Introduction: The Significance of Intentionality

Bertram F. Malle, Louis. J. Moses, and Dare A. Baldwin

Considerations of intentions and intentionality permeate human social life. Picture a first date, in which the partners try to find out their own and the other's desires, or a business negotiation, in which proclaimed intentions must be separated from hidden ones. Scan the human affairs columns for stories about conflicting desires and surmised intentions, or for legal cases about intent and insanity. Or simply read literature to see that human social interaction fundamentally requires that people infer and avow intentions as well as probe and affirm the intentionality of actions. If one took a Kantian approach to social cognition, searching for the fundamental concepts without which such cognition is impossible, intentionality would be one of those concepts, on par with space, time, and causality in the realm of non-social cognition.

Intentionality is a foundation for social cognition in several ways. For one, the concept of intentionality unlocks a central part of the folk ontology of mind, because intentionality's constituent components represent basic mental categories, such as belief, desire, and awareness. Moreover, the concept of intentionality brings order to the perception of behavior in that it allows the perceiver to detect structure—intentions and actions—in humans' complex stream of movement. Further, the intentionality concept supports coordinated social interaction by helping people explain their own and others' behavior in terms of its underlying mental causes. And intentionality plays a normative role in the social evaluation of behavior through its impact on assessments of responsibility and blame.

Intentionality is thus a tool with manifold functions, ranging from the conceptual to the interpersonal and even to the societal, and it is a tool with various domains of application, ranging from perception to explanation to

interaction. Contemporary research on the role of intention and intentionality in human social cognition has touched on all these functions and domains, but findings from this research are often discussed in isolation from one another. For example, much philosophical work has been devoted to analyzing the conceptual components of intentional action, but these analyses have rarely guided psychological research on the social perception of intentional behavior. Within psychology, the role of intentionality in explanations and in assignments of responsibility has been studied in developmental psychology within the paradigm of "theory of mind" and in social psychology within the paradigm of "attribution theory," but little communication has occurred between these paradigms. The central aim of this volume is to bring together the various disciplines, approaches, and traditions that have examined intentions and intentionality and to integrate current knowledge of this central facet of human cognition.

With this integrative aim in mind, we have organized this introductory chapter and the book by the major research questions being asked about intentionality across disciplines. A first set of questions concern how the concept of intentionality is defined within the folk theory of mind, what components make up this concept, how the components are related, and how they are acquired between infancy and adulthood. Earlier attempts to answer these questions can be found both in philosophy (e.g., Brand 1984; Mele and Moser 1994; Schueler 1995; Searle 1983) and in psychology (e.g., Astington and Gopnik 1991; Malle and Knobe 1997a; Maselli and Altrocchi 1969; Moses 1993). A second set of questions concern how people perceive human action and detect its underlying intentions and motives. These issues, which are a major focus of current research in developmental psychology (e.g., Barresi and Moore 1996; Baldwin and Baird 1999; Zelazo, Astington, and Olson 1999), have attracted considerable attention in philosophy (Bogdan 2000; Carruthers and Smith 1996; Davies and Stone 1995), as they did in early social psychology (Heider 1958). A third set of questions concern the role of intentionality in people's explanations of behavior, and especially its role in distinguishing "reason explanations" from "cause explanations." Here, too, pertinent work spans the disciplines of development (Bartsch and Wellman 1989; Kalish 1998), social psychology (Buss 1978; Malle 1999; White 1991), and philosophy (Audi 1993; Davidson 1963; Lennon 1990). A final set of questions concern the role of

judgments of intentionality in the evaluation of human action in terms of responsibility and blame. These issues have been explored in legal, philosophical, and psychological writings (e.g., Duff 1990; Hart 1968; Shaver 1985; Wallace 1994; Williams 1993).

Conceptual Elements of Intentionality

If the study of intentionality is to be a successful cross-disciplinary enterprise, it will require conceptual clarification to establish a common conceptual language and a shared map with which to scout the territory. This is not merely a methodological desideratum. Conceptual clarity represents a necessary step toward a model of what intentionality consists of, what it means to people, and how it functions in the social world.

Intentionality has two quite different meanings. Brentano (1874) introduced it as a technical term that could be used to refer to the property of all mental states as being directed toward something. Desires, for example, may be directed toward attractive objects, and beliefs toward states of affairs (Searle 1983; Lyons 1995). Second, intentionality is the property of actions that makes ordinary people and scholars alike call them purposeful, meant, or done intentionally. The focus of this volume is on this second sense—more specific, on people's conceptions of this sense of intentionality.

Another important distinction is that between *intentionality* and *intention*. The two terms are sometimes equated in psychological writing, even though folk use and philosophical analysis mark them as distinct. Intentionality, as we have mentioned, is a quality of actions (those that are intentional or done on purpose), whereas intention is an agent's mental state that represents such actions. This type of mental state often precedes its corresponding action or even occurs without it. One can therefore ascribe intention to an agent without making a judgment of intentionality. The reverse is not true, however: The judgment of an action's intentionality typically implies the ascription of an intention to the agent. Whether this implication holds under all circumstances is still debated among philosophers (Adams 1986; Bratman 1987; Harman 1976; Mele, this volume), but studies of folk use have thus far supported it (Malle and Knobe 1997a).

Both intention and intentionality are complex concepts in that people apply them only when a number of conditions are met. The ascription of

an intention to A requires minimally that one grants the agent a desire for some outcome O and a belief that A will likely lead to O (Malle and Knobe 1997a). But an intention cannot be reduced to this belief-desire pair nor are intention ascriptions just shorthands for elaborate desire ascriptions. Three chapters in part I explore what unique conditions underlie people's ascriptions of intention and how these conditions distinguish the folk concepts of intention and desire.

In chapter 2, Bertram Malle and Joshua Knobe contend that the two concepts are distinguished by three features: First, intentions are directed at the intender's own action whereas desires can be directed at anything. Second, intentions are based on some amount of reasoning whereas desires are typically the input to such reasoning. Third, intentions come with a characteristic commitment to perform the intended action whereas desires do not. Malle and Knobe provide conceptual arguments and empirical data to support the validity of this tripartite model. They also speculate about the psychological functions of the folk distinction between desire and intention.

In chapter 3, Louis Moses argues that, although preschool children have some appreciation of motivational aspects of intention, they are not especially cognizant of the fact that agents' beliefs constrain their intentions. Such belief constraints further distinguish intentions from desires. For example, although agents cannot intend what they believe to be impossible (Davis 1984; Grice 1971; Velleman 1989), there is nothing that prevents them from desiring it. Moses concludes that children aged 3 and younger may collapse desire and intention within a generic pro-attitude concept.

Janet Astington, in chapter 4, also traces developmental stages in mastering the folk concept of intention and disentangling it from desire. Paradoxically, intention can be thought of as either an early-arriving or a late-arriving concept within children's developing theory of mind: Infants seem to be able to detect intentions by age 1 (see part II), but not until age 5 do children reliably master the distinction between intentions and desires. The complex adult concept of intention must therefore be acquired gradually, with some aspects (e.g., object-directedness) acquired well before others (e.g., self-referentiality). Genuine understanding of intention, Astington hypothesizes, depends on the emergence of "metarepresentational" understanding—the child's understanding that people's beliefs and desires are

mental representations of the world that mediate their actions in the world (Bartsch and Wellman 1995a; Perner 1991). Astington's hypothesis converges here with Moses's claim that an intention concept requires a belief concept, as the latter also rests on metarepresentational understanding.

The importance of metarepresentational understanding after age 3 poses a puzzle, however. Even 2-year-olds seem to have an understanding of mental states like goals or desire (Bartsch and Wellman 1995a; Wellman and Woolley 1990); thus, either children have already acquired a representational theory by age 3 or else their early understanding of desire, although mentalistic, is not representational. The latter would be an unusual combination of features in light of the philosophical tradition of defining mental states as representational states (Brentano 1874; Chisholm 1981; Searle 1983). In chapter 10, as one part of his larger argument, Alvin Goldman challenges the evidence and logic in support of such a non-representational desire concept.

The conditions for ascribing intentionality are even more complex than those for ascribing intention, as Alfred Mele demonstrates in chapter 1. Following Aristotle and Hume, philosophers have focused on desire and belief as primary features of intentionality; however, the mere presence of appropriate belief and desire states is not sufficient for an action to be intentional. For one thing, intentions are considered an additional condition of intentionality. A behavior may be performed in accordance with a belief-desire pair, but it would not count as intentional unless it was brought about by an intention grounded in that belief-desire pair (Brand 1984; Bratman 1987; Searle 1983; Thalberg 1984). Suppose Brenda fouls an opponent during a basketball game. Suppose further that we are certain Brenda wants to win the game and believes that fouling her opponent would help her win. Still, we can't be certain that she committed the foul intentionally unless we know of Brenda's specific intention, her *decision* to act on her desire and belief. Moreover, even a behavior that was based on appropriate desires and beliefs plus an intention may not count as intentional: The intention must also cause the action via skill rather than luck. For instance, a golf novice may (at most[1]) intend to hit a hole in one, but few would call the accomplished feat *intentional*. Both conceptual analyses and empirical studies have converged on identifying skill as a further necessary condition for intentionality (Malle and Knobe 1997a; Mele and Moser 1994). Finally,

the folk concept of intentionality appears to require a particular kind of awareness on the part of the agent, namely, the awareness of acting as intended (informants call it "knowing what she is doing" (Malle and Knobe 1997a)). This condition shares some similarity with Searle's (1983) notion of an "intention-in-action" even though, somewhat surprisingly, Searle did not characterize intentions-in-action as conscious. Future research should address the stringency of the awareness condition as it applies to habitual and automatic daily actions (e.g., eating or driving). In his analysis of these conditions of intentionality, Mele poses a number of additional questions about the intentionality concept that invite further research, such as the generality of the intentionality-intention implication, the boundaries of skill, and the relevance of moral considerations for judgments of intentionality (as opposed to the relevance of intentionality considerations for moral judgments—see chapter 16).

Despite some disagreement over details, there is consensus across disciplines that intention and intentionality are complex states that are ascribed only if a set of simpler component states are present (in contrast to models of "direct perception" of intentionality, which are discussed in the following section). The ascription conditions for intention minimally include the presence of desire, belief, and some form of commitment; those for intentionality minimally include the presence of desire, belief, intention, skill, and awareness. This does not mean, of course, that perceivers always compute each and every component before they ascribe the resulting complex state. Many routine actions and familiar social contexts permit the spontaneous judgment of intentions or intentionality without explicitly checking as to whether each component is present. However, the constituent components are very likely to be considered when such judgments are difficult to make, or when they are debated (as in interpersonal conflict or in a court of law).

Analysis of the components of intention and intentionality offers many advantages. For one, it helps us to separate concepts and phenomena that adult social perceivers distinguish, such as desires, intentions, and intentionality. As a result, we can ask precise questions, such as "How does children's theory of mind develop from broad category distinctions to differentiated component concepts?" Moreover, unlike the full-blown concepts themselves, their constituent components can be relatively easily

grounded in lower-order precursors, such as belief in perception, desire in bodily needs, and intention in acts of reaching—precursors that children may perceive both in themselves as agents (Russell 1996) and in others through social interactions (Bruner 1981; Dunn 1991). Analysis of components also sharpens our understanding of extraneous variables that influence judgments of intention and intentionality, among them emotions and stereotypes. Stereotypes may provide default assumptions about certain components, such as expectations about an agent's stable desires or beliefs, which may bias the perceiver toward or against ascribing intentions and intentionality. Finally, component models highlight a fundamental feature of folk theories of mind: that they consist of conceptual networks systematically relating beliefs, desires, intentions, and other mental states to one another and to observable behavior (Gopnik and Wellman 1994). These networks provide tools for parsing and organizing what might otherwise be a chaotic stream of mental experiences (in the case of self) and behaviors (in the case of both self and others).

Traditionally, intentions have been regarded as private mental states that one ascribes to individual persons. However, intentions can also be ascribed to pairs or groups of people, who may have a joint intention to see a movie together, win a game, or publish an edited volume. Recent work in philosophy and psychology (e.g., Abelson, Dasgupta, Park, and Banaji 1998; Bratman 1993; Gilbert 1989; O'Laughlin and Malle 2000; Searle 1990; Velleman 1997) has begun to explore the nature of such joint intentions and the intentionality ascribed to whole groups and even nations. In addition, psycholinguists have examined the emergence of shared meaning out of individual intentions, a necessary process for successful conversation and social coordination (Clark and Brennan 1991; Gibbs 1998; Krauss and Fussell 1996). Interesting questions about the "location" of joint intentions and the "location" of shared meaning arise. One wonders, for example, whether there actually exist group minds that "have" mental states or whether social perceivers merely metaphorically extend their folk ascriptions of mental states to group agents. These puzzles notwithstanding, the ascension from individual to shared mental phenomena is essential to human relations. Along these lines, Raymond Gibbs (chapter 5) examines the function of shared meaning in communication and argues that at least some intentions are not in the head but rather are emergent

social phenomena. In addition, Daniel Ames, Eric Knowles, Michael Morris, Charles Kalish, Andrea Rosati, and Alison Gopnik (chapter 15) review recent psychological evidence that people comfortably apply their theory of mind to individual agents as well as to group agents and, moreover, that cultures seem to differ in their tendency to designate either groups or individuals as the primary agent category.

Reading Intentions and Intentionality

People typically read the intentions underlying the behavior of others readily and with little conscious effort. Of course failures occur, and these are sometimes serious enough to give rise to argument, legal action, or international conflict. However, such interpretive failures are rare when measured against the countless actions to which perceivers smoothly assign relevant intentional meanings—actions such as tooth brushing, newspaper reading, and kitchen cleaning. Even actions motivated by complex and potentially obscure intentions, such as the casting-about behavior occasioned by a search for a television remote control, often pose little interpretive difficulty for perceivers. The same can be said of novel actions. On first viewing a skiboarder in action, for instance, it is easy to recognize the intention of thrill seeking.

How do people so effortlessly detect intentions within the dynamic behavior stream and so readily apprehend their content, and how is such skill acquired in children's development? Surprisingly, these questions received little systematic examination before the relatively recent attempts by social psychologists. For example, following ideas that Asch (1952) borrowed from gestalt psychology, Newtson and his colleagues (see, e.g., Newtson 1973; Newtson and Engquist 1976) argued that people *directly perceive* others' intentions on witnessing their actions. Intentionality, and the specific intentions at play, are thought to be there within the behavior stream, waiting to be detected. Working within a similar direct-perception framework, Premack and his colleagues (see, e.g., Premack 1990; Premack and Premack 1995) provided nativist speculations about the origins of the human ability to read others' intentions. They suggested that infants arrive in the world biologically prepared to perceive certain kinds of animate

motion (in particular, self-propelled motion) as intentional, and that infants and can recognize at least a small set of specific intentions (e.g., helping vs. hurting) on the basis of the different behavioral patterns associated with them.

Here it is important to separate possible claims about detecting intentionality (how perceivers recognize *that* an intention is being enacted) from those concerning the *content* of an agent's intention (how perceivers recognize which specific intention, or set of intentions, is being enacted). The direct-perception framework seems at least potentially fruitful in accounting for the former, but seriously flawed as an approach to the latter. Let us consider these points in turn. Concerning the detection of intentionality, organisms wired to read "intentional" anytime they encounter self-propelled motion would be off to something of a start, because of the correlation between such motion and intentional action (see also Mandler 1992 and Wellman and Phillips in this volume). Of course, they would then need to learn to suppress an intentionality reading for countless important exceptions involving self-propelled motion or the appearance of it: involuntary behaviors such as sneezing, accidental and incidental motions such as the inadvertent knocking of objects off counters, and the motion of many inanimates for which the cause of motion has either been missed or is not yet understood (e.g., feathers and leaves blown by the wind, falling rocks, cars and trains, computer cursors). Clearly, then, the direct-perception framework requires substantial embellishment to successfully account for the full spectrum of intentionality judgments that adult perceivers actually make. Nonetheless, it might well capture the essence of how infants get their start in the business of detecting intentionality.

In contrast, the direct-perception framework seems fundamentally unworkable as an account of how perceivers detect the *specific* intentions motivating others' behavior. Behavior patterns and intentions stand in a many-to-many relation (Baird 1999; Baldwin and Baird 1999; Searle 1984): One and the same action (e.g., pressing a hypodermic needle into another's arm) admits of multiple intentional interpretations (e.g., the intent to heal vs. harm), and one and the same intention (e.g., to heal) can give rise to many possible actions (e.g., referral, advice, medication, surgery). Moreover, an infinite number of possible intentions are consistent with any given action,

yet only one of these candidates (or at most a very small set) is actually relevant and usually considered by perceivers. Searle (1984, p. 58) captures this beautifully:

> If I am going for a walk to Hyde Park, there are any number of things that are happening in the course of my walk, but their descriptions do not describe my intentional actions, because in acting, what I am doing depends in large part on what I think I am doing. So for example, I am also moving in the general direction of Patagonia, shaking the hair on my head up and down, wearing out my shoes, and moving a lot of air molecules. However, none of these other descriptions seems to get at what is essential about this action, as the action it is.

Searle's example makes obvious that the content of agents' intentions cannot be recovered directly from the behavior stream itself; too many possible intentions are recoverable in any given case. In other words: From the standpoint of the perceiver, the content of agents' intentions is radically underdetermined by their behavior.

An alternative approach views the detection of intentions and intentionality as the outcome of an inferential system. (See, e.g., Baldwin 1993b; Baldwin and Baird 1999; Dittrich and Lea 1994; Meltzoff 1995; Tomasello, Kruger, and Ratner 1993.) Unlike the direct-perception account, this inferential framework readily accommodates, at least in principle, our ability as perceivers to deal with the complex link between actions and intentions. The specific intention motivating a given action is thought to be inferred not just from the flow of behavior itself but also from external information, including other cues in the immediate context (e.g., a medical setting such as a clinic, the presence of doctors, nurses, and medical supplies), prior knowledge about the agent (e.g., a physician vs. a violent offender), and the script within which the agent's motions are embedded (e.g., a physical exam vs. a session of interrogation and torture). Sensitivity to such "extrabehavioral" characteristics could enable perceivers to constrain their inferences about intentions in the face of the limitless possibilities.

In addition to providing a possible account for the ability to interpret the content of agents' intentions, the inferential approach seems amenable to explaining how people distinguish intentional from unintentional or incidental behavior. Because behaviors from these different classes seem structurally different in many cases (e.g., self-propelled motion is often intentional whereas motion caused by direct physical contact with another moving body rarely is), processing of the behavior stream can play an important role in

making these distinctions. This is why the direct-perception framework also can offer something in accounting for such ability. However, the inferential approach again has the advantage in that it also has the potential to explain the finer judgments about intentionality that social perceivers make—for example, the ability to recognize the behavior of sleepwalkers and "zombies" as unintentional despite the surface similarity of such motion to that of conscious agents. In such cases, an inferential account would point to information external to the behavior stream (e.g., night-time setting, history of night-time talking and walking, lack of response to questions, known ingestion of mind-altering substances) as crucial in shaping the perceiver's inferences about the agent's intentionality.

The chapters in part II all speak (at least implicitly) to the inferential framework. Each offers new ideas and new evidence regarding the processes involved in detecting intentions. In chapter 9, Jodie Baird and Dare Baldwin highlight some of the qualities requisite to an inferential system for intentional understanding. In particular, they propose that such an inferential system crucially depends on the operation of a low-level structure-detection mechanism capable of analyzing the dynamic behavior stream into relevant units—units coinciding with the initiation and the completion of intentions—for further analysis. Moreover, they present new evidence that adults as well as 10–11-month-old infants spontaneously parse continuous intentional action in terms of just such "intention-relevant" units.

To date, the preponderance of work within the inferential approach has focused on development, exploring infants' and young children's emerging abilities to detect and interpret the intentions motivating others' behavior. Many of the chapters in part II reflect this trend, as they deal primarily with developmental evidence. In chapter 7, Amanda Woodward, Jessica Somerville, and José Guajardo present research indicating that infants as young as 9 months understand the goal-oriented quality of some intentional actions and can distinguish between intentional and unintentional action in at least some cases. Further, they find that at this early age infants already use information external to the behavior stream to determine the relevance of a goal object. For example, previously provided information about an agent's interest in the contents of a box led infants to construe a subsequent box-grasping action as goal-oriented; in the absence of such prior information, infants failed to register the action's goal-oriented quality.

This "action-in-context" effect meshes nicely with the predictions of the inferential framework described above, which is grounded in the idea that intentions are inferred by interpreting action within its larger context.

Inferential theorists concerned with how children come to detect intentions and intentionality typically think in constructivist—as opposed to purely nativist—terms. To the extent that judgments about intentions are derived through complex inferential processes depending on experience and world knowledge, conceptual change within this arena is to be expected as development proceeds. Many of the chapters in part II manifest this partiality for constructivist speculation regarding the origins of intentional understanding. For example, in discussing the origins of early intentional understanding, Henry Wellman and Ann Phillips (chapter 6) offer an analysis of the kind of input regarding intentions and intentionality that might be available to infants through observation of others' behavior. They go on to present new evidence that infants as young as 12 months are sensitive to two perceptible features of behavior—object-directedness and action-connectedness—that are typical of intentional action. As Wellman and Phillips suggest, early recognition of these features may not actually reflect a genuine understanding of the agent's intentions; rather, it may represent crucial steps toward such understanding. In a similar vein, Woodward and colleagues (chapter 7) have found that infants at 9 months process some actions (such as grasping) in ways that are relevant to intentions but fail to process other actions (such as pointing) in these terms. This is among the first evidence suggesting that abilities enabling the detection of intentional content are constructed through infants' world experience. In chapter 9, Baird and Baldwin suggest that a low-level mechanism for analyzing action plays a crucial role in making possible developmental change of the kind that Woodward et al. demonstrate.

In chapter 8, Andrew Meltzoff and Rechele Brooks provide important ballast to the constructivist stance by embracing a hybrid account that credits its newborns with crucial skills for interpreting others' actions. Such an account helps to explain how typically developing children so easily and naturally come to understand others' intentions. Meltzoff and Brooks propose a "starting-state nativism" that consists of innate foundations that are modified by extensive development and practice in social interactions. The innate foundations are at work, for example, when a newborn imitates

someone's actions, thus translating a perceived act into an act of its own. This imitation relies on a mapping between others and self—at birth, on the level of actions. By 18 months, this mapping occurs on the level of goals. Meltzoff and Brooks describe experiments in which infants observed another person engaging in action that failed with respect to goal attainment (e.g., an attempt to open a device was unsuccessful) but nevertheless inferred the goal and spontaneously performed the action that successfully led to the goal (opening the device). Meltzoff and Brooks argue that the pivotal element in early imitation, in later goal inference, and in many other achievements of infant social cognition is the "like me" analogy—the tendency to see others' acts as being like acts the infants can produce themselves. As they experience their own attempts to control behavior, infants build maps that link effort experiences, goals, and their own actions, and they use these maps to infer others' goals from observed actions.

The idea of an innate analogical process shares important similarities with a simulation theory account such as the one suggested by Alvin Goldman in chapter 10. Like Meltzoff and Brooks, Goldman offers his account as an alternative to a purely inferential framework. A key element of his model is the distinction between first-person and third-person attributions of mental states. Goldman argues that first-person attributions are based not on inference but rather on a form of direct perception, or introspection. He defends this proposal against recent skepticism regarding the possibility of introspection (e.g., Gopnik 1993). Goldman goes on to argue that social perceivers make third-person attributions of mental states with the help of first-person access, using simulation processes to re-create and thereby represent others' mental states. He then musters various lines of evidence for the plausibility of simulation as a core mechanism of intention reading. This evidence includes the intriguing possibility of "mirror neurons" in primates that seem to fire both when the organism perceives certain actions performed by others and when it performs that action itself—a mechanism not unlike the supramodal representation system discussed by Meltzoff and Brooks in chapter 8.

Any attempt to account for the ability to read others' intentions must grapple with the extent to which such skill is special to humans. On the one hand, there is an obvious gulf dividing humans from other species in terms of the complexity of reasoning about others' intentions. Imputing to others

complex intentions with multiple subparts, such as intentions to embezzle funds for charitable purposes or to run for president of the United States in good faith but with little hope of success is everyday fare for humans, but as far as we can tell nothing of comparable complexity has ever been seen in other species. On the other hand, behavior that powerfully suggests the essence of intention-reading ability is ubiquitous in interactions between humans and other species and in interactions among members of other species. A dog's suspicion at signs indicating an intention to bathe him and his ecstasy when he notes preparations for a walk are examples familiar to many. Other examples are easy to find throughout modern literature, even at the academic level; witness two influential volumes concerning the social understanding and the "Machiavellian intelligence" of higher primates (Byrne and Whiten 1986; Whiten and Byrne 1997).

In chapter 11, Daniel Povinelli describes a substantial body of work investigating these issues. Based on the evidence, he suggests that chimpanzees lack a genuine appreciation for the intentions motivating others' actions, despite their evident skill at processing such actions in ways that enable them to predict and influence others' behavior. To account for this apparent paradox, Povinelli proposes the operation of two independent mechanisms. One is the skill of analyzing, correlating, and predicting complex patterns of behavior on the merely behavioral level; the other comprises genuinely mentalistic reasoning about intentions and intentionality underlying behavior. On this proposal, both humans and chimpanzees are skilled at behavior analysis but only humans are capable of mentalistic reasoning. And Povinelli speculates that the developmental progress from infants' behavior analysis to preschoolers' mentalistic reasoning may be a qualitative step—the emergence of a separate mechanism—rather than a gradual elaboration of a single mechanism from its incipient to its mature stage.

Intentionality and Explanations

Explanations of behavior are regarded by many as a key function of folk psychology, and the concept of intentionality plays a pivotal role in the construction of such explanations. Behavior explanations take cognitions of behavior (including judgments about the behavior's intentionality) as input

and render as output a model of what generated the behavior, often including reference to mental states such as beliefs, desires, and intentions. Such a model, in turn, influences judgments of responsibility, predictions of future behavior, and attempts to change the behavior.

Various traditions of explanation research across disciplines can be organized along two dimensions: whether or not the role of intentionality in explanations is considered, and whether or not explanations are studied in their natural context of conversation and social interaction. Figure 1 shows the resulting four combinations.

The first cell contains approaches to explanation that do not consider the role of intentionality and analyze explanations independently of the interactive context in which they occur. According to this *causal judgment* approach, explanations are cognitive processes (often unconscious) that apply equally to all objects of explanation, be they physical or behavioral, intentional or unintentional events. Prime representatives of this approach include Kelley's (1967) ANOVA model of causal attribution and its variants (e.g., Hewstone and Jaspars 1987), the cognitive study of causal reasoning (Spellman 1997; Cheng and Novick 1990), and normative models of scientific explanation (Hempel and Oppenheim 1948). The strength of this approach is that it focuses on the common cognitive principle of all explanations: that they identify causal antecedents of the explanandum according to rules of logic and evidence. At the same time, this focus is also its major weakness, because distinctions among explananda (e.g., actions vs. physical events) and corresponding distinctions between causal models

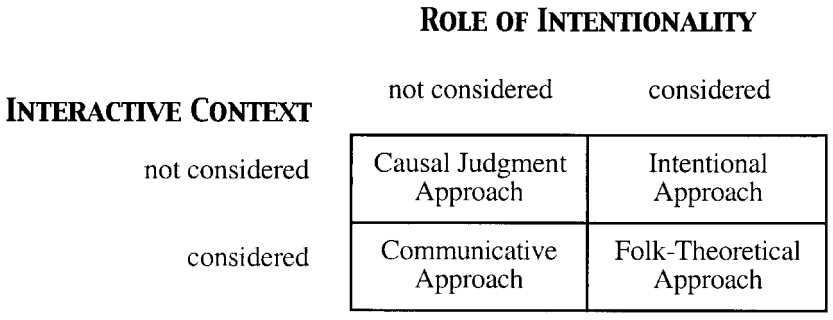

Figure 1
Four approaches to the study of explanation.

are overlooked, as is the social context that gives explanations their function, regardless of how they are cognitively (or physiologically) realized.

The second cell contains paradigms that still analyze explanations independently of their interactive context but that consider the role of intentionality by distinguishing between two types of explananda (intentional human action vs. all other events) and their corresponding types of explanation (often labeled *reasons* and *causes*). Representatives of this *intentional* approach include the hermeneutic movement in the social sciences (e.g., Gadamer 1989; Harre and Secord 1972; von Wright 1971) and the substantial contingent of philosophers who consider reason explanations of intentional action to be unique and irreducible (e.g., Audi 1993; Davidson 1963; Mele 1992a; Taylor 1964). The strength of this approach is that it recognizes important conceptual differences between two types of explanation and tries to work out the implications for theories of science, motivation, and action. However, the psychological reality of the two types of explanation has not been explored within this tradition, because it regards explanations primarily as logical entities rather than as verbal behaviors. Philosophers of action sometimes assume that ordinary people, too, distinguish between reasons and causes, but no systematic empirical data are ever mustered in support of this assumption.

The third cell contains the *communicative* approach, which emphasizes the social context and function of explanations, especially their dialogical nature in both everyday and scientific settings (Antaki 1994; Bromberger 1965; Hilton 1990; Kidd and Amabile 1981; Turnbull and Slugoski 1988). Explanations are seen as answers to why-questions, filling a knowledge gap exhibited by the questioner. (The explainer and the questioner are typically separate individuals, but in the case of private explanations they are identical.) Scholars within this approach—e.g., Slugoski, Lalljee, Lamb, and Ginsburg (1993)—have recognized that causal judgments are responsive to the social demands created by why-questions, such that explainers search for and present different explanations depending on their inferences about the questioner's background and the particular knowledge gap that is to be filled. For example, when Q asks E "How come Mary bought a Mercedes?" E might answer "Because it's a good car." However, if E considers that Q asks the question because Q knows Mary is poor, a more appropriate answer would be "Because she inherited a load of money" (McClure and Hilton

1997). Variations in explanations are therefore understood not just in terms of different causal perceptions but also in terms of different social demands that the explainer tries to meet (Malle, Knobe, O'Laughlin, Pearce, and Nelson 2000). Perceived gaps in knowledge present the most obvious demands; authority and accountability have also been examined (Edwards and Potter 1993; Scott and Lyman 1968; Tedeschi and Reiss 1981).

The communicative approach shares with the causal-judgment approach its major strength and its major weakness: It identifies general principles of explanation but does not distinguish between different types of explananda; hence, it still assumes that humans have a uniform conceptual model of causality. The consideration of social context, however, is a distinct strength of the communicative approach. Important questions follow from this consideration, such as whether the "truth" of an explanation depends solely on a speaker's and an audience's assumptions and to what extent social demands may modify not only verbal behavior but actual causal perceptions.

The fourth cell contains approaches emphasizing both the role of intentionality in explanations and the social-interactive context in which explanations are embedded. We call this the *folk-theoretical* approach because it considers explanations as an integral part of folk theories for core domains of cognition, such as psychology, physics, and biology (Carey 1995; Malle 1997; Wellman, Hickling, and Schult 1997). Within folk psychology, explanations heavily implicate the concept of intentionality in that all human behavior is classified as either intentional or unintentional and (depending on the classification) explained in conceptually distinct ways (Buss 1978; Locke and Pennington 1982; Kalish 1998; Malle 1999; Read 1987; White 1991). In particular, unintentional behavior is explained by mere causes, which are seen as simply bringing about the effect in a mechanical way (sadness causes crying; sunshine causes happiness). In contrast, intentional behavior is explained by reasons—the beliefs and desires in light of which the agent formed an intention to act.

Unlike scholars who take the intentional approach, those who endorse the folk-theoretical approach do not try to clarify the nature of explanation in a philosophical sense, nor do they necessarily postulate the objective existence of intentionality. Instead, they analyze explanations and intentionality as cognitive tools that guide people's perception, prediction, and control of their environment. Explanations are thus assigned a psychological

reality that is grounded in a shared folk-conceptual framework. In addition, and in agreement with the communicative approach, some researchers consider explanations also to be a social tool, expressed and strategically used in social interaction (Bartsch and Wellman 1995a; Malle and Knobe 1997b; Malle et al. 2000).

A possible weakness of the folk-theoretical approach is that, in appreciating the complexity and the variability of explanations in particular domains, it may lose the generality that other approaches seek. However, domain specificity is not incompatible with generality. Some functions of explanations (e.g., the need to anticipate and control one's environment (Heider 1958) and the desire to propel one's theoretical understanding of it (Gopnik 1998)) hold across domains; other functions (e.g., the face-saving and evaluative character of behavior explanations) are specific to particular domains. The same might be said of the cognitive processes underlying explanations, of which some may be domain-general (e.g., considerations of temporal order in causal processes) and others domain-specific (e.g., considerations of rationality in explanations of intentional behavior). Future research will have to clarify the exact similarities and differences among different types of explanation, but the contribution of the folk concept of intentionality to some of these differences is already apparent.

Several recent advances in folk-explanation research are driven by the assumption that people learn to master not one but a variety of modes of explanation (e.g., Kalish 1998; Malle 1999; McClure and Hilton 1997; Wellman et al. 1997). One question to be settled is whether these modes of explanation can be distinguished solely by their domain of application (Wellman et al. 1997; Schult and Wellman 1997) or whether they exhibit distinct conceptual structures (Kalish 1998; Malle 1999). Broad domain distinctions (e.g., physics, biology, psychology) provide useful approximations of explanatory types, but they may be less helpful for distinguishing modes of explanation within the richest of domains: that of human behavior. Because humans can be characterized as physical, biological, and psychological systems, explanations of human behavior encompass all the forms of explanation employed in other domains but also make use of unique modes that apply only to intentional action.

All the chapters in part III deal in one way or another with the interplay and the distinctions among the various modes of explaining human behav-

ior. In chapter 12, G. F. Schueler takes a close look at the tension between two ways of explaining intentional action: by means of the agent's own reasons, and by means of objective causal processes that seem to leave no room for genuine reasons. He argues that reason explanations are distinct from and not reducible to standard causal explanations because they fundamentally incorporate normative features. Even reason explanations offered by an observer must mimic the normative reasoning process the agent went through when deciding to act.

In chapter 13, Bertram Malle argues that folk explanations of intentional behavior encompass three different modes, with *reason* explanations the primary one and *causal history of reason* explanations and *enabling factor* explanations the secondary ones. Malle proposes a model that distinguishes these modes by their conceptual, linguistic, and functional features and contrasts this model with alternative theories of attribution and explanation. He also explores what implications this plurality of explanation modes has for the well-known debate between "theory theorists" (who assume that people ascribe mental states by relying on an organized set of generalizations) and "simulation theorists" (who assume that people ascribe mental states by relying on their capacity to simulate these states in their own mind).

In chapter 14, Andrea Rosati, Eric Knowles, Charles Kalish, Alison Gopnik, Daniel Ames, and Michael Morris explore the possible reconciliation between behavior explanations referring to the agent's mental states and behavior explanations referring to the agent's enduring traits. Rosati et al. identify mental-state components in trait concepts and demonstrate that trait inferences rely crucially on mental-state inferences. In the past, mental-state explanations were featured in the theory of mind tradition and trait explanations in the social-psychological attribution tradition. Rosati et al. attempt to integrate these rather disparate paradigms of research into social cognition.

Intentionality, Responsibility, and Social Context

Interpersonal perception encompasses not only cold assessments of others' mental states but also affective and moral responses to those states and to the social actions the states generate. These responses, which include praise, blame, pride, shame, resentment, and gratitude, may well be unique to

humans. They stake out the evaluative and corrective functions of the folk theory of mind, and they place issues of intentionality in a larger societal context. Moral responses are inextricably linked to the constituents of intentionality, to mental-state inferences, and to explanations of action, the topics discussed in parts I–III of this volume. But exactly how these connections should be understood—that is, exactly how intentionality relates to moral responsibility—is a matter of some debate in psychology, philosophy, and the law. Below we provide a conceptual framework for studying the intentionality–responsibility connection, and we locate some recent literature as well as this volume's contributions within that framework.

Responsibility has many meanings (cf. Frey and Morris 1991; Hamilton and Sanders 1992; Hart 1968), and so the role of intentionality in responsibility will differ depending on the meaning in view. *Responsibility* or *being responsible* always refers to a socially ascribed relation that holds either (1) between an agent and a specific action or outcome, (2) between an agent and that agent's general capacity for acting, or (3) between a cause and an effect. The third relation casts responsibility merely as causality (e.g., "The hurricane was responsible for 10 deaths"[2]), a derivative meaning that exports the first (agent/outcome) meaning of responsibility from the psychological realm into the realm of natural events but removes its implication of intentional agency. We therefore focus here on the two principal meanings, which might be labeled *normative responsibility* and *responsibility as agency*. (For similar distinctions, see Bratman 1997, note 7.)

Normative responsibility establishes a normative relation between an agent and a specific action or outcome. This relation is cast either as *duty* ("Mulder and Scully are responsible for investigating every paranormal event in the country") or as *liability* (liableness to blame[3]) ("That constitutes child neglect and the parents should be held responsible!"). Duty and liability are normative in that they refer to social rules or expectations that dictate what the agent should do or should have done (Hamilton 1978). Responsibility as duty is typically used in a forward-looking manner to direct the agent's future actions ("What are your responsibilities at your new job?"). One *has* such responsibilities because they are *given* or *assumed*. Responsibility as liability is typically used in a backward-looking manner to respond to negative outcomes ("Police and public officials are not responsible for the attack"). One is *held* responsible or *accepts* respon-

sibility for those outcomes (which usually means that one is blamed or accepts blame). In practice, the two normative relations are often blended, as pre-existing duties are considered when assigning liability (Haidt and Baron 1996; Hamilton 1978).

Responsibility as duty presupposes the capacity for intentional action. Assigning specific duties to a person is pointless unless one assumes that the person can intentionally fulfill them. Responsibility as liability, too, presupposes the capacity for intentional action: Liability is assigned when the agent could and should have acted so as to prevent the outcome but didn't (Hamilton and Sanders 1992; Weiner 1995). "Could have" corresponds to an assumption of preventability or intentional controllability; "should have" corresponds to an assumption of duty. Liability does not, however, presuppose factual intentionality—agents can be liable to blame even for outcomes they did not bring about intentionally but rather caused through negligence or recklessness (Duff 1990; Hart 1968). Intentionality amplifies liability, to be sure, and lack of intention can ameliorate it (Heider 1958; Schlenker, Britt, Pennington, Murphy, and Doherty 1994). But liability holds even for unintentionally caused outcomes, as long as the agent had the intentional capacity and duty to prevent that outcome.

Beginning with the classic work of Piaget (1932), research on children's developing moral reasoning has emphasized the transition from judgments based solely on outcome severity to judgments incorporating or even focusing on the agent's intentions and motives. This transition occurs in the preschool period (Nelson-LeGall 1985; Yuill and Perner 1988), apparently after a child begins to distinguish intentional from unintentional behavior (Shultz 1980) and to ascribe motives to agents (Bartsch and Wellman 1995a). Not surprisingly, it seems to take a child some time to learn to apply the intentionality concept to the new function of distinguishing a blameworthy from a blameless state of mind.

An even more complex step occurs when a child learns that sometimes even unintentional behaviors leave the actor subject to blame. In chapter 17, Michael Chandler, Bryan Sokol, and Darcey Hallett explore this intriguing developmental step. As was mentioned above, adults hold agents responsible when they are perceived to have been capable of preventing the behavior and to have a duty to do so. Chandler et al. demonstrate that by age 5 children begin to take these counterfactuals into account when making

judgments of blame. Furthermore, they argue that this advance in moral reasoning is tied to the onset of a "constructivist" theory of mind—one that fully appreciates that human agents interpret and construct what they know about the world but that their interpretations and their ensuing moral conduct can be more or less justified.

Responsibility as agency, the second principal meaning of responsibility, refers to an agent's general capacity to perform autonomous, rational action (e.g., "The successful performance of chores is another way that patients can demonstrate they are ready for greater financial responsibility"). To act with responsibility in this sense requires the ability to consider the consequences of one's actions and to choose an action with desirable consequences ("Taking responsibility means accepting the consequences of your own choices"). Responsible agency thus presupposes intentionality in the form of planning, deliberation, and reasoning (Bratman 1997; Hart 1968). Indeed, these planning features of intentionality define a particular version of responsible agency: acting with good judgment ("If you choose to drink alcohol, drink responsibly"). However, exactly what goes into the concept of responsible agency beyond the capacity for intentional action is still debated among philosophers (see, e.g., Bratman 1997; Fischer 1994; Wallace 1994). Proposed criteria include rationality, communicative capacity, and responsiveness to reasons. The choice of these criteria has great practical importance; for example, it affects sentiments and decisions about the responsibility of children and that of mental patients. However, there have been no empirical studies exploring what the criteria for responsible agency might be. Psychological research (see, e.g., Fincham and Jaspars 1980; Shaver 1985; Weiner 1995) has focused almost exclusively on the conditions under which people assign liability, leaving discussions of factors that constitute responsible agency to legal and philosophical scholars.

Normative responsibility, responsibility as agency, and intentionality are closely intertwined concepts. Responsible agency presupposes intentionality (the capacity to deliberate about and choose one's course of action). Normative responsibility in turn presupposes responsible agency (hence intentional capacity), for unless one considers an agent equipped with responsible agency in general one cannot assign to this agent any specific duties or any liability for specific outcomes. Thus, the folk-psychological assumption that humans are capable of intentional action underlies both

of the principal meanings of responsibility; as a result, judgments of intention and intentionality permeate the social practices of praise, blame, reward, and punishment (Marshall 1968; Williams 1993).

The fundamental role that the assumption of intentionality plays in responsibility is further highlighted when we consider the social role and function of responsibility attributions. The practice of assigning responsibility—in all its meanings—serves the coordination and organization of social activities, the maintenance of social order, and the enforcement of social rules (Heider 1944; Schlick 1966; Semin and Manstead 1983). Normative responsibility, in particular, lays the foundation for a social feedback system in which desirable outcomes yield a premium and undesirable outcomes are sanctioned. But this normative system applies only to those outcomes that are in principle controllable by intentional agency (excluding, for example, natural disasters) and involves only those agents who are equipped with responsible agency (excluding, for example, young children and some mental patients). This feedback system has often been discussed in theoretical terms, but few scholars have explored it in detail. In chapter 16, Bernard Weiner does exactly that, detailing some of the system's cognitive, emotional, and behavioral elements and using a model centered on the concept of controllability to account for moral sentiments (such as anger or pity), philosophies of punishment, and individual differences in political ideology.

Two chapters put intentionality and responsibility in their larger social and cultural contexts. In chapter 15, Ames, Knowles, Morris, Kalish, Rosati, and Gopnik try to integrate considerations of norms and context into the often purely cognitive models of social perception. They examine, in particular, how people's social and cultural contexts shape their judgments of intentionality and responsibility as well as their mental-state ascriptions and their behavior explanations. In chapter 18, Leonard Kaplan analyzes how conceptions of intentional agency and responsibility interrelate and, more important, how they are in tension with expectations of justice in the modern state. Kaplan examines several models of moral action and identifies their varying assumptions about intentional agency, responsibility, and justice. Because all these models center on responsibility as the ethical duty to be responsive to another's needs, classic issues of social cognition arise when the models specify to what extent an individual is capable of recognizing the

suffering of others and distinguishing it from deception or exploitation. Even though the particular ethical problem of enabling justice in the modern state must remain unsolved, Kaplan's analysis illustrates how ethical discourse presupposes the agent's capacity to act intentionally and to perceive and interpret the social world and other beings within it.

Conclusions

Theories and research programs on the role of intention and intentionality in social cognition are distributed over many scholars, traditions, and disciplines. These individual efforts, though united by the goal of elucidating interpersonal understanding, have often remained isolated from one another. A unifying theory of how humans understand other humans will have to emerge from communication and collaboration across the traditional boundaries of paradigms and disciplines. The research brought together in this volume, we hope, both attests to the fundamental role of intentionality in human social cognition and offers noteworthy progress toward a broad and interdisciplinary account of human social relations.

Acknowledgments

Preparation of this chapter was supported by a National Science Foundation CAREER award (No. 9703315) to Bertram Malle and by a National Science Foundation New Young Investigator Award (No. 9458339) and a John Merck Scholars Award to Dare Baldwin. The chapter was prepared while Dare Baldwin was a fellow at the Center for Advanced Study in the Behavioral Sciences; she is grateful for the financial support provided by the William T. Grant Foundation under award 95167795.

Notes

1. On the appropriateness of ascribing to such an agent only an intention to try to A, not an intention to A, see Mele 1989.

2. Quoted examples were found by searching the World Wide Web and various newspapers for sentences containing the words *responsibility* or *responsible*.

3. This meaning of responsibility is also labeled accountability, answerability, or blameworthiness in the literature. Its legal version is liableness for punishment.