recall of prose material. *Journal of Personality and Social Psychology, 43*, 674-688.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology, 24*, 358-365.

Zajonc, R. B. (1960). The process of cognitive tuning in communication. *Journal of Abnormal and Social Psychology, 61*, 159-167.