1 History: Past Research on Attribution and Behavior Explanation

Many reviews have been written about the productive and well-known area of attribution theory (e.g., Shaver 1975; Fiske and Taylor 1991; Försterling 2001; D. T. Gilbert 1998; Hastorf Schneider, and Polefka 1970; Kelley and Michela 1980; Ross and Fletcher 1985; Weary, Stanley, and Harvey 1989). Some of these reviews incorporate critical thoughts about the classic theories but, by and large, they represent the "standard view" of attribution theory, both in terms of its history and its substantial claims. A composite sketch of this standard view looks something like this:

1. Heider (1958) argued that people try to identify the dispositional properties that underlie observed behavior and do so by attributing behavior either to external (situational) or internal (dispositional) causes.

2. Jones and Davis (1965) built on Heider and focused on the conditions under which people observe an agent's behavior and either do or do not attribute a correspondent disposition to the agent. Later, other researchers elaborated on Jones and Davis's approach by studying the precise cognitive processes that underlie dispositional attributions.

3. Kelley (1967) theorized in detail about the information processing people engage in when explaining social events. His model describes the rational analysis of patterns of covariation among three elements—a *person* acting toward a *stimulus* in particular *circumstances*—and derives the conditions under which people make attributions to the person or the stimulus.

4. In studying attributions for achievement outcomes, Weiner and colleagues (1972) found that people rely not only on the person–situation dimension of causality but also on the dimensions of stability and controllability, and these three-dimensional causal judgments mediate some of people's emotions and motivations in response to social outcomes.

In the present chapter, I contrast this standard view with quotes and interpretations of the classic attribution works by Heider (1958), Jones and Davis (1965), and Kelley (1967), complemented to a smaller extent by analyses of some of the more recent work on explanations. Because there are so many excellent reviews available on the standard view, I will spend relatively little time recounting it. My goal is, rather, to point out those aspects of prior attribution theories that are not generally emphasized. In so doing I will point to what seem to me to be historical misunderstandings and theoretical difficulties that have not been adequately resolved. The subsequent chapters then introduce a theory of behavior explanations that builds on previous theories but also tries to remove some of their difficulties and, in particular, tries to be a genuine theory of behavior explanations, not one of trait inferences, causal judgments, or responsibility ascriptions.

1.1 Attribution as Perception: Heider

Psychological research on attribution began with the work of Fritz Heider, who developed models of attribution for both object perception and person perception. His theory of object perception (first described in his 1920 dissertation) is rarely cited today, but it serves as the foundation for his later theory of person perception, so I will briefly review it (see also Malle and Ickes 2000).

1.1.1 Object Perception

Heider's early theorizing was an attempt to solve one of the core philosophical problems of phenomenology: the relation between sense qualities and real objects. That is, Heider asked how it was possible that we take sense qualities to be qualities of objects, given that sense qualities are "here" in the mind, whereas object qualities are "out there" in the physical world. Heider's answer began with the distinction between *things* (physical objects) and the *media* by which things affect the perceiver (Heider 1920, 1925; see also 1959). For example, a ticking watch (thing) causes systematic air vibrations around it (medium), which in turn engage the eardrum (another medium) and lead to perception. Heider argued that media have a considerable degree of variance but are shaped by the relative invariance of things. The perceptual apparatus reconstructs things from their effects on the me-

dia, and Heider termed this reconstructive process *attribution.* In Heider's theory of object perception, then, attribution generates representations of the relatively invariant qualities of things from the characteristic variances they cause in their media. Perceivers faced with sensory data thus see the perceptual object as "out there" because they attribute the sensory data to their underlying causes in the world (Heider 1920).[1]

1.1.2 Person Perception, Dispositions, and Personal Causality

After his early work on object perception Heider turned to the domain of social interactions, wondering how people perceive each other in interaction and, especially, how they make sense of each other's behavior. Heider proposed that a process of attribution is also involved in *person perception,* but he recognized that person perception is more complex than object perception—due to the manifold observational data available and the manifold causes (e.g., beliefs, desires, emotions, traits) to which these data can be attributed. In addition, it was clear to Heider that persons are targets of perception very different from inanimate objects. Persons are "perceived as action centers and as such can do something to us. They can benefit or harm us intentionally, and we can benefit or harm them. Persons have abilities, wishes and sentiments; they can act purposefully, and can perceive or watch us" (1958, p. 21). Heider repeatedly refers here to the intentionality of persons, which he considered a core assumption in the conceptual framework that underlies social perception. With the help of such concepts as intentionality and the inference of wishes, purposes, sentiments, and other internal states, Heider argued, perceivers bring order to the massive stream of behavioral data.

Even though, in one sense, person perception is like object perception—a process of extracting invariance out of variance—Heider saw crucial differences between the two (and these differences are often glossed over by Heider interpreters). The first difference is that in the social domain, *variance* refers to the agent's behavior and *invariance* refers to the inferred perceptions, intentions, motives, traits, and sentiments, which are all relatively invariant against the stream of ongoing behavior. Subsequent attribution scholars often focused on only one type of invariance, namely traits, because they interpreted Heider's notion of *disposition* as referring to stable personality factors (i.e., traits, attitudes, or abilities).[2] But it was the agent's

motives that occupied a special role in Heider's model: "The underlying causes of events, *especially the motives of other persons,* are the invariances of the environment that are relevant to [the perceiver]; they give meaning to what he experiences" (1958, p. 81; emphasis added). But even though Heider (1958) occasionally referred to traits and abilities when talking about dispositions (e.g., pp. 30, 80), he considered "motives, intentions, sentiments . . . the core processes which manifest themselves in overt behavior" (p. 34). In the study of social perception, then, Heider's terms *disposition* and *invariance* referred primarily to mental and motivational states, and the practice in social psychology of considering dispositions to be stable traits is quite different from Heider's original theory.

The second crucial difference between person perception and object perception is that when people perform a causal (i.e., "attributional") analysis of human behavior, their judgments of causality follow one of two conceptual models (ibid., chap. 4). The first is a model of *impersonal causality,* applied to unintentional human behaviors (such as sneezing or feeling pain) as well as physical events (such as stones rolling or leaves falling). The second is a model of *personal causality,* applied only to human agents who perform an intentional action (such as cleaning the kitchen or inviting someone to dinner). "Personal causality," Heider wrote, "refers to instances in which p causes x intentionally. That is to say, the action is purposive" (ibid., p. 100).

1.1.3 Heider and the Person–Situation Distinction

Ensuing attribution research set aside Heider's distinction between personal and impersonal causality and claimed instead that Heider had argued for a distinction between *person* (or internal) causes and *situation* (or external) causes of behavior.[3] That is, when people try to explain behavior, they attempt to find out whether the behavior was caused by factors internal to the person (e.g., mood, motives, personality) or by factors in the external situation (e.g., physical environment, other people). But even though Heider at times referred to these two classes of internal/person and external/situation causes (see below), the distinction he cared far more about was that between personal and impersonal causality, a distinction that refers to two kinds of behaviors (intentional and unintentional) and the different ways people think about them.

There are two problems associated with this misunderstanding that the personal–impersonal distinction is a person–situation distinction. First, not all internal causes fall under personal causality because, "unless intention ties together the cause–effect relations, we do not have a case of true personal causality" (Heider 1958, p. 100). Thus, those internal causes that do *not* involve intentions (e.g., tiredness, moods, emotions) belong to the impersonal class: "[E]ffects involving persons but not intentions . . . are more appropriately represented as cases of impersonal causality" (p. 101). A driving blunder due to tiredness, for example, was subsumed by Heider under impersonal causality but would be considered an internal or person factor within post–Heiderian attribution theory.

Second and far more important is the problem that researchers working with the person–situation distinction began to omit a major element of human social cognition: the distinction between intentional and unintentional behavior. Because it seemed so easy to classify *any* explanation as either referring to the person or the situation, researchers stopped tracking whether the behavior explained was actually intentional or unintentional—but that was exactly what the personal–impersonal distinction was supposed to capture. This omission mattered a great deal because the intentionality distinction plays an essential role in the interpretation and social control of behavior and in the evaluation of morality. The dichotomy between person and situation causes fails to capture all these important roles.

But why has it been believed that Heider proposed a person–situation dichotomy in attribution? One section in particular may have spawned this belief (ibid., 1958, pp. 82–84). There Heider endorsed Lewin's famous equation that characterizes any "action outcome" (the result of an action) as "dependent upon a combination of effective personal force and effective environmental force" (p. 82). As is clear from elaborations of this claim (pp. 83–87), Heider argued that for an action outcome to occur (which is sometimes just the performance of the action itself), there needs to be a concomitance of two elements: the agent's attempt to perform the action (*trying*) and supporting factors (*can*) that lie in the agent (e.g., ability, confidence) or in the environment (e.g., opportunity, luck, favorable conditions). Heider catalogued here the necessary elements that have to join together for an intentional action to succeed in producing its desired outcome—the "conditions of successful action" (p. 110). Consequently, when people point

to the presence of these elements, they clarify what enabled the action outcome to be attained (Malle et al. 2000; McClure and Hilton 1997). Such *enabling factor explanations* answer the specific explanatory question of *how it was possible* that an action outcome was attained (see chapter 3 of this book).

The section in question thus expresses Heider's belief that people distinguish between internal and external causal factors *when they explain how action outcomes were attained.* Answers to this question can indeed refer to either person factors (e.g., effort and ability) or situation factors (e.g., task difficulty and luck), but there is no indication in the text that Heider thought people use the internal–external distinction when explaining behavior in general. On the contrary, Heider stated that people explain *why* a person is acting by referring to the "reasons behind the intention" (1958, p. 110; see also pp. 125–129). The contrast between these two types of explanations can be illustrated with the following passage from Daniel Gilbert (1998, p. 96):

> If a pitcher who wishes to retire a batter (motivation) throws a burning fastball (action) directly into the wind (environmental influence), then the observer should conclude that the pitcher has a particularly strong arm (ability). If a batter tries to hit that ball (motivation) but fails (action), then the observer should conclude that the batter lacked coordination (ability) or was blinded by the sun (environmental influence).

The observer's reasoning in this passage is entirely focused on accounting for successful or failed outcomes; the question *why* the batter and the pitcher acted as they did is not answered by reference to either arm, wind, or sun. The why-question is in fact already answered by mentioning the pitcher's obvious desire to retire the batter and the batter's wish to hit the ball and get a run. As in virtually all cases of enabling factor explanations, in this case too, the question *why* the agent acted is not at issue (because the answer is obvious); what is at issue is the question *how the outcome was attained* (Malle et al. 2000).

In an interview with Bill Ickes (1976, p. 14), Heider explicitly distinguished between these two types of question, and hence between two types of explanation:

1 the attribution of *outcomes* to causal factors (i.e., enabling factor explanations);[4]

2. the attribution of *intentional actions* to the actor's motives (i.e., reasons for acting).

Heider himself never developed a model of motive attributions (or *reason explanations*), and he in fact felt that these explanations had not been adequately treated by contemporary attribution work (Ickes 1976, p. 14; see also Buss 1978; Fiske and Taylor 1991). What Heider did develop—in the passages and sections of his book that describe action attainment as a function of *trying* and *can*—was the core of a model of outcome attribution, and he felt that this issue was later advanced in Bernard Weiner's work (e.g., Weiner et al. 1972).

It seems likely, then, that scholars who claimed Heider proposed the external–internal dichotomy as the fundamental dimension of explanation in fact mistakenly applied Heider's model of outcome attribution to the domain of motive attribution or action explanation. The following passages from Hastorf et al. (1970) illustrate the confusion between the two types of explanation (indicated in square brackets):

> Presumably the outcomes of action are caused by some combination of personal characteristics and environmental forces [outcome attribution]. The person may have done something because he had to do it . . . or because he wanted to do it [action explanation]. (p. 64)

> When we infer that the combination of ability and effort was stronger than the external forces, we infer that internal causality was present [outcome attribution]. Only then do we say such things as "he did it because he wanted to" [action explanation]. (p. 89)

In both of these passages, the authors treat two different explanatory questions as if they were one and the same. The judgment whether "he did it because he wanted to" or "because he had to do it" clarifies the agent's motives for acting (by means of a reason explanation). These reasons can be given even before the agent tries to perform the action (because reasons explain the intention, whether or not it gets fulfilled). By contrast, the judgment as to whether ability, effort, or external forces enabled the action outcome clarifies how it was possible that the action outcome was attained (by means of an enabling factor explanation). Enabling explanations can be given only after the agent[5] tried to perform the action—if she succeeded, for example, one might say it was because of her ability.

Because these two explanation types—reason explanations (motive attributions) and enabling factor explanations (outcome attributions)—answer such different questions, it is unfortunate that the attribution literature after

Heider collapsed them into one (cf. Zuckerman 1978). What makes this collapse even more unfortunate is that only enabling factor explanations can be classified into the traditional internal–external (person–situation) scheme, whereas reason explanations make very different conceptual assumptions and have a very different linguistic surface (Malle 1999; Malle et al. 2000). Much confusion in the attribution literature resulted from this collapse, and I will propose an alternative theory in chapters 4 and 5.

1.1.4 Summary

The textbook view of Heider's attribution theory differs from the theoretical position Heider took in his 1958 book. Even though Heider's whole analysis was predicated on the distinction between *personal causality* (intentional events) and *impersonal causality* (unintentional events), he is consistently credited with introducing the person–situation dichotomy in attribution. Heider indeed claimed that people explain outcomes and all unintentional events by reference to causes (which can be located either in the person or the situation); but, more importantly, he claimed that people explain intentional events (cases of personal causality) by reference to reasons. The dichotomy between person and situation causes thus applies to some but not all modes of behavior explanation, with explanations of actions by reasons being the critical exception. Reason explanations, though very frequent in everyday life, were not treated in detail by Heider and, perhaps as a result, were long overlooked by attribution researchers.

1.2 Attribution as Trait Inference: Jones and Davis

Two years after its publication, Heider's (1958) attribution work was lauded in a book review by Harold Kelley (1960). However, attribution theory's launch toward public prominence came several years later, after Edward Jones and Keith Davis (1965) published their acclaimed "theory of correspondent inference."

1.2.1 Action Explanation versus Trait Inference

The first few pages of Jones and Davis's (1965) paper appeared to address just the issue that Heider had left open: exactly how people explain intentional action by means of motives and reasons. The authors wrote that their theory was attempting to account for:

- "a perceiver's inferences about what an actor was trying to achieve by a particular action" (p. 222);
- "the attribution of intentions" (p. 220);
- the process of finding "sufficient reason why the person acted" (p. 220).

These statements appear to usher in a theory of explanations for intentional action. And indeed, Jones and Davis's section I was entitled "The Naive Explanation of Human Action: Explanation by Attributing Intentions." There the authors argued that "the perceiver's explanation comes to a stop when an intention or motive is assigned that has the quality of being reason enough" (p. 220). However, page 220 was the only one Jones and Davis devoted to action explanations. In actuality, their chapter offered an account of the conditions under which perceivers infer traits (such as arrogance or dominance) from single behavioral events.[6] Even though the beginning of the chapter mentioned both inferences of intentions and inferences of dispositions (by which they specifically meant stable traits and attitudes, straying from Heider's broader use of the term), the chapter quickly developed an exclusive focus on traits and attitudes. Likewise, all of the empirical studies Jones and Davis reviewed in support of their theory featured trait ratings as dependent variables. Not surprisingly, then, the paper's summary section stated:

> To say that an inference is correspondent, then, is to say that a disposition is being rather directly reflected in behavior, and that this disposition is unusual in its strength or intensity. Operationally, correspondence means ratings toward the extremes of trait dimensions which are given with confidence. (Jones and Davis 1965, p. 264)

Jones and Davis thus sidestepped the social perceiver's task of inferring the agent's reasons for acting and instead provided a theory of inferring traits. As David Hamilton (1998) put it, "correspondent-inference theory was an important theory of how people make dispositional inferences, but not really a theory of how people make causal attributions" (p. 107). Why Jones and Davis moved from a theory of action explanation, promised in their introductory remarks, to a theory of trait inference is not entirely clear, but clues can be found in their decision to entitle the whole chapter "From Acts to Dispositions: The Attribution Process in Person Perception" and in their characterization of traits as that "toward which the perceiver presses in attaching significance to action" (Jones and Davis 1965, p. 222). Jones and Davis regarded trait inferences as the ultimate aim of the "attribution

process" and action perception in general, a position that would soon dominate the field (see, e.g., Shaver 1975).

1.2.2 A Saving Effort

Daniel Gilbert (1998) attempted, rather heroically, to extract more out of the Jones and Davis chapter than can be found there at first blush. Specifically, he tried to show that the theory of correspondent inference in fact accounts for people's explanations of action via intentions (even though Gilbert concurs with Jones and Davis that traits are ultimately what perceivers are after). To this end, Gilbert adopted Jones and Davis's uncommonly broad definition of *intentions* as referring to a "constellation of beliefs, desires, plans, and goals" (D. T. Gilbert 1998, p. 105). He also adopted Jones and Davis's two principles that guide diagnostic inferences about an actor's dispositions: the principle of noncommon effects (inferences reveal something about an agent if they rely on an action's effects that are unique to that action, not shared by alternative actions) and the principle of desirability (inferences reveal something about an agent if they rely on those action effects that are not obviously socially desirable). Finally, Gilbert applied the two principles to a simple action ("Why did Frank cross the room and turn on the television?") and argued that these principles would allow the perceiver to infer the actor's "intention," yielding an answer such as "because he wanted to watch the news."

Does this reconstruction salvage Jones and Davis's attempts to account for action explanations? I think not. First, attention to noncommon and undesirable effects will yield only one type of "intention," namely goals (because the principles are concerned only with desired or undesired effects of actions). This leaves out a major element in the explanation of action, namely references to beliefs, such as when Frank turned on the television because "he thought that the news was on." No analysis of act–effects can yield a straightforward belief reason explanation.

Second, the act–effects analysis works alright so long as the action in question is a choice between clearly demarcated options that have a manageable set of effects. But many human actions are not like that, which causes problems for the analysis. For one thing, we have to assume that the perceiver selects the agent's relevant options of acting from sheer infinite possibilities, but correspondent inference theory is silent on how this selection might work. In addition, we have to assume that the perceiver considers each po-

tential action's relevant effects, and here, too, the theory is silent on how this selection from another set of infinite possibilities might work.

Third, Jones and Davis's model of intention inferences (via noncommon and undesirable action effects) will typically yield only an answer to the question of *what* the person was doing, not *why* she was doing it—as the authors themselves point out (1965, pp. 222, 228). Granted, sometimes a redescription of a movement pattern in terms of action verbs (e.g., "He was walking toward the window") will be informative and hint at possible explanations, but it will not itself supply these explanations. That is, it will not answer the question "*Why* was he walking toward the window?" Subsequent models of correspondent inference (Gilbert and Malone 1995; Quattrone 1982; Trope 1986) also did not incorporate people's answers to why-questions. For example, in D. T. Gilbert's (1989) multistage model of attribution, the early process of intention inference is called "action identification" (what is the person doing?) and is not credited with explanatory force. The later stage is called "attributional," but it is concerned with either inferring or not inferring an extraordinary disposition—a process quite distinct from ascribing motives or reasons for why the person acted.

1.2.3 Summary

Jones and Davis (1965) introduced an important issue in social perception by asking under what conditions people infer traits from (single) behaviors. Their theorizing about these correspondent inferences was highly influential, leading to research on the "fundamental attribution error" (Ross 1977), stereotypes (e.g., Gilbert and Hixon 1991; Yzerbyt, Rogier, and Fiske 1998), and the cognitive underpinnings of impression formation (e.g., Gilbert and Malone 1995; Trope 1986). However, even under the most charitable reconstruction, Jones and Davis (or theorists in their wake) have not offered a theory of how ordinary people explain behavior, only how they infer traits from behavior—two processes that are just not the same (Hilton, Smith, and Kin 1995; Hamilton 1998).

1.3 Attribution as Causal Judgment: Kelley

Kelley's (1967) paper, "Attribution theory in social psychology," is generally considered the first systematic and general treatment of lay causal explanations. Kelley's self-ascribed goal in the paper was "to highlight some of the

central ideas contained in Heider's theory" (p. 192). Specifically, the two central ideas on which Kelley focused were:

1. In the attribution process "the choice is between external attribution and internal [. . .] attribution" (p. 194).
2. The procedure of arriving at these external or internal attributions is analogous to experimental methodology.[7]

Two issues require discussion here, one historical, the other substantive. The historical one concerns Kelley's claim that the two ideas just listed were indeed central to Heider's theory. The substantive issue concerns the claim that the two ideas together provide a strong foundation for a theory of behavior explanation.

1.3.1 The Historical Issue: Kelley Representing Heider

Evidence for the claim that Heider considered the attribution process a choice between external and internal causes appears strong if we consult secondary literature on attribution, but, as argued earlier, this appearance is based on a misrepresentation of Heider's theory. For Heider, the personal–impersonal distinction was more fundamental than the external–internal distinction because it identified two very different domains of causality. Only when there is no intention causing the event (i.e., in the case of impersonal/unintentional events and outcomes) does the external–internal dichotomy apply. Explanations of intentional action, by contrast, are based on the conceptual framework of personal causality, which involves intentionality and the agent's reasons.

Support for Kelley's second claim, that Heider considered the attribution process analogous to experimental methodology, lies in a quote from the very end of Heider's book (Heider 1958, p. 297), which is itself largely based on the section "Attribution of Desire and Pleasure" (ibid., pp. 146–160). But in that section Heider focuses entirely on the attribution of unintentional events (such as enjoyment); and to such unintentional events, both the external–internal distinction and the strategy of covariation assessment apply. Heider never claimed, however, that all behavioral events are explained that way. In particular, nowhere did he argue that the external–internal distinction and the strategy of covariation assessment provide a model of how people explain intentional action.

1.3.2 The Substantive Issue: Kelley's Theory of Behavior Explanations

Whether or not Kelley correctly represented Heider, the more important question is whether Kelley's theory accounts for people's explanations of behavior. As a starting point, consider the following example Kelley offers to illustrate the attribution process:

> Am I to take my enjoyment of a movie as a basis for an attribution to the movie (that it is intrinsically enjoyable) or for an attribution to myself (that I have a specific kind of desire relevant to movies)? The inference as to where to locate the dispositional properties responsible for the effect is made by interpreting the raw data (the enjoyment) in the context of subsidiary information from experiment-like variations of conditions. (Kelley 1967, p. 194)

This example features an actor's wondering about the meaning or explanation of enjoyment—an unintentional event. Indeed, throughout the chapter Kelley applies his attribution analysis to "*effects* such as experiences, sensations, or responses" (p. 196), "impressions" (p. 197), as well as arousal states and evaluative reactions (pp. 231–232). All of these events are unintentional, and the person–situation causal analysis applies quite well to this type of event—but to this type only. Kelley himself, it appears from the text, believed that his model also extended to the case of "inferring a person's intentions from knowledge of the consequences of his actions" (p. 196; see also p. 193), but no theory, empirical data, or examples clarify how this extension might work.[8] Of course, the absence of such clarification is not proof of its impossibility. So let me illustrate some of the difficulties one quickly runs into when using the person–situation dichotomy for intentional actions. Consider the following 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. (By assumption, Fred's action is intentional, so we rule out the possibility that Fred unwittingly put the note in the wrong mailbox.) What does Kelley's attribution model have to say about this situation? The theory would claim that Pauline's choice is between a person attribution (something about Fred caused the action) and a situation attribution (something about her or the circumstances caused the action). But right away, this is

a confusing choice. Surely something about Fred must have been present in order for him to put the note in her mailbox: motives, an intention, a fairly controlled movement—all inescapable implications of Fred's action being intentional. And so it goes with all such actions (Kruglanski 1975). Intentional actions ought to elicit person attributions because people perceive them as caused by the agent's intention and motives (D'Andrade 1987; Heider 1958; Malle and Knobe 1997a). Nonetheless, the situation probably played some role in Fred's choice as well—but the situation as *subjectively represented* by Fred: He wouldn't have put this note in Pauline's mailbox if he hadn't *expected* her to check her mailbox in due time and if he hadn't *thought* about the coordination of time and place for the lunch and had not *hoped* her response to the invitation to be positive.

Consequently, a theory that portrays explainers of intentional actions as making a choice between person and situation attributions is amiss. We need a theoretical instrument that captures the explainer's interpretation of the agent's considerations and deliberations that motivated his action. For if Pauline knew Fred's deliberations, she would at once understand and be able to explain why he wrote the note.

But perhaps we were too quick in dismissing Kelley's approach to the mailbox scenario. Is there not a sense in which the "experimental methodology" Kelley has in mind could prove useful? If so, Pauline would have to ask the three questions about consensus, distinctiveness, and consistency and thus arrive at a plausible explanation of Fred's action. But this will generate few answers if we play it by the book. No other faculty member has so far, on Pauline's first day on the job, left a note in her mailbox (low consensus). What can she conclude from that? Fred may have wanted to welcome her, or go out on a date with her, or discuss some common research ideas with her—there are just too many possibilities. All of these explanations might be labeled "person attributions," because they are possible goals/desires Fred had when leaving the note. But the inference of a person attribution is uninformative in this case. Pauline does not doubt that Fred had some goal; she wonders rather which goal Fred had.

Similar problems arise with other covariation questions. Does Fred perform this kind of action toward other people too? Pauline won't know, but assuming she finds out that this is the first time Fred did it (high distinctiveness), she learns only that his action has something to do with her, but

she still does not know *why* he did it. And if Fred has left this kind of note with other people as well (low distinctiveness), Pauline merely learns that Fred shows some habit, which is also of limited use. She would want to know specifically whether his habit is to invite all new faculty members, or only women, or members of her research area, and so on. Systematic collection of such covariation information (if available) may at times prove helpful in constructing explanations of another's actions. But when it does, the explainer will not try to choose between person versus situation attributions but rather infer specific goals, beliefs, and the like, that were—in the explainer's assessment—the reasons for the agent's action.

Over the years, Kelley's covariation model and its refinements were tested empirically and appeared to receive reasonable support (e.g., Cheng and Novick 1990; Försterling 1992; McArthur 1972).[9] However, all that these tests showed was that people can take covariation information into account when it is made available to them; none of the tests showed that people actually seek out covariation information on their own (cf. Ahn et al. 1995). People may seek out covariation information for such unintentional events as headaches or moods and such outcomes as success or failure. But covariation reasoning about person and situation causes is surely not the exclusive process by which people go about explaining behavior, and it is actually quite ineffectual in the case of explaining intentional actions. (For a continued discussion of covariation reasoning, see 5.4.)

1.3.3 Summary

Kelley's (1967) model of attribution contains two core propositions: (a) that attribution is a choice between external and internal causes and (b) that the cognitive procedure by which people arrive at this choice is covariation assessment. Both propositions are problematic. First, the internal–external dimension cannot be the foundation of a theory of behavior explanation because, though it may be an important distinction in the explanation of unintentional events, it simply does not capture people's explanations of intentional action. Second, covariation assessment is not the only method by which people arrive at explanations. In the straightforward causal model that underlies explanations of unintentional events, covariation reasoning may be useful (though not essential; see Ahn et al. 1995; Johnson, Long, and Robinson 2001; Lalljee and Abelson 1983; Read, 1987). But the causal

model of intentional action is far more complex as it involves intentions, subjective reasons, and rationality (Malle 1999). Covariation reasoning can, at best, assist in constructing reason explanations of intentional action.

1.4 Subsequent Attribution Research

The three classic works by Heider, Jones and Davis, and Kelley were of course not the only important contributions to the study of behavior explanation. A number of scholars proposed theoretical additions, refinements, and extensions of attribution theory. I discuss these contributions under four headings: expanded causal dimensions, refined trait inference models, reasons and goals, and conversational processes. Despite their partial success, these contributions still left some old questions unanswered and raised several new ones. By the end of this review, then, we will be able to gather a list of desiderata that a theory of behavior explanations needs to satisfy.

1.4.1 Expanded Causal Dimensions

In the early 1970s, Bernard Weiner analyzed the domain of achievements and, in particular, the emotions and motivations people have toward others who succeed or fail. He relied on Heider's early insights and introduced the causal dimension of stability to complement the common one of externality–internality (Weiner et al. 1972). Empirical studies showed that people who failed because of lack of effort (unstable internal) were evaluated more negatively than those who failed because of inability (stable internal). (For a review see Weiner 1986.) Later Weiner also analyzed other outcomes that happen to people, such as illnesses, unemployment, or obesity. To account for the systematic differences in people's emotions and evaluations toward these outcomes, Weiner introduced the dimension of controllability (1995). Accordingly, empirical studies showed that people are more angry at agents who suffer negative outcomes brought about by controllable causes (e.g., an illness because of risky behavior) than agents who suffer negative outcomes brought about by uncontrollable causes (e.g., an illness because of a genetic precondition). Finally, Abramson, Seligman, and Teasdale (1978) analyzed the cognitive processes underlying helplessness and depression and proposed globality as a further dimension of causes. They suggested that attributing negative outcomes to global causes (especially if they are also internal and stable) was associated with higher degrees of helplessness

and depression. Empirical research incorporating this dimension was not flawless, however (Deuser and Anderson 1995), and studies showed the various causal dimensions to be so highly correlated as to make distinctions among them very difficult (e.g., Fincham and Bradbury 1992, table 1).

What all these proposals have in common is that they deal primarily with explanations of and emotional responses to *outcomes,* which are unintentional events. It may well be important to distinguish causes of unintentional events along a variety of dimensions (such as internality, stability, etc.), but for a theory of behavior explanation, these causal dimensions are not the whole story, because they do not apply to reason explanations of intentional behavior (Malle 1999). People are very concerned with explaining intentional behaviors (Malle and Knobe 1997b), and the moral and interpersonal implications of intentional behaviors are typically more significant than those of unintentional events. A theory of behavior explanation must therefore account for how people explain intentional behavior.

1.4.2 Refined Trait Inference Models

I described earlier how Jones and Davis's seminal paper from 1965 subtly turned attention away from action explanation and toward trait inferences. This shift had a lasting impact on attribution research, as is still visible in the numerous theoretical models on how and when people infer traits from single behaviors (e.g., Carlston and Skowronski 1994; Gilbert, Pelham, and Krull 1988; Newman and Uleman 1989; Quattrone 1982; Ross, Amabile, and Steinmetz 1977; Trope 1986). These models describe with great sophistication the process sequence from observing a behavior to inferring a correspondent trait (and adjusting or not adjusting this inference by considering situational forces). But it would be a mistake to assume that these trait inference models describe the process sequence of behavior *explanations*. Trait inferences and behavior explanations are plainly different processes, with different cognitive and social functions and different conceptual requirements (Fein 2001; Hamilton 1998; Hilton, Smith, and Kin 1995; Malle 1999). The methodology, too, of traditional trait inference studies does not reveal anything about behavior explanations. In these studies, by and large following a classic paradigm (Jones and Harris 1967), participants are asked to indicate high or low ratings on trait, attitude, or ability scales; they are never asked to explain why the target person acted as she did.

A new trend, however, promises to lift some of these restrictions on trait inference work. Research that began with Read, Jones, and Miller (1990) has shown that many trait inferences are based on reason explanations or motive ascriptions (Ames in press; Kammrath, Mendoza-Denton, and Mischel 2003; Reeder et al. 2002; Shoda and Mischel 1993). For example, when inferring an agent's morality, aggressiveness, or helpfulness from a given action, perceivers consider the agent's motives and reasons for her action and thus appear to construct behavior explanations before (or while) drawing a trait inference from it. The temporal ordering of these processes is not yet solidly established, but the changes in methodology that these studies introduced (presenting intentional stimulus behaviors and asking people to ascribe motives) represent a critical step forward in reconnecting the two strands of attribution—work on behavior explanations and work on trait inferences.

1.4.3 Reasons and Goals

In 1978, Allen R. Buss wrote a controversial paper in which he criticized mainstream attribution theory at a fundamental level. He argued that ordinary people do not explain all behavior with causes (as Kelley had suggested) but rather use *reasons* to explain intentional behavior. Reasons and causes are fundamentally different types of explanation, Buss, maintained and attribution theory created a good deal of confusion by equating the two. Buss's (1978) paper drew rather negative responses (Harvey and Tucker 1979; Kruglanski 1979), perhaps because his argument was flawed in its details or because he rattled a central pillar of attribution theory, which at the time lay at the heart of social psychology. Either way, mainstream attribution theory remained rather unaffected by this critique. Over the next decade or so, other scholars launched similar critiques, arguing that reasons are an autonomous form of explanation (Locke and Pennington 1982; see also Kalish 1998; Lennon 1990; Schueler 1989) and that attribution theories must incorporate reasons and goals into their conceptual repertoire (e.g., Lalljee and Abelson 1983; Read 1987; for a review see McClure 2002). Edward E. Jones, too, in an interview in 1978, admitted that the reason concept had been missed by early attribution work (Harvey, Ickes, and Kidd 1978, p. 379).

But it remained unclear exactly how the emerging conceptions of reasons and goals could be integrated with the traditional conception of causal

attribution. Most proposals relied on a two by two scheme with reasons versus causes on one side, and person versus situation on the other (e.g., Buss 1978; White 1991). This proposal raises serious problems, however. First, reasons are always person causes in that they are the agent's mental states that motivated her action (Davidson 1963; Kruglanski 1975; Locke and Pennington 1982). So what does it mean that a reason can be either a person or a situation factor? A satisfactory theory of behavior explanation needs to clarify whether or not reasons can be classified into a person–situation dichotomy and what such a classification would mean.

The second problem with the early conceptions of reasons and causes is that they tell us nothing about what determines when people use one or the other mode of explanation. The intentionality of the behavior must certainly be involved here (Buss 1978; Heider 1958; White 1991), but do people automatically offer reasons for all intentional behaviors? This cannot be true, as the following examples,[10] taken from student conversations, show:

(1-1) Why did she reveal the guy's name?—**She was just . . . she's like that. She has nothing to hide.**

(1-2) Why did your roommate cook all her food in the dorm room this year?—**Well, she had all of her food with her and her hot pot and toaster oven.**

In each of these cases, the agent performs an intentional action, but in none of them could we say that the explanation cites the agent's reasons for performing that action. For example, explanation (1-1) does not suggest that the agent thought "I am like that, I have nothing to hide; I should therefore reveal his name." Nonetheless, reference to these character traits somehow helps explain why the agent revealed the name. Similarly, in explanation (1-2), having her food, pot, and toaster with her was not the agent's reason for cooking in the dorm room (it wasn't as though she discovered the equipment and became motivated to cook). Even so, the presence of the cooking equipment explains an important aspect of her behavior. We need a theory that tells us both when people use alternatives to reason explanations and what the nature of these alternatives is.

Another fundamental problem that an adequate theory of behavior explanation needs to resolve is why reasons are used to explain intentional behavior in the first place. It may seem to some as obvious that they are, but

what makes intentional behaviors so different that they require a unique mode of explanation?

1.4.4 Conversational Processes

An important expansion of attribution research was achieved by a series of papers on the conversational nature of explanations (Kidd and Amabile 1981; Hilton 1990; Turnbull 1986). In these contributions, explanations are characterized not as cognitive processes in the social perceiver's mind but rather as publicly observable speech acts. More specifically, they are question–answer pairs, with "Why?" being the question and the explanation being the answer. Even though such pairs sometimes occur in people's own minds, more often they occur as an actual conversational exchange between a questioner and an explainer. This conversational analysis comes with the important implication that, in answering a why-question, explainers must take into consideration (a) exactly what the questioner finds puzzling or abnormal (Hilton and Slugoski 1986; Turnbull 1986) and (b) what information the questioner already has available (Slugoski et al. 1993). In a sense, the explainer anticipates what kinds of possible answers the questioner has in mind when asking the question (Bromberger 1965). This process of tailoring an explanation to the audience with whom one is communicating is evident in the following example:

(1-3) Q: But why did you have to leave [the football game]?

A: Because that was the time when . . . , it was like Saturday and I was coming back on Sunday, right? [she was there just for the weekend]

Q: Yeah, so you just wanted to pack.

A: Yeah, I had to pack, and I had to get ready, so during that night we could go out.

Q: Oh!

A: . . . so I wouldn't have to spend my whole night packing.

Q: Ya, ahah.

At each step of the way, the questioner engages with the explainer, chronicling, as it were, the process of discovering the answer to the why-question—

from hypothesis (Yeah . . .) to surprise over new information (Oh), to eventual understanding (ahah).

The insight that explanations are subject to conversational processes was a minor revolution, because it pulled attributions out of their cognitive isolation and highlighted the fundamentally social nature of explanations (which any new theory of explanation must grapple with). However, research into the conversational features of explanations did not expand on the conceptual apparatus of *person–situation causes,* inherited from Kelley and, still today, defining the textbook attribution model. This conservative stance is all the more surprising in light of the paradoxes that result from forcing conversationally situated explanations into the categories of person and situation causes (Antaki 1994; Monson and Snyder 1976; Ross 1977). Consider the following examples.

(1-4) I did my senior research paper in high school on homophobia **because it was just interesting.**

(1-5)* I did my senior research paper in high school on homophobia **because I was just interested in it.** (See note 10.)

According to the classic approach (e.g., Nisbett et al. 1973), (1-4) would be classified as a situation attribution and (1-5), as a person attribution. But this is puzzling because the two explanations do not seem to tell a different causal story; rather, they appear to be just linguistic variations of each other. Traditional attribution theorists can only shrug in light of such cases and insist that, *in general,* the person–situation dichotomy makes sense. The folk-conceptual theory of behavior explanations, as we will see in chapter 4, can easily account for both the linguistic surface difference and the deeper similarity between these two explanations.

Consider another striking example, in which grandmother is about to purchase a car and grandson explains her decision making.

(1-6) [The car's color] wasn't a problem any more, she decided, **because grandpa was dead, and he was the one that was anal retentive about cars.**

Despite surface appearances, here too we do not have a situation attribution, because grandfather's being dead and having been anal retentive in the past can hardly cause grandmother's decision in the present. So what kind of explanation is this?

Many more such examples can be found of behavior explanations that are sensible but that traditional causal attribution theories, focusing merely on person–situation categories, cannot adequately describe, even less so account for. The more mature models of dispositional attribution or trait inference (Gilbert 1989; Trope 1986) cannot come to our aid here, nor can the models of responsibility attribution (e.g., Shaver 1985; Weiner 1995), because none of them directly deal with *explanations* of behavior. Traditional attribution theories simply do not provide an adequate model of the tools and functions of folk behavior explanations. These explanations, however, are an essential element of social cognition and social interaction. Through behavior explanations, people find meaning in social behavior, form impressions, and influence other people's impressions; through behavior explanations, they blame and praise, coordinate interaction, and negotiate status and identity; and through these explanations, they tie together social events into narratives, bolster choices and preferences, and justify attitudes. In short, explanations are ubiquitous in social thinking and social behavior. There is no question, then, that a comprehensive theory is needed for this important social-cognitive tool.

1.5 Desiderata for a Theory of Behavior Explanation

The preceding review of classic and contemporary attribution theories has pointed to a number of shortcomings and unanswered questions in the extant literature. To resolve these problems and build a comprehensive scientific model of behavior explanation I propose a two-pronged approach. First, we need to recognize that explanations are people's way of finding meaning in both intentional and unintentional behavior. This meaning, however, emerges not in the ascription of person and situation causes but in the application of reasons, causes, and a variety of other modes of explanation, which are embedded in a conceptual framework called the *folk theory of mind and behavior* (chapter 2). By locating behavior explanations in this folk-conceptual framework, we are able to identify the various modes of explanation, their conditions of occurrence, and the cognitive processes underlying them (chapter 4 and 5). In so doing, we can show that the traditional person–situation dichotomy in attribution theory obscures a number of important distinctions (chapters 6 and 8) and that person–situation ef-

fects in attribution research—a seemingly impressive body of findings—may be based on serious misinterpretations of the evidence (e.g., chapter 7).

Second, we need to recognize that behavior explanations are a social tool that people use for a variety of social-interactive purposes. In offering explanations in social contexts, people try to manage social interaction—manage, that is, both the audience's impression of a given behavior and any joint future actions that explainer and audience might perform. These social uses of explanation, too, will be guided by people's folk-conceptual framework, because altering impressions of behavior occurs against the backdrop of shared fundamental assumptions about mind and behavior (chapters 3 and 6). Thus, the program of this book is to introduce the elements of people's conceptual framework of mind and behavior in which explanations are embedded and to offer a scientific theory of behavior explanations that recognizes this framework and takes into account both functions of behavior explanations: to find meaning and to manage interactions.