Attributions as Behavior Explanations: Toward a New Theory

Bertram F. Malle, University of Oregon

Attribution theory has played a major role in social-psychological research. Unfortunately, the term *attribution* is ambiguous. According to one meaning, forming an attribution is *making a dispositional (trait) inference* from behavior; according to another meaning, forming an attribution is *giving an explanation* (especially of behavior). The focus of this paper is on the latter phenomenon of behavior explanations. In particular, I discuss a new theory of explanation that provides an alternative to classic attribution theory as it dominates the textbooks and handbooks—which is typically as a version of Kelley's (1967) model of attribution as covariation detection. I begin with a brief critique of this theory and, out of this critique, develop a list of requirements that an improved theory has to meet. I then introduce the new theory, report empirical data in its support, and apply it to a number of psychological phenomena. I finally conclude with an assessment of how much progress we have made in understanding behavior explanations and what has yet to be learned.

Attribution theory is a core element of social-psychological thinking. Thousands of articles have been published in 40 years of research, and textbook and handbooks of social psychology typically devote a chapter or large section to attribution phenomena. This body of research, however, can be distinguished into a *general approach* to social-psychological phenomena—an "attributional" analysis, as Bernard Weiner has called it—and *theories of specific attribution phenomena*, such as Kelley's (1967) theory of explanation or Jones and Davis' (1965) theory of dispositional inference. The general attributional approach recognizes that humans try to make sense of their surroundings and themselves and that this sense-making activity (explanations, finding meaning, creating stories) is an integral part of the social phenomena under investigation. This approach has made countless contributions to the literature, shedding light on achievement motivation, responsibility judgments, helplessness, sleeplessness, obesity, depression, emotion, and well-being research (e.g., Abramson, Seligman, & Teasdale, 1978; Jones, Kanouse, Kelley, Nisbett, Valins, & Weiner, 1972; Schwarz, & Clore, 1983; Weiner, 1995), and it has become a central part of social psychology as a whole.

Attribution theories, by contrast, are theories of more specific attribution phenomena. Unfortunately, the term *attribution* is quite ambiguous. According to one common meaning, forming an attribution is *giving an explanation* (especially of behavior); according to another common meaning, forming an attribution is *making a dispositional (trait) inference* from behavior (Hamilton, 1998; Hilton, Smith, & Kin, 1995; Malle, 2004). Even though explanations and trait inferences are occasionally related, they are distinct in many ways. For example, explanations sometimes refer to traits but often do not; trait inferences can be explanatory but usually are not; traits can be inferred from any behavior, whereas explanations are triggered only

by surprising or confusing behavior; explanations are answers to why questions, trait inferences are not.

My focus in this chapter is the phenomenon of behavior explanations.[1] In particular, I discuss a new theory of explanation that provides an alternative to classic attribution theory as it dominates the textbooks and handbooks—which is typically as a version of Kelley's (1967) model of attribution as covariation detection. I begin with a brief critique of this theory and, out of this critique, develop a list of requirements that an improved theory has to meet. I then introduce the new theory, report empirical data in its support, and apply it to a number of psychological phenomena. I finally conclude with an assessment of how much progress we have made in understanding behavior explanations and what has yet to be learned.

Classic Attribution Theory

Kelley's (1967) original theory—and many others after him (Cheng & Novick, 1990; Fiedler, Walther, & Nickel, 1999; Försterling, 1989; Hewstone & Jaspars, 1987)—made two core claims:

> (PS) The causal concepts on which people rely when forming behavior explanations consist of a dichotomy of internal vs. external (or person vs. situation) causes.
>
> (COV) The cognitive process that underlies explanations is covariation analysis.

When we examine each of these core claims of attribution theory in turn, we see that there is shockingly little support for either. First, what is the empirical evidence for claim (PS)? In most attribution studies, the truth of (PS) was assumed, not tested. Participants had to fill out scales for "person/disposition" and "situation" causes (e.g., Storms, 1973) or researchers classified free-response explanations into person and situation categories (e.g., McGill, 1989; Nisbett, Caputo, Legant, & Marecek, 1973). In neither case was assumption (PS) falsifiable. The few attribution studies that did not automatically make the person-situation assumption found several other dimensions of explanation to be of greater importance than that of person and situation (e.g., Fletcher, 1983; Lewis, 1995; Passer, Kelley, & Michela, 1978). Likewise, substantial research in developmental psychology, which did not make the person-situation assumption, found no evidence for the truth of (PS). The data show that even 3-4 year old children distinguish between intentional and unintentional behavior and explain the latter by referring to the agent's beliefs and desires he or she had for acting (e.g., Bartsch & Wellman, 1989; Wellman et al., 1997).

If there is no empirical evidence for (PS), is there at least good theoretical reason to believe in (PS)? Unfortunately, no. The person-situation assumption is not derived, for example, from a model of people's conceptual framework of behavior; in fact, it runs counter to pertinent analyses in the philosophical literature on human action and action explanation (e.g., Davidson, 1963; Mele, 1992; Searle, 1983; Mischel, 1969), which distinguish between

intentional and unintentional behavior and identify a unique mode of explaining intentional behavior in the form of the agent's reasons. A simple example should suffice for now to illustrate that the person-situation dichotomy just doesn't capture the nature of people's explanations of intentional action. Consider this scenario:

> Having just arrived in the department as a new Assistant Professor, Pauline finds in her mailbox a note that says "Let's have lunch tomorrow. Faculty club at 12:30? *–Fred*." Pauline is a bit surprised. She met Fred W. during her interview, but she wouldn't have expected him to ask her out for lunch.

Pauline now tries to explain Fred's action of leaving the note in her mailbox. Kelley's attribution model would claim that Pauline's choice is between a person attribution (something about Fred caused the action) and a situation attribution (something about herself or the circumstances caused the action). But right away, this is a confusing choice. Surely something about Fred *must* have been causing the action (his intention, his motivation…) if putting the note in her mailbox was intentional. But of course, the situation figured into the action as well, or at least the situation as seen by Fred—he may have thought Pauline would like to have some company, or expected her to be an ideal collaborator. Either way, a satisfying answer for Pauline will never be "it was something about Fred" or "something about the situation." What the search for person-situation attributions misses entirely is what the explainer *actually* does when faced with a situation like this. Pauline will simply try to find out Fred's *reasons* for leaving the note—his specific goals, beliefs, and assumptions. A theory of behavior explanation must therefore incorporate the concept of reasons into its theoretical repertoire.

Finally, what is the historical basis for (PS)? Here, we encounter two major misunderstandings. To begin, Lewin (1936) as one historical source of the person-situation dichotomy, meant it as a sketch of the *reality* of social behavior—that scientists can start out with the assumption that behavior is a function of the person and the situation (including all their complex interplays). But Lewin at no point argued that *ordinary people* see social behavior in this way.

What is perhaps more surprising is that Heider (1958), the most widely cited historical source for (PS) didn't claim either that lay people divide the world into person and situation causes (Malle & Ickes, 2000). Instead, Heider argued that the fundamental distinction people make when trying to explain events in the social world is between "personal causality" and "impersonal causality." What he referred to with these terms are distinct causal models that ordinary people bring to social perception (Heider, 1958, pp. 100-101). One causal model is applied to the domain of intentional behavior, for which people assume the involvement of an intention as the critical force that brings about the action. The other causal model is for all other domains (i.e., unintentional human behavior as well as physical events), in which causes simply bring about effects—without any involvement of intentions.

This confusion between Heider's distinction of personal and impersonal causality (or intentional and unintentional behavior) on the one hand and the traditional person-situation

dichotomy on the other is not just a curious historical accident[2]; it had massive theoretical consequences. Attribution theories after Heider ignored the intentional-unintentional distinction and built models that applied to all behaviors alike. But it was precisely Heider's (1958) point that not all behaviors are explained the same way. He specifically stated that, whereas unintentional behaviors were explained simply by causes, intentional actions were explained by the "reasons behind the intention" (Heider, 1958, p. 110; see also pp. 125-129). But even in 1976, around the peak of attribution research, Heider observed that explanations of intentional action by way of reasons had not been adequately treated in contemporary attribution work (Ickes, 1976, p. 14). Sadly, nothing much seems to have changed in this regard, if we take social psychology textbooks and major surveys of attribution research as barometers.

Perhaps social psychology has held on to the simplified model of person-situation attribution because, for a long time, there was no alternative available? This can't be quite right, because alternative viewpoints have been voiced repeatedly (e.g., Buss, 1978; Lalljee & Abelson, 1983; Locke & Pennington, 1982; Read, 1987; White, 1991). It is true, however, that these alternative viewpoints did not resolve the contradictions between the various models and did not provide an integrative theory of behavior explanation. Such an integrative theory is what I hope to offer in this chapter, but first I briefly discuss the second core claim of classic attribution theory.

Kelley's (1967) claim that covariation analysis underlies the construction of lay explanations (COV) is problematic as well. First off, the covariation claim is poorly supported empirically. The available evidence shows that people can *make use of* covariation information when it is presented to them by the experimenter (e.g., Försterling, 1992; McArthur, 1972; Sutton & McClure, 2001; Van Kleeck, Hillger, & Brown, 1988). But there is no evidence that people *spontaneously search for* covariation information when trying to explain behavior. In fact, very few studies even examined whether and when people actively seek out covariation information in natural contexts. As a rare exception, Lalljee, Lamb, Furnham, and Jaspars (1984) asked their participants to write down the kind of information they would like to have in order to explain various events, and covariation information was in low demand under these conditions. A few additional studies examined people's choices between receiving covariation information and some other information, and there, too, explainers were less interested in covariation information than in information about generative forces or mechanisms (Ahn, Kalish, Medin, & Gelman, 1995).

The theoretical foundation for (COV) is dubious as well. The notion of covariation analysis was a creative analogy to scientific and statistical reasoning, but it wasn't grounded in any model of either cognitive inference or causal learning. The covariation thesis also contradicts what we know about behavior explanations as communicative acts (i.e., one person's clarification for another person; Hilton, 1990; Kidd & Amabile, 1981; Turnbull & Slugoski, 1988). In constructing communicative explanations, the speaker's choice of a particular causal factor is guided, not by covariation analysis, but by impression management (i.e., selecting a

cause that puts the agent or explainer in a certain evaluative light; Tedeschi & Reiss, 1981) and audience design (i.e., selecting a cause that meets the listener's wondering or expectation; Slugoski, Lalljee, Lamb, & Ginsburg, 1993). So even if there are some contexts in which covariation analysis is important, it is clearly not the only cognitive process by which explanations are constructed.

Besides the lack of support for its two core claims, classic attribution theory and its successors has two additional limitations. For one thing, it treats explanations as a purely cognitive activity, so there is no accounting for the social functions of explanation, such as clarifying something for another person or influencing an audience's impression's. Moreover, classic attribution theory does not specify any psychological factors besides raw information that influence the construction of explanations. Specifying these factors would allow us to predict such important phenomena as actor-observer asymmetries, self-serving biases, and the like.

Demands on a New Theory of Explanation

The difficulties with standard attribution theory imply a number of desired features that a new theory of explanation must have. First, instead of allowing a reduction to person and situation causes, the theory has to capture the concepts that actually underlie people's thinking and reasoning about human behavior, such as agency and intentionality.

Second, the new theory must identify additional cognitive processes besides covariation analysis that are recruited to construct explanations. It should also begin to specify the conditions under which one or the other cognitive process are used.

Third, the theory has to integrate the social-communicative aspect of explanations with the cognitive one. It must be made clear in what way the social and the cognitive aspects of explanation are tied together and in what respects they differ.

Fourth, the new theory has to identify psychological factors that govern the construction of explanations, processes that can be used to predict actor-observer asymmetries and related phenomena of behavior explanation.

An Alternative: The Folk-Conceptual Theory of Explanation

A theory of behavior explanation that my colleagues and I have developed appears to meet these demands and may be able to supersede attribution theory as an account of people's behavior explanations (Malle, 1999, 2001; Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000). I call it the *folk-conceptual theory of behavior explanation* because its basic assumptions are grounded in people's folk concepts of mind and behavior.

The theory has three layers. The first layer concerns the **conceptual framework** that underlies behavior explanations (and helps meet the first demand specified above). The starting point is Heider's insight that people distinguish sharply between intentional and unintentional behavior and conceptualize intentional behavior very differently from unintentional behavior (Malle, 2001; Malle & Knobe, 1997a). As I will show in more detail, this conceptualization

implies three distinct modes of explaining intentional behavior, along with a fourth mode of explaining unintentional behaviors, as well as distinct explanation types within each mode (Malle, 1999). For example, the mode of reason explanations breaks down into three reason types: beliefs, desires, and valuings.

The next layer of the theory concerns the **psychological processes** that govern the construction of explanations (helping to meet the second and fourth demand.). There are two different challenges people face: The first is to choose among their various explanatory tools (i.e., modes and types of explanations). The three factors that influence those choices are: features of the behavior to be explained (e.g., intentionality, difficulty), pragmatic goals (e.g., impression management, audience design), and information resources (e.g., stored information, perceived action context). The second problem in constructing explanations is that people must select *specific* reasons, causes, etc. (not just "a belief reason" or "a situation cause"), and they do so by relying on a number of cognitive processes separately or jointly (e.g., retrieving information from knowledge structures, simulation, projection, rationalization, and occasionally covariation analysis).

The third layer of the theory is a linguistic one, which identifies the specific **linguistic forms** people have available in their language to express behavior explanations. (This layer helps meet the third and fourth demand.) Some of these linguistic forms can be usefully exploited by people to use explanations as a tool of social influence, such as to distance oneself from an agent's reason (e.g., Why did she refuse dessert?—"Because she's been gaining weight" *vs.* "Because *she thinks* she's been gaining weight"; Malle et al., 2000).

The three layers can be depicted in a hierarchy (see Figure 1) that considers the conceptual framework as the foundation, the psychological processes as operating on this foundation, and the linguistic layer as operating on both layers underneath. I now develop the first two layers in detail and report supporting empirical evidence.

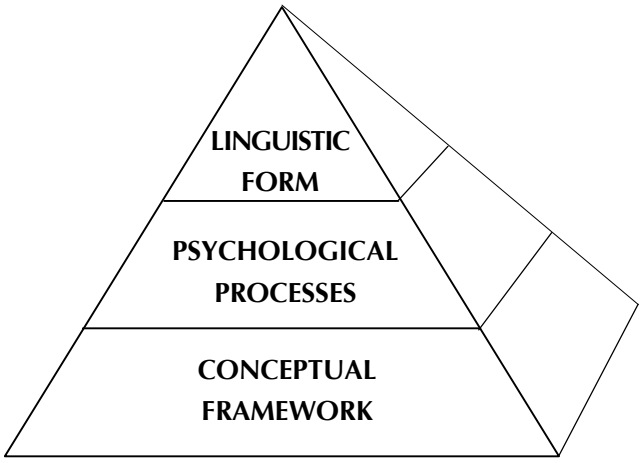

Figure 1. Three layers of the folk-conceptual theory of behavior explanation

The Conceptual Framework of Behavior Explanations

Traditional attribution theory postulated a very simple set of concepts that was supposed to underlie lay behavior explanations. There were effects (behaviors, outcomes, or events) and causes, and those causes were classified into personal (dispositional or internal) and situational (external) types.[3] This framework is wholly incompatible with what we know about children's emerging theory of mind and behavior—the conceptual network that already 4-5 year-olds have when interpreting and thinking about human behavior (Perner, 1991; Wellman, 1990; Gopnik & Meltzoff, 1997). We see there the importance of a concept of intentionality, of mental states contrasted with observable behaviors, and of specific mental states such as intentions, beliefs, and desires that are used to explain intentional behavior. It is rather unlikely that people forget these concepts and distinctions when they grow up and, instead, begin to explain behavior in terms of person and situation causes. Even though the attribution framework was criticized repeatedly for omitting mental-state concepts such as reasons, goals, or motives (e.g., Buss, 1978; Read, 1987, White, 1991), no comprehensive revision of the attribution framework has yet been offered.

The folk-conceptual theory of explanation takes seriously the complex network of assumptions and distinctions that underlie people's thinking about behavior—whether in early childhood or adulthood—and thus integrates important concepts such as reasons and goals into a revised model of how people explain behavior. The theory can be formulated in terms of a number of conceptual postulates.

Intentionality

The first conceptual postulate of the theory is that when people deal with human behavior, they distinguish sharply between intentional and unintentional behavior (Heider, 1958; Malle, 1999; White, 1991). Social perceivers show a high level of agreement ($\alpha = 0.99$) in their intentionality judgments (Malle & Knobe, 1997a, Study 1), and they do so by relying on a shared folk concept of intentionality. This concept normally includes five requirements for a behavior to be judged intentional: It must be based on an intention, on skill, and awareness, and the intention itself must be based on appropriate beliefs and desires (Malle & Knobe, 1997a, Studies 2-4). Subsequent studies showed that an intention is seen as a commitment to act that flows form a reasoning process (Malle & Knobe, 2001). In this reasoning process, the agent weighs a number of beliefs and desires and settles on a course of action. The intention does not guarantee, of course, that the intended action will be successfully performed; the agent's skill and facilitating circumstances are required for a successful outcome.

Reasons vs. Causes

The second conceptual postulate is that people explain intentional behavior differently from the way they explain unintentional behavior.[4] Specifically, whereas unintentional behavior is explained by causes, intentional behavior is primarily explained by reasons (Buss, 1978;

Davidson, 1963; Donellan, 1967; Locke & Pennington, 1982; Malle, 1999; Mele, 1992; Read, 1987; Searle, 1983). Reasons and causes are both seen as "generating factors," but reasons are a unique kind of generating factor. They are representational mental states, that is, states like beliefs and desires that represent a specific content (*what* is believed or *what* is desired). For beliefs or desires to be the agent's reasons, their content had to be (in the explainer's eyes) part of a reasoning process that led the agent to her decision to act. When an explainer claims that "Anne invited Ben to dinner because he had fixed her car," then the explainer must presume[5] that Anne actually considered Ben's fixing her car and *for that reason* invited him to dinner. The notion of a "reasoning process" doesn't require that the agent goes through an extended and full-fledged deliberation; but it does require, in people's folk theory of mind, that (a) the agent considered those reasons when deciding to act and (b) regarded them as grounds for acting. These two conditions (which I have called subjectivity assumption and rationality assumption; Malle, 1999, 2004) define what it is to be a reason explanation. It isn't enough for some mental states to be grounds for acting if the agent wasn't aware of (didn't consider) them when deciding to act—those mental states may be good reasons in general, but they weren't the *agent's reasons* (Malle et al., 2000). Likewise, it isn't enough for some mental states to be on the agent's mind while deciding to act if she doesn't regard them as grounds for acting—again, they wouldn't be the reasons for which she decided to act.

Causes, of course, don't have to meet a subjectivity or rationality requirement. There can be unconscious or irrational causes; all that counts is that the cited causes are presumed to be factors that brought about the behavior in question. Because unintentional behavior presupposes neither an intention nor awareness on the part of the agent, the way by which causes bring about unintentional behavior is independent of the agent's reasoning and will. In that sense, causes are "impersonal," as Heider (1958) put it.

To illustrate the difference between reasons and causes, consider the following two explanations:

> (1) Lee was nervous about the math test because she wanted to be the best in class.
>
> (2) Lee studied for the math test all day because she wanted to be the best in class.

In the first case, the desire to be the best caused Anne's nervousness; but that desire didn't figure as part of a reasoning process, nor did Anne regard it as grounds for being nervous. In fact, it's possible Anne wasn't even aware of her desire to be the best. The situation is very different in (2). To understand this explanation is to assume Anne decided to study in light of her desire to be the best, regarding such as desire to be grounds for studying all day. Anne's awareness of her own reason, and the rational role it plays in her decision to study are implied by reading explanation (2) as a reason explanation.

Other Modes of Explaining Intentional Behavior

Reasons are the default mode by which people explain intentional actions (with a frequency of about 70%). But at times (and I will clarify shortly at what times), people use one of two alternative explanation modes. One such alternative is to explain actions not with the agent's reasons but with factors that preceded those reasons and presumably brought them about (see Figure 1). Whereas reasons capture what the agent herself weighed and considered when deciding to act, causal history explanations capture the various causal factors that led up to the agent's reasons. These causal history of reason (CHR) explanations literally describe the causal history, origin, or background of reasons (Malle, 1994, 1999; see also Hirschberg, 1978, Locke & Pennington, 1982), and such a history could lie in childhood, cultural training, traits, or in situational cues that triggered, say, a particular desire.

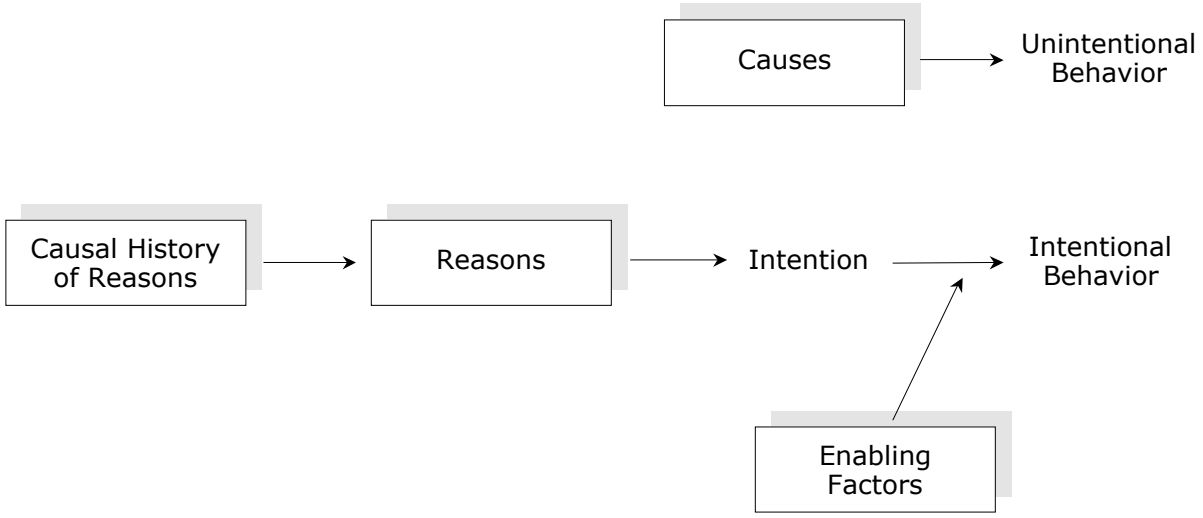

Figure 2. Four modes of folk explanation, with arrows indicating perceived causal connections

If we wanted to offer a CHR explanation for Lee's studying for a test all day, we might say, "She is achievement-oriented" or "She comes from a family of academics" or "That's typical in her culture." CHR explanations can cite something about the agent, even a trait, or something about the situation—but the "locus" of the cause can vary widely. What defines an explanation as a causal history explanation is (a) that it explains an intentional action and (b) that it clarifies why the person decided to act as described, and (c) that it isn't a reason explanation. Condition (c) implies that neither the subjectivity nor the rationality assumption holds for CHR explanations. In the eyes of the explainer, Lee didn't reason, "I am achievement-oriented, therefore I should study all day" or "It's typical in my culture to study all day, so therefore I will too." Causal history factors exert their causal power independently of the agent's awareness of

those powers and outside of the agent's own reasoning about what rationally favors her action. In many cases, we can construct a chain such that the agent's reasons are what led her to act, and the CHR factors are what brought about or strengthened those reasons (see Figure 2). It's likely, for example, that one's desire reason to be the best is a result of one's being achievement-oriented or coming form a family of academics.

There is one more alternative mode of explanation for intentional behaviors. This one doesn't clarify what motivated the agent to act (as reasons and causal histories do) but rather what enabled the action to succeed, hence we call it *enabling factor explanation*. Whereas reason explanations and CHR explanations explain both the agent's action and her intention (even before she implements the action), enabling factor explanations apply only when the action was actually completed. What is explained, then, is how it was possible that the action succeeded, and the explainer cites important causal factors (e.g., abilities, effort, opportune circumstances) that presumably helped turn the intention into a successful action (see Fig. 1). For example, if we wonder how it was possible that Lee, in fact, studied all day for the math test, we might say that "she'd made a big pot of coffee," or if we wondered how she could complete the eventual test in just 15 minutes, we might say that "she had one of those new calculators" or "she works very efficiently."

People's conceptual toolbox for explaining behavior thus contains four modes of explanation: one for unintentional behavior (causes), and three for intentional behavior (reasons, causal histories, enabling factors). These modes of explanation can be reliably discriminated when coding naturally expressed behavior explanations (with $\kappa > .80$; see F.Ex, 2002, for details), and we have identified many of the psychological conditions under which each mode occurs. Even though these conditions are in the purview of layer 2 of the folk-conceptual theory of explanation, I'd like to foreshadow those results.

Obviously, cause explanations are used whenever the explainer considers the behavior as unintentional. If the explainer considers a behavior as intentional, a further choice has to be made between the default option of citing one or more reasons or an alternative modes of explanation. CHR explanations supplant reasons either when the explainer doesn't have enough information to construct a reason or has pragmatic goals that are better met by CHR explanations (e.g., weakening blame for negative actions), or when the behavior to be explained is a trend (across time or across different people) that is more parsimoniously explained by a causal history factor (a factor that may underlie most or all specific reasons within the action trend). Enabling factor explanations supplant reasons or causal histories when the action in question was particularly difficult or got accomplished despite countervailing forces, and when the question asked about an action is *how it was possible*, not *why the person decided to perform it* (Malle et al., 2000; McClure & Hilton, 1997, 1998).

The Nature of Reasons

The fourth conceptual postulate consists of a set of specific claims about the types and features of reasons. As already mentioned, reasons are representational mental states. We can therefore distinguish between the specific mental state that is cited in the explanation and the content of that state, and that distinction yields three classifications of reason features. (1) On the mental state side, there are three classes of states that can function as reasons: *beliefs* (including knowledge, thinking, etc.), *desires* (including wants, needs, and tryings, etc.), and *valuings* (including likes/dislikes, enjoyment, etc.). (2) On the content side, we can classify contents into the traditional person-situation categories (e.g., "…because he wanted *a car"* [situation]; "…because he wanted *to be rich*" [person]) or into alternative categories, such as desirable-undesirable. (3) When examining the linguistic form of reason explanations, finally, we see that some reasons are formulated with mental state markers—verbs such as "think," "believe," "want," "need," or "like" that indicate explicitly what kind of mental state the reason is—whereas other reasons lack such markers. Table 1 contrasts pairs of reason explanations that are either in their marked or unmarked form.

Table 1. *Reasons in Their Marked and Unmarked Form*

Behavior	Reason type	Marked form	Unmarked form
Why did they sell their car?	Belief	*They felt* it was too small for the family	It was too small for the family
Why did he go to the coffee shop?	Desire	*He wanted* to have a real Italian espresso	To have a real Italian espresso
Why did she stay until after 10?	Valuing[6]	*She liked* the show.	The show was fun.

Thus, when we examine reasons in detail, we find three features along which reasons can vary: the type of mental state the reason is; the content of that reason; and the presence or absence of mental state markers. We can now ask what psychological functions and processes are associated with each of these three features.

The choice between belief reasons and desire reasons has at least two psychological functions. First, belief reasons are favored over desire reasons when the explainer tries to present the agent in a rational light (Malle et al., 2000). Second, belief reasons are more dominant over desire reasons among actors than among observers. Our data so far suggest that this asymmetry is not a result of actors presenting themselves in a more desirable (e.g., rational) light, because a manipulation of such impression management did not alter belief-desire reason frequencies either among actors or observers (MacCionnaith, 2003). Instead, we have evidence that the belief-desire asymmetry is due to differential information resources in actors and

observers, because when we examine observers who know the agent well (and have been present at the time of the action), the asymmetry disappears: Observers provide no fewer belief reasons than the agents themselves do (Malle, Knobe, & Nelson, 2003).

Mental state markers may appear to be a trivial linguistic variation, but they have several important functions as well. Pairs of explanations such as those in Table 1 not only differentiate the two ingredients of reasons—the type of mental state the reason is (belief, desire, or valuing) and the content of that reason; they also demonstrate that reason explanations always refer to the agent's mental states, even if their linguistic surface doesn't explicitly mention a mental state (cf. Gordon, 2001). That is, an explanation such as "My father never lets us go out *because something might happen to us"* refers to the father's *belief* that something might happen to his daughters, and therefore he never lets them go out. Without the theoretical categories we have proposed, one might falsely assume that the explanation refers to some objective situation cause that could be subsumed under the classic person-situation categories. But nothing actually has happened at the time of the explanation, so the objective situation cannot itself be the cause of the father's action. Instead, the father *thinks* that something might happen, and everybody who hears or reads the explanation will infer that he does.

Why might explainers omit mental state markers? For one thing, omitting a mental state marker makes the reason *sound* more objective and true. By stating that "something might happen," the daughter refers to a potential reality that appears to justify the father's action. Conversely, adding a mental state marker allows explainers to distance themselves from an agent's reason. By stating that the father never lets them go out "because *he thinks* something might happen," the daughter would indicate disagreement with the father's belief and cast some doubt on its plausibility.[footnote on a third possibility: child before false-belief phase]

Consider another example, which we presented to 91 undergraduate students (Malle et al., 2000, Study 6). Cliff and Jerry are at a dinner party. Cliff asks Jerry, "Why did your girlfriend refuse dessert?" Jerry responds by saying either "She thinks she's been gaining weight" (marked belief) or "She's been gaining weight" (unmarked belief). After reading the vignette, participants rated (on a scale from 0 to 8) how happy Jerry was with his girlfriend's current weight. As predicted, Jerry was seen as happier with his girlfriend's weight when he used the marked belief ($M = 5.4$) than when he used the unmarked belief ($M = 2.6$), $F(1, 88) = 21.9$, $p < .01$, $\eta^2 = 20\%$.[7]

Reason contents, finally, have not proven to carry any clear psychological function, at least as long as they are classified according to traditional person-situation categories. For example, actors and observer do not differ in their contents of reasons (Malle et al., 2003); explanations of group actions and individual actions do not differ in their contents of reasons (O'Laughlin & Malle, 2002); and impression management or self-servingness do not have a reliable impact on reason contents. There may well be a psychological function associated with other aspects of reasons contents (such as their social desirability), but this possibility remains to be investigated.

In sum, reasons have a complex conceptual and linguistic structure that is not reducible to any traditional causal categories. To understand reasons is to understand their nature as representational mental states and their resulting three features: the type of mental state they are (belief, desire, valuing), the content of that state, and the linguistic form as being marked or unmarked. At least two of these features are associated with important psychological functions or processes.

Types of Causal Factors

The fifth conceptual postulate of the folk theory of explanations concerns the types of causes, causal histories, and enabling factors that people construct. These explanatory modes all refer to causal factors that can in principle be classified by their locus, following the person-situation dichotomy in classic attribution research. However, that dichotomy was ambiguous in that the person category sometimes referred to stable traits, usually labeled "dispositions," whereas at other times it referred to all causal factors internal to the agent, whether stable or not. A more precise way to classify these causal factors is to use the label *person* as an overarching category that refers to all causal forces inside the agent and reserve the word *trait* for person factors that are stable parts of the agent's personality. That way we break up the causal forces into two orthogonal contrasts: person vs. situation and, among person factors, traits vs. nontraits.[8] Supporting this finer classification, we have found that actors and observers do not differ in their use of person vs. situation causes when explaining unintentional behavior, but they do differ in their use of trait vs. nontrait person causes. Beyond this actor-observer asymmetry (which is restricted to unintentional behaviors and only when explainers know the agent well) we have identified no predictive validity of either the person-situation or the trait-nontrait classification, either in studies comparing explanations of groups and individuals (O'Laughlin & Malle, 2002) or in studies exploring impression management (MacCionnaith, 2003; Malle et al., 2000).

These postulates about the folk-conceptual structure of behavior explanations suggest a complex picture in which explainers have to choose (consciously or not) between multiple modes of explanations, different types within each mode, and alternative linguistic forms. Attribution theory after Heider failed to distinguish between all these different explanatory tools, confounding intentional and unintentional behavior, collapsing four distinct modes of explanation (causes, reasons, causal histories, and enabling factors) into one "causal attribution," and ignoring many finer-grained types and forms of explanation. Many of these modes, types, and forms of folk explanations have predictable psychological significance, as in actor-observer asymmetries, group-individual differences, and impression management. By contrast, the classic attribution categories of person/disposition and situation show very limited predictive power, because they simply do not carve up explanatory phenomena at their joints. (I will explore later how it was possible that attribution research nevertheless uncovered a number of interesting findings.)

The question I now turn to is which psychological processes guide the choice among the multiple explanatory tools, constituting the second layer of the folk-conceptual theory of explanation.

Psychological Processes

When examining the psychological processes that guide folk explanations of behavior, we have to separate two challenges the folk explainer faces. For one thing, he must choose a tool from the large toolbox of explanatory modes, types, and forms (e.g., a marked belief reason; a trait enabling factor). In addition, he must provide a *specific* instance of any explanatory tool. No ordinary explanation stops at the level of conceptual categories—one cannot explain an everyday behavior by saying "She had a reason" or "There was some trait enabling factor." Instead, folk explanations of behavior must be tailored to the agent, action, and context, so that an action like "moving the furniture" is explained by a specific reason such as "*because she expected a lot of people for the party*" or a specific trait such as *anxiousness*.

Of these two problems—the choice of an explanatory tool and the provision of a specific explanation—the first is scientifically more tractable. In fact, it is unlikely that a psychological theory will ever predict the precise contents that an explainer provides in an explanation. But the process of *searching* for such contents may well be predictable, and this issue promises to be an intriguing domain for future research. Before I sketch this future research, however, I summarize what we know about the first problem, the choice of explanatory tools.

Determinants of Choosing Explanatory Tools

Our research findings on people's differential use of explanation modes, types, and forms, is best accounted for by three factors: attributes of the explained behavior; pragmatic goals; and available information resources.

1. Behavior attributes. Before explaining a given behavior, social perceivers make several (often implicit) judgments about that behavior. To begin, they judge the behavior's *intentionality* and, as a result, either provide cause explanations (for unintentional behavior) or reasons, causal history, or enabling factor explanations (for intentional behavior). The evidence unequivocally supports this pattern. In one study (Malle, 1999, Study 2), a group of participants made intentionality judgments for 20 behaviors whereas a second group of participants offered explanations for these same behaviors. The behaviors' judged intentionality correlated at $r = -.92$ with the probability of cause explanations given for these behaviors (Malle, 1999, Table 1).

A second important attribute is the *difficulty* of intentional behaviors. If the behavior is considered difficult to accomplish, the explainer will often provide enabling factors; otherwise, he is likely to choose reasons or causal history explanations (Malle et al., 2000; McClure & Hilton, 1997, 1998).

A third attribute is whether the social perceiver explains a singular behavior or a behavior trend (across time or agents). If the behavior is judged to be a trend, the rate of CHR

explanations increases significantly, compared to singular behaviors (O'Laughlin & Malle, 2002). That is because each behavior within the trend may have a different reason explanation, and citing those reasons would be extremely cumbersome. One or two causal history factors may suffice to indicate the background that triggered the full array of differential reasons. For example, a mother who was asked to explain why she went shopping many times a week answered: "Because I have three children." The series of actions in question is parsimoniously explained by offering the causal history of having three children because it underlies the variety of specific reasons she has for shopping each time (e.g., buying more milk, a new supply of diapers, or a special carpet cleaner for crayon stains).

2. Pragmatic goals. When social perceivers explain behaviors in communicative contexts, they have a variety of smaller or larger projects they try to accomplish with their explanations, such as to lessen another person's confusion, manage their own status in the interaction, or fend off blame. Two groups of goals can be distinguished. In *audience design*, the explainer tailors the explanation to the audience's needs and existing knowledge (Hilton, 1990; Slugoski et al., 1993) A clear case of such design is when the explainer matches an explanation mode to the type of question asked (Malle et al., 2000; McClure & Hilton, 1998). The question by which someone requests an explanation can inquire either about the agent's immediate motivation ("What did she do that for?") and demand a *reason* explanation; or it can inquire about the background of that motivation ("How come?") and invite a *causal history* explanation; or it can inquire about the factors that enabled successful action performance ("How was it possible that she did that?"), demanding an *enabling factor* explanation.

In *impression management*, the explainer is engaged in an act of social influence, using the behavior explanation to create certain beliefs, perceptions, or actions in the communication partner. For example, people increase their use of causal history explanations when accounting for negative actions (Nelson, 2003); they increase their use of belief reasons when trying to appear rational (Malle et al., 2000); and they explicitly add a mental state marker to their belief reasons when they want to distance themselves from the agent (e.g., "Why is he looking at apartments?"—"He thinks I am moving in with him"; cf. Malle et al., 2000).

3. Information resources. Different explanation modes and types have different information demands. Reason explanations, for example, require relatively specific information about the agent, the behavior, and the context, whereas causal history explanations and enabling factor explanations may get by with less context-specific information. Similarly, belief reasons often require fairly idiosyncratic information about the agent's deliberations, whereas desire reasons can sometimes be constructed from the nature of the behavior alone. Supporting this difference between desire and belief reasons, we found that observers normally provide fewer belief reasons (and more desire reasons) than actors do. However, when observers know the agent and/or were present when the action took place, their rate of belief reasons increases to equate that of actors (Malle et al., 2003).

The three determinants of explanatory choice along with the conceptual nature of modes and types of explanations jointly provide the theoretical basis to predict and account for a number of important phenomena. Elsewhere, we have successfully applied this approach to impression management in explanation (Malle et al., 2000), to the difference between explaining group and individual behaviors (O'Laughlin & Malle, 2002), and to the variety of asymmetries between actor and observer explanations of behavior (Malle et al., 2003). [More?] Other domains are open to further investigation, such as relationships, negotiation, psychopathology, and cross-cultural comparisons (see Malle, 2004).

Constructing Specific Explanations

The process of constructing explanations with a specific content has remained largely unexplored in 40 years of attribution research. One reason for this omission was that standard attribution models tried to predict only whether an explainer would give a "person attribution" or a "situation attribution." By describing explanatory work at such a general level, the process of constructing specific explanations doesn't come up (cf. Kruglanski, 1979). Another reason for the omission was the assumption that explainers use only one cognitive process to arrive at their attributions, namely, covariation analysis. Unfortunately, this assumption was never adequately supported, as all such tests showed merely that people can *respond to* covariation information if it is presented by the experimenter. The few studies that examined whether people spontaneously search for covariation information in more natural contexts cast serious doubt on the ubiquity of covariation analysis (Ahn, Kalish, Medin, & Gelman, 1995; Lalljee, Lamb, Furnham, & Jaspars, 1984).

Several critics of standard attribution theory have proposed cognitive processes that exist alongside covariation analysis and help the explainer to construct specific explanations. First, explainers recruit event-specific, agent-specific, or general *knowledge structures* (Abelson & Lalljee, 1988; Ames, 2003; Lalljee & Abelson, 1983; Read, 1987). Second, they use the two related processes of *simulation* (i.e., imaginative representation of the agent's mental states; Goldman, 1989, 2001; Gordon, 1986, 1992; Harris, 1992) and *projection* (i.e., assuming that one's own mental states is the same as the agent's; Ames, 2003; Krueger & Clement, 1997; Van Boven & Loewenstein, 2003). In addition, there are two principles that aim knowledge recruitment and simulation into certain directions: (a) the "method of difference," which contrasts the event in question with an alternative event and tries to identify the critical difference (e.g., Cheng & Novick, 1990; Hilton & Slugoski, 1986; Kahneman & Miller, 1986; McGill, 1989) and (b) a premium on identifying generative forces or mechanisms (Ahn et al., 1995; Ahn & Kalish, 2000; Cheng, 2000). These two principles do not themselves generate specific explanations but put constraints on knowledge recruitment and simulation.

Elsewhere I have applied this set of proposed processes to naturally occurring explanations and developed hypotheses about the relationship between particular processes and

particular explanatory tools (Malle, 2004). Because of space constraints I can only summarize the main results of this exploration, grouped by the four modes of folk behavior explanations.

Cause explanations. From the observer perspective, when the explainer has little familiarity with the agent and/or did not directly observe the unintentional behavior, reliance on general knowledge, including stereotypes and cultural scripts is likely to be high (Ames, 2003). As familiarity increases, and especially when the explainer directly observes the behavior, the use of simulation and projection will increase. From the actor perspective, knowledge structures will be dominant (especially recall of events immediately preceding the unintentional behavior), but for private wonderings about recurring and puzzling experiences, covariation analysis may be recruited, such as when one wonders about a recurrent headache.

Causal history explanations. For both perspectives, the predominant process in generating CHR explanations is the recruitment of knowledge structures relevant to the context, the agent, or the action. Simulation processes may come into play when an observer searches for potential experiential causal history factors, and covariation analysis will become dominant when the explainer (from either perspective) searches for a common causal history behind a trend of actions.

Enabling factor explanations. The construction of enabling factor explanations, too, relies primarily on specific or generic stored knowledge (e.g., about the kinds of facilitating forces that enable particular actions), but in domains of achievement (e.g., grades, sports victories), covariation analysis can becomes relevant as well. Simulation is largely absent when constructing enabling factor explanations, because people cannot easily simulate abilities, opportunities, or other facilitating forces.

Reason explanations. When selecting reason explanations, actors never rely on covariation calculation. When being truthful, they have (or believe they have) access to the reasons that initially prompted their intention, relying on a process of direct recall (Brewer, 1994; Cowan, 1995; Russell & D'Hollosy, 1992). This process should get decreasing use, in favor of general knowledge structures, the less deliberation went into the action (because the memory trace for the action's reasons is weak) and the longer the time span between action deliberation and the explanation (because the memory traces may have washed out). When actors alter their explanation for impression management purposes, they will also not use covariation analysis but rather recruit reasons from knowledge structures that best meet their impression goals. Observer may occasionally use covariation analysis when they wonder about an agent's repeated choice among well-defined options, but in most cases of singular actions, knowledge structures and simulation/projection will be dominant.

Besides testing these hypotheses about processes involved in constructing specific explanations, future research might also examine the interrelationship between these construction processes and the more general determinants of explanatory choice (i.e., behavior attributes, pragmatic goals, and information resources). In some cases, the determinants will first favor a particular mode and type of explanation and then favor a construction process suitable for this

mode and type of explanation. In other cases, the determinants may directly favor a construction process, which in turn provides a specific explanation content, cast as a particular mode and type.

Accounting for Human Explanations of Behavior

For a long time, the decisive word on lay behavior explanations came from attribution theory. In this chapter, I have tried to convince the reader that this decisive word was wrong. Not always, but often. When dealing with explanations of unintentional behaviors and outcomes, attribution theory provides a fine conceptual framework, though some improvements were recommended here as well (e.g., the distinction between traits and other person causes, and the multiple cognitive processes to construct specific explanations). Where attribution theory fails is in dealing with explanations of intentional behaviors, both at the conceptual level and at the level of psychological processes.

The folk-conceptual theory of behavior explanation identifies the conceptual framework that underlies lay explanations of intentional behavior, including the key role of intentionality and the resulting distinctions between modes of explanation (reasons, causal histories, enabling factors) and their specific features (e.g., beliefs, desire, mental state markers). This first, conceptual layer of the new theory—directly tested and supported in recent work (Malle, 1999; Malle et al., 2000)—precisely describes the tools people use to explain behavior and brings order to the complexity of naturally occurring behavior explanations. This layer also unites two aspects of explanation that are often separated in the literature: explanations as cognitive (private) events and explanations as social (public) acts (Malle & Knobe, 1997b; Malle, 2004). Despite their different implementations, antecedents, and consequences, these two kinds of explanations are built from the same conceptual framework. This framework specifies the modes and types of explanation that are available both for the private and the public explainer.

A second layer of the folk-conceptual theory concerns the psychological processes that give shape to explanations as cognitive and social acts. For one thing, three primary psychological determinants (judged behavior features, pragmatic goals, information resources) guide people's choices of modes and features of explanation. Moreover, explainers must select specific contents of explanations in specific situations, and they do so by relying on a variety of cognitive processes, including knowledge structures, direct recall, simulation, and covariation analysis.

The folk-conceptual theory not only describes but also *accounts for* many of the regularities of behavior explanations, from their conditions of occurrence to some of their specific functions (e.g., Malle, 1999; Malle et al., 2000). Furthermore, we have been able to show that the concepts and distinctions in this new theory have predictive power when it comes to investigating impression management (Malle et al., 2000), asymmetries between individual and group targets (O'Laughlin & Malle, 2002), asymmetries between actor and observer perspectives (Malle et al., 2003), and self-servingness (Nelson & Malle, 2003). And whenever we analyzed the explanation data in terms of a person-situation attribution model, virtually no

such predictive power was found. These findings suggest that the folk-conceptual theory better captures the breakpoints, or joints, in the framework that underlies people's behavior explanations.

Previous attempts at replacing attribution theory were not very successful, as the textbooks still describe mostly Kelley's covariation model, Jones' trait inference model, and person-situation attributions by perspective, self-servingness, and the like. Perhaps this is because the person-situation distinction is so seductively simple and so easily researched in the lab (using a pair of rating scales) that the field has been reluctant to abandon this tried approach in favor of any alternative, especially a more complex model of explanation. But good science must study the phenomena as they exist, and that is where attribution theory's greatest weakness lies. In imposing a conceptual framework on people's folk explanations that just isn't *people's own* framework, much of attribution research has provided data that are simplified, difficult to interpret, and have led to false conclusions.

But how can it be that previous research on attribution phenomena failed to uncover its own limitations? I believe that this failure arose from two methodological biases. First, participants were typically asked to express their explanations on pre-defined person/situation rating scales rather than in the form of natural verbal utterances. As a result, people had to transform their complex explanatory hypotheses into simple ratings, which probably invited guessing strategies as to how the ratings were to be interpreted and surely led to severe ambiguities in the data that ensued. A high "person" rating, for example, could have indicated a confident judgment of intentionality, a reason explanation, a person causal history factor, and many more.

Second, in the few cases in which free-response explanations were analyzed, the coding was very limited, often picking up no more than trends in the linguistic surface of explanations, such as in the use of mental state markers (McGill, 1989; Nisbett et al., 1973; for evidence and discussion see Malle, 1999, Study 4, and Malle et al., 2000, Study 4).

These serious problems of ambiguity and misinterpretation apply to attribution research that examined intentional behavior, because in confounding it with unintentional behavior and subsuming explanations for both behaviors under the person-situation dichotomy, the portrayal of intentional behavior explanations became seriously distorted. Research that was focused entirely on unintentional events remains largely valid. For example, the analysis of self-serving biases in explaining achievement outcomes does not involve reason explanations, so the person-situation dichotomy captures these explanations reasonably well. However, nothing that was found about these biases can be extended to explanations of intentional events. In fact, recent research using the folk-conceptual theory suggests that the degree of self-servingness, and the tools of achieving it, when people explain intentional behavior differ significantly from the classic picture (Nelson & Malle, 2003).

Other classic attribution findings were assumed to apply to both unintentional and intentional behavior, such as the actor-observer asymmetry. But when we examined, in a series

of studies, actor and observer explanations for both unintentional and intentional behaviors (Malle et al., 2003), very little remained valid about the classic Jones and Nisbett (1972) thesis. The person-situation dichotomy predicted no actor-observer asymmetries; the most consistent asymmetries held for the choice between reasons and causal histories, beliefs and desires, and marked vs. unmarked beliefs; and the only finding that supported classic claims was that observers used more trait explanations than actors did, but only for unintentional behavior and only when they knew the agent well (Knobe & Malle, 2002; Malle et al., 2003).

The shortcomings of traditional attribution theory extend to the process level as well. Many of the classic findings in attribution research were not accounted for by reference to identifiable psychological processes, such as the ones I enumerated—behavior attributes, pragmatic goals, and information resources. Also, the central proposition that explanations are constructed from covariation assessments (Kelley, 1967) has garnered no supportive evidence except for demonstrations that people can *respond to* covariation information if provided by the experimenter. I have suggested, instead, that people rely on multiple psychological processes to construct explanations, including retrieval of general and specific knowledge, mental simulation, and occasional covariation analysis. This is an issue that has yet to be settled by empirical research, but the textbook tenet that explanations are constructed from covariation assessments is almost certainly false in its general form.

Despite my critical analysis, I recognize of course that classic attribution research has contributed much to social psychology. It posed questions and pointed to phenomena that had simply not been considered before—such as the power of behavior explanations (Heider, 1958; Jones et al., 1972; Quattrone, 1985); the many interesting factors that create systematic variations in explanation, such as actor-observer differences (Jones & Nisbett, 1972), self-servingness (Bradley, 1978; Heider, 1958; Miller & Ross, 1975), and impression management tactics (Tedeschi & Reiss, 1981); and the larger network of cognitive and social antecedents and consequences of behavior explanations (Anderson, Krull, & Weiner, 1996).

But these impressive results and insights of attribution research emerged *in the context* of attribution theory, not as *predicted results* of that theory. What attribution theory does explicitly predict is that people form explanations as ascriptions of person (disposition) causes vs. situation causes and do so on the basis of covariation assessment. Nothing in this theory predicts that there must be actor-observer asymmetries, much less that these asymmetries are of a particular kind (Knobe & Malle, 2002), and nothing predicts impression-management tactics, a self-serving bias, or other interesting phenomena. The most celebrated insights and findings of attribution research were developed under the influence of creative early models of attribution, but it is time that we allow theoretical advances to both account for existing and predict new findings. The folk-conceptual theory is one attempt to make such advances.

References

Abelson, R. P., & Lalljee, M. (1988). Knowledge structures and causal explanations. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 175-203). Brighton: Harvester.

Abramson, L. Y., Seligman, M. E. P., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87*, 49-74.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54,* 299-352.

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 199-226). Cambridge, MA: MIT Press.

Ames, D. R. (2003). *Mental state inference in person perception: Everyday solutions to the problem of other minds.* Manuscript under review.

Anderson, C. A., Krull, D. S., & Weiner, B. (1996). Explanations: Processes and consequences. In E. T. Higgins, & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 271-296). New York: Guilford Press.

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

Brewer, W. F. (1994). Autobiographical memory and survey research. In N. Schwarz & S. Sudman (Ed.), *Autobiographical memory and the validity of retrospective reports* (pp. 11-20). New York: Springer.

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

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365-382.

Cheng, P. W. (2000). Causality in the mind: Estimating contextual and conjunctive power. In F. C. Keil and R. A. Wilson (Eds.), *Explanation and cognition* (pp. 227-253). Cambridge, MA: MIT Press.

Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.

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

Donellan, K. S. (1967). Reasons and causes. In B. Edwards (Ed.), *Encyclopedia of philosophy* (Vol. 7, pp. 85-88). New York: Macmillan.

F.Ex (2002). *Coding scheme for people's folk explanations of behavior* (Version 4.1). Retrieved April 20, 2003 from darkwing.uoregon.edu/~bfmalle/fex.html

Fletcher, G. J. O. (1983). The analysis of verbal explanations for marital separation: Implications for attribution theory. *Journal of Applied Social Psychology*, *13*, 245-258.

Försterling, F. (1989). Models of covariation and attribution: How do they relate to the analogy of analysis of variance? *Journal of Personality and Social Psychology, 57*, 615-625.

Försterling, F. (1992). The Kelley model as an analysis of variance analogy: How far can it be taken? *Journal of Experimental Social Psychology, 28*, 475-490.

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

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

Gordon, R. M. (1992). The simulation theory: Objections and misconceptions. *Mind and Language, 7*, 11-34.

Gordon, R. M. (2001). Simulation and reason explanation: The radical view. *Philosophical Topics, 29,* 175-192.

Hamilton, D. L. (1998). Dispositional and attributional inferences in person perception. In J. M. Darley & J. Cooper (Eds.), *Attribution and social interaction: The legacy of Edward E. Jones* (pp. 99-114). Washington, DC: American Psychological Association.

Harris, P. (1992). From simulation to folk psychology: The case for development. *Mind and Language*, *7*, 120-144.

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

Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A logical model of the intuitive analysis of variance. *Journal of Personality and Social Psychology, 53,* 663-672.

Hilton, D. J., Smith, R. H., & Kin, S. H. (1995). Processes of causal explanation and dispositional attribution. *Journal of Personality and Social Psychology*, *68*, 377-387.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*, 75-88.

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.

Ickes, W. (1976). A conversation with Fritz Heider. In J. H. Harvey, W. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1, pp. 3-18). Hillsdale, NJ: Erlbaum.

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., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79-94). Morristown, NJ: General Learning Press.

Jones, E. E., Kanouse, D., Kelley, H. H., Nisbett, R. E., Valins, S., & Weiner, B. (Eds.). (1972). *Attribution: Perceiving the causes of behavior.* Morristown, NJ: General Learning Press.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136-153.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 129-238). Lincoln: University of Nebraska Press.

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.

Knobe, J., & Malle, B. F. (2002). Self and other in the explanation of behavior: 30 years later. Special issue on self-other asymmetries: *Psychologica Belgica*, *42,* 113-130.

Krueger, J., & Clement, R. W. (1997). Estimates of social consensus by majorities and minorities: The case for social projection. *Personality and Social Psychology Review*, *1*, 299-313.

Kruglanski, A. W. (1979). Causal explanation, teleological explanation: On radical particularism in attribution theory. *Journal of Personality and Social Psychology*, *37*, 1447-1457.

Lalljee, M., & Abelson, R. P. (1983). The organization of explanations. In M. Hewstone (Eds.), *Attribution theory : Social and functional extensions* (pp. 65-80). Oxford: Basil Blackwell.

Lalljee, M., Lamb, R., Furnham, A. F., & Jaspars, J. (1984). Explanations and information search: Inductive and hypothesis-testing approaches to arriving at an explanation. *British Journal of Social Psychology, 23,* 201-212.

Lewin, K. (1936). *Principles of topological psychology.* (F. Heider & G. M. Heider, Trans.). New York: McGraw-Hill.

Lewis, P. T. (1995). A naturalistic test of two fundamental propositions: Correspondence bias and the actor-observer hypothesis. *Journal of Personality*, *63*, 87-111.

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

MacCionnaith, K. (2003). *Accounting for actor-observer asymmetries in explanation: The role of impression management.* Unpublished Senior Honors' Thesis, University of Oregon.

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

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. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Cambridge, MA: MIT Press.

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. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101-121.

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. (2003). *Actor-observer asymmetries in folk explanations of behavior: 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.

McArthur, L. Z. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, *22*, 171-193.

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.

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.

McGill, A. L. (1989). Context effects in judgments of causation. *Journal of Personality and Social Psychology*, *57*, 189-200.

Mele, A. R. (1992). *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.

Mischel, T. (1969). *Human action: Conceptual and empirical issues.* New York: Academic Press.

Nelson, S., & Malle, B. F. (2003). *Self-serving biases in explanations of behavior.* Manuscript in preparation.

Nelson, S. E. (2003). *Setting the story straight: A study of discrepant accounts of conflict and their convergence.* Unpublished Doctoral Dissertation, University of Oregon

Nisbett, R. E., Caputo, C., Legant, P., & Marecek, J. (1973). Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology*, *27*, 154-164.

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

Passer, M. W., Kelley, H. H., & Michela, J. L. (1978). Multidimensional scaling of the causes for negative interpersonal behavior. *Journal of Personality and Social Psychology, 36,* 951-962.

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

Quattrone, G. A. (1985). On the congruity between internal states and action. *Psychological Bulletin, 98,* 3-40.

Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality & Social Psychology*, 52, 288-302.

Russell, E. W., & d'Hollosy, M. E. (1992). Memory and attention. *Journal of Clinical Psychology*, *48*, 530-538.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513-523.

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

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.

Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality Social Psychology*, *27,* 165-175.

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.

Turnbull, W., & Slugoski, B. (1988). Conversational and linguistic processes in causal attribution. In D. J. Hilton (Ed.), *Contemporary science and natural explanation* (pp. 66-93). Brighton, Sussex: Harvester Press.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, *29*, 1159-1168.

Van Kleeck, M. H., Hillger, L. A., & Brown, R. (1988). Pitting verbal schemas against information variables in attribution. *Social Cognition*, *6*, 89-106.

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

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. M. (1990). *The child's theory of mind*. 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.

Endnotes

[1] Because of this focus, I will not discuss Jones and Davis (1965) correspondent inference theory. This theory has had a major impact on social psychology, but it does not represent a theory of behavior explanation. For detailed arguments why it does not, see Malle (2004, chapter 1) and also Hamilton (1998).

[2] Part of the blame for this accident may accrue to Heider himself. As his 1958 book was conceived and written over several decades, Heider wasn't entirely consistent in his use of terms. In one section, for example, he speaks of causes in the person and in the environment, and it seems that he actually made the person-situation dichotomy claim (Heider, 1958, pp. 82-84). But, in fact, Heider's analysis there concerned only one particular mode of explanation—when a social perceiver tries to make sense of an "action outcome" and wonders how it could be accomplished (e.g., a weak man rowing across the river; a rookie pitcher getting 12 strike outs in a row). In this situation, the perceiver is not interested in clarifying the agent's motivation for acting but rather wonders *how it was possible* that the agent accomplished the desired action outcome. When explaining such accomplishments, Heider argued, the social perceiver considers two elements: the agent's attempt to perform the action (*trying*) and supporting factors (*can*), which lie in the agent (e.g., ability, confidence) or in the environment (e.g., opportunity, luck, favorable conditions). Heider thus catalogued here the "conditions of successful action" (p. 110), which serve as explanations of accomplishments, as answers to how-possible questions. In this catalogue, Heider made use of the person-situation (or internal-external) dichotomy; but he never claimed that all lay explanations of behavior are organized around a split into person and situation causes. For further details on this historical analysis, see Malle and Ickes (2002) and Malle (2004, chapter 1).

[3] Additional cause types were postulated in the domains of achievement, responsibility, and depression, namely, stable-unstable, specific-global, and controllable-uncontrollable. But these distinctions were never clearly integrated into a theory of causal attributions.

[4] The more precise way of speaking would be to use the term *event* instead of *behavior.* Intentional events include not only observable actions, such as to write, make a phone call, or turn the radio on, but also unobservable mental states, such as to decide, calculate a price, or imagine a new carpet. Likewise, unintentional events include observable behaviors such as fidgeting, tripping, or a spontaneous frown, and also unobservable mental states, such as feeling sad, hearing a dog bark, or having a flashback (see Malle & Knobe, 1997b; Malle & Pearce, 2001). For simplicity, I use the term *behavior* to refer to any of these events.

[5] In communicative settings, explainers may of course lie and merely try to convince their audience that the agent acted for the reason stated. But such lying is only possible against

the background of the normative case in which the explainer actually presumes that the agent acted for the reasons stated.

[6] Among valuings, unmarked forms are extremely rare. Moreover, the unmarked forms cannot be created by omitting the mental state verb (as with most beliefs and desires); instead, unmarked valuings are expressed by an evaluative claim about the content of the reason (e.g., "it's fun!").

[7] I should note that these effects of distancing oneself from (or embracing) an agent's reasons operate reliably only with belief reasons. Unmarked belief reasons have no surface indicator that they are a reason (and therefore can give the impression of a reality that underlies the agent's action), whereas unmarked desire reasons (e.g., "To lose weight," "so she'll stick to her diet") still have the grammatical structure of desires. As a result, unmarked desire reasons cannot easily create a different impression than marked desire reasons.

[8] In our classifications of free-response behavior explanations we routinely distinguish a variety of subclasses (e.g., among nontrait person factors: behaviors, mental states, group memberships, etc.) to explore any predictive validity of such classes. (See F.Ex, 2002).