13

Folk Explanations of Intentional Action

Bertram F. Malle

How much easier it is to do psychology than philosophy! If I had to provide philosophical arguments for the unique modes of explanation that intentional action demands, I would swiftly be trapped in the hoary thicket of causality, free will, and the mind-body problem. As a social psychologist, I face a lighter task: I merely need to demonstrate that people explain intentional actions differently from how they explain other events. Such a demonstration, however, goes against a firmly established thesis in social psychology: that people explain all behavior (whether intentional or unintentional) by citing causes that are either internal or external to the agent. Even though this thesis is, on a basic level, correct (behavior explanations do provide, among other things, causal information), it grossly simplifies the conceptual structure and social functions of folk explanations of behavior. In particular, it fails to acknowledge the central role that people's concept of intentionality plays in shaping their explanations of behavior.

In this chapter I try to describe in detail how people explain intentional behavior, spelling out the unique conceptual, linguistic, and social features of these explanations. I argue that people explain intentional actions primarily with reasons (the explanation mode traditionally associated with intentionality) but also with two other modes that are conceptually and functionally distinct. I then compare the resulting model of folk explanation of behavior to classic attribution theory and recent developmental work on explanations. Finally, I apply the model to the debate between simulation theorists and theory-theorists, seeking an integration between the two approaches.

Intentionality and the Three Modes of Explaining Action

The social practice of behavior explanation is embedded in the folk theory of mind and behavior (Bruner 1990; Gopnik 1998; Heider 1958). Explanations both rely on and reveal this framework's central concepts—first and foremost, the concept of intentionality (Malle 1999; Malle and Knobe 1997b). Therefore, an account of how people explain behavior naturally begins with an analysis of the folk concept of intentionality.

The Folk Concept of Intentionality

Intentionality is a complex folk concept that specifies under what conditions people judge a behavior to be intentional (Malle and Knobe 1997a). A behavior is judged intentional when the agent has, at least, a desire for an outcome, a belief that the action leads to that outcome, an intention to perform the action, the skill to perform the action, and awareness of fulfilling the intention while performing the action. These are the minimal conditions for folk judgments of intentionality, but they do not yet tell us how people *explain* intentional behavior. To identify actual modes of explanation, we must separate the components of intentionality into three domains of intentional action that people find worthy of explanation (figure 1).

The first domain refers to factors that enabled the action to come about as the agent intended it (Malle 1999). Without such enabling factors, the intention would merely lead to an attempt, not to a completed action. The folk concept of intentionality specifies only one necessary enabling factor: the agent's skill.[1] That is, if a behavior is performed and fulfills an intention, the agent must have brought about that behavior with skill (rather than luck) for the action to count as intentional (Malle and Knobe 1997a; Mele, this volume). There are other enabling factors (among them effort,

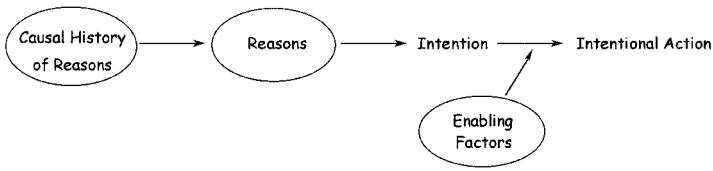

Figure 1
Domains of explanation (circled) within the folk concept of intentionality.

persistence, opportunities, and removed obstacles) that are not necessary for an action to count as intentional but are necessary for the action to be performed in the first place (McClure and Hilton 1997; Turnbull 1986). What all enabling factors have in common is that they explain how it was possible that the agent turned her intention into the intended action.[2] For example, "She hit her free-throws because she had practiced them all week." For many social actions, a "How possible?" question does not even come up; however, when it does come up (under conditions that will be discussed later), *enabling factor* explanations are the mode of choice to answer this question.

The second domain worthy of explanation refers to the reasons the agent had for acting (Audi 1993; Buss 1978; Davidson 1963; Locke and Pennington 1978; Malle 1999; Read 1987). Reasons are seen as representational mental states (desires, beliefs, valuings) that the agent combines in a (sometimes rudimentary) process of reasoning that leads to an intention and, if all goes well, to the intended action. The concept of intentionality specifies two minimal reasons for an action to be intentional: that the agent have a desire for an outcome and that the agent have a belief that the intended action leads to that outcome. These minimal reasons are sometimes explicitly mentioned in explanations of intentional action. For example, a student explained why she chose psychology as her major by saying "I want to go to graduate school in counseling psychology; I think psychology is the right major to have as background for counseling psychology." In many naturally occurring explanations, however, other reasons are mentioned, such as desires for avoiding alternative outcomes, beliefs about the context, beliefs about consequences, and valuings of the action itself.[3]

The intention underlying an action seldom serves an explanatory function by itself because the propositional content of an intention to A is the action A itself, which is then still left to be explained: "Why are you going shopping now?" "Because I intend to go shopping." Some scholars' claims about the explanatory function of intentions typically rely on a confounding of desires and intentions (Malle and Knobe, this volume; Moses, this volume). Intentions do answer an important question, namely, *what* the agent is doing ("She is trying to fix the computer"). But in answering this question, the intention describes the action at the right level (from the agent's perspective) without actually explaining it (Malle 1999, note 1).

The third domain of explanation refers to factors that lie in the causal history of reasons and thus clarify what led up to these reasons in the first place. The folk concept of intentionality is silent on the causal history of reasons. No matter how an agent's reasons originated, what counts toward intentionality is whether the reasons include a desire for an outcome and a belief that the action leads to that outcome. But under some conditions folk explainers are interested in this causal history, and I will discuss these conditions for *causal history of reason* explanations in more detail later.

The concept of intentionality thus allows us to locate three domains of explanation and their corresponding modes of explanation: how it was possible that a given action occurred (enabling factor explanations), why the agent intended to act that way (reason explanations), and what lay in the background of these reasons (causal history of reason explanations). For research purposes it is important to note that these three explanations can be reliably distinguished from one another when coding naturally occurring folk explanations of behavior ($\kappa = 0.72–0.88$).[4]

I will now discuss each of these explanatory modes in detail, beginning with reason explanations. They are the most frequently used explanation mode; they also have unique conceptual and linguistic features that differentiate them from all other explanations.

Reason Explanations

Reasons are mental states that help produce an intentional action. In this sense, they are considered to have "causal power." However, they are quite distinct from mere causes, because they perform a very specific function in bringing about intentional behavior. Mental states count as reasons only if they played a role in the agent's reasoning toward forming an intention to act, and this role is characterized by two essential features: *agent subjectivity* and *rationality*.

Two Essential Features of Reason Explanations

Agent Subjectivity

When providing reason explanations, folk explainers cite those mental states in light of which, to their best knowledge, the agent formed an intention to act. They try to reconstruct the decisive deliberations that the agent

underwent when forming her intention and thus take the agent's subjective viewpoint in explaining her action. There may be other good reasons for acting the way the agent did, but what counts as an explanation must refer to her own subjective reasons in deciding to so act. Because of the agent subjectivity of reason explanations, and because of the assumption that an agent undergoes some sort of reasoning process, folk explainers assume that the agent had at least minimal awareness of (the content of) her reasons; otherwise they would not be *her* reasons.

Consider the explanation "Shanna ignored her brother's arguments because they were irrelevant." If folk explainers consider this a reason explanation, they should assume that the agent was aware of the explanation's content and consequently should reject as senseless the added claim "even though she was not aware of the fact that they were irrelevant." We designed a study that would test this prediction across different behaviors and explanations (Malle, Knobe, O'Laughlin, Pearce, and Nelson 2000).

We first constructed several behaviors and, for each, various explanations that contained no obvious linguistic markers of reasons. Thus, we removed mental state markers such as "he wanted" or "she thought," and we excluded all desire reasons (because they have a characteristic linguistic structure of "so (that) . . ." or "(in order) to . . ."). Then we presented these behavior-explanation pairs to undergraduate students and selected seven pairs that were clearly judged to be reason explanations by a majority of the students (ranging from 78 percent to 98 percent per explanation). We had also included explanations of the same actions that were not reasons but causal histories of reasons, which provided a judgmental and statistical contrast to the reasons. Each of the seven reason explanations was then paired with a statement that negated awareness, of the form "even though [agent] was not aware that/of [explanation content]"—for example, "Carey watered her plants because the leaves were wilting (even though she was not aware that the leaves were wilting)." These statements were presented to a second group of students, and when asked whether these reason explanations made sense or not, an average of 77 percent said they did not make sense (versus 28 percent for causal history of reason explanations). A follow-up study clarified why 23 percent of students claimed that a reason explanation could still make sense even though the agent's awareness was negated. Virtually all of them assumed that the offered explanation described

some preceding cause, not a reason: when given an opportunity to clarify their judgments, they spontaneously offered alternative reasons for which the agent in fact acted (Malle et al. 2000, study 2).

Rationality

Besides agent subjectivity, folk explainers assume a second essential feature of reason explanations: a rational link between the reasons and the intended action. They try to cite only mental states that would make it appear rational for the actor to form her intention. The constraint of rationality excludes beliefs and desires that bring about an intentional behavior in a merely causal way and are therefore not the agent's reasons for acting that way (e.g., "She came to the party 'cause she didn't know that her ex was gonna be there"). The rationality assumption also explains why one reason typically implies a host of other reasons. For example, the explanation "Anne was driving above the speed limit because she knew the store closed at 6 o'clock" cites a belief reason that readily entails the desire reason "and she wanted to get to the store." Similarly, the explanation "Anne was driving above the speed limit because she wanted to get to the store before 6 o'clock" cites a desire reason that readily entails the belief reason "and she thought that only by driving fast could she be there by 6 o'clock." Nothing in the desire (as stated) entails the belief, and nothing in the belief (as stated) entails the desire. Only under rationality constraints—the assumption that reasons combine rationally in the formation of an intention—do these implications follow.

Of course, folk explainers need not share or approve of the agent's reasons; they need only acknowledge that, given those reasons, it is rational (reasonable, intelligible) for the agent to form her intention. That also implies that an agent who avows her behavior as intentional subjects herself to the scrutiny of rational criticism (Schueler, this volume). Conversely, if the agent tries to appear rational, she will portray her behaviors as intentional and, in particular, explain it with reasons (Malle et al. 2000).

It is still a matter of debate what the rationality assumption exactly entails (cf. Føllesdal 1982). Some scholars (e.g., Davidson 1980c) require logical consistency within the agent's reasoning chain; others (e.g., Collingwood 1946) require only intelligibility for the explainer. Yet others invoke a more

general normativity rather than a specific rationality norm (Schueler, this volume). Clearly there is a need for empirical work on the exact features of rationality that ordinary people assume when they ascribe reasons to an agent.

The Grammar of Reasons

Reason explanations are often expressed in natural language, either when the agent gives an account of her own action or when an observer gives an account of another's action. To understand the function of reasons in social interaction, we must identify the "grammar" of reasons—the conceptual and linguistic parameters that differentiate reasons from each other. I will focus on three such parameters: what type of mental state constitutes the reason, whether that type is linguistically marked, and what the mental state represents in its content.

Reason Types: Desires, Beliefs, Valuings

Consider the following reason explanations for why the agent teased another person:

(1) because she wanted to make the other kids laugh

(2) because she disliked the way the person looked

(3) because she thought that the boy was too feminine.

All three of these reasons are clearly marked as subjective mental states (of wanting, disliking, thinking), and each implies a larger network of mental states that were involved in the agent's reasoning (e.g., the belief that the boy was too feminine implies in this context that the agent didn't like that and wanted to make fun of it). Reasons such as these can usually be classified as desires (1), valuings (2), or beliefs (3).

Desire reasons reveal the action's desired outcome, which is often called the action's goal, aim, end, or purpose. Consequently, desire reasons are answers to the question "For what purpose?" or "What for?" An unfulfilled desire is the paradigmatic instigator of action, as it represents something the

agent lacks and tries to get through acting. Mentioning a desire reason thus portrays the agent as deficient (i.e., "wanting") in some respect and as driven toward removing the deficiency. Moreover, mentioning a desire as a reason for acting (not just as a description of a mental state) portrays the agent as endorsing the desire as worth pursuing (Schueler, this volume). Thus, merely by placing an outcome in the content of a desire reason, the explainer can indicate the outcome's worthiness, at least from the agent's subjective perspective: "Why did she turn up the volume?" "To make her brother mad."

Valuings, like desires, indicate positive or negative affect toward the representational object. This affect can be absolute (liking, hating, missing something) or relative (preferring one thing over another). Valuings are primarily used to indicate the inherent desirability of an action (e.g., "Why did she go dancing?" "She loves dancing"), whereas desires typically indicate the desirability of an outcome (Malle and Knobe, this volume). Like desires, valuings ascribe to the agent an evaluative attitude toward an object (or the action itself); however, unlike desires, valuings do so explicitly and without highlighting the agent's deficiency.

Belief reasons encompass a broad range of knowledge, hunches, and assessments that the agent has about the outcome, the action, their causal relation, and relevant circumstances. Beliefs are aimed at representing reality and thus are not, by themselves, apt to instigate action. But they are essential in choosing worthwhile outcomes to pursue and actions to select as means. They help the agent track feasible paths of action, consider the consequences of those actions, and navigate around obstacles, and they can represent other people's wishes and reactions. The latter is crucial in coordinating one's actions with others.

My co-workers and I have recently begun to explore the determinants and functions of explainers' choice of reason type. Across thousands of free-response behavior explanations, we are finding a base rate distribution of roughly 50 percent desires, 10 percent valuings, and 40 percent beliefs. We are also finding systematic variations in the use of reason types. For example, when people explain their own behavior, they use belief reasons to present themselves as rational (Malle et al. 2000). Furthermore, there is a reliable actor-observer difference: Observers tend to use more desire

reasons and fewer belief reasons than actors (Malle, Knobe, Nelson, and Stevens, in preparation), presumably because desire reasons are more generic and easier to guess whereas belief reasons require situated information that may be idiosyncratic to the agent.

Mental State Markers

The nature of reasons as subjective mental states can be linguistically highlighted with verbs such as "I thought," "he wanted," and "she likes." If no such mental state markers are used, only the reason's propositional content is cited in the explanation. For example, when explaining why Anne waters her plants with vitamin B, we may cite a desire reason that is marked ("because she wants them to grow faster") or unmarked ("so they will grow faster"). Similarly, we may cite a belief reason that is marked ("because she thinks they will grow faster") or unmarked ("because they will grow faster").

The linguistic device of mental state markers has a number of interesting implications for reason explanations. First, unmarked beliefs downplay their nature as mental states. For example, "Joan canceled the party because it was raining" looks, on the surface, identical to "Joan got wet because it was raining." But of course in the first explanation the rain did not directly cause Joan's behavior—she decided to cancel the party in light of her belief that it was raining. In accordance with the subjectivity assumption for reasons, the explanation is shorthand for "Joan decided to cancel the party because *she thought* it was raining." However, omission of the marker and direct reference to the content make the action seem to be more rational and a "natural" response to the situation. Indeed, actors who attempt to appear rational do so primarily by increasing their use of unmarked belief reasons (Malle et al. 2000). Second, by marking a belief reason with a mental state verb, social perceivers can emphasize that this was the *agent's* belief and was not necessarily shared by the perceiver: "She's quitting her job because she thinks her pay sucks." Thus, explainers can distance themselves from an agent's reasoning by using mental state markers or can embrace this reasoning by omitting them. With this subtle linguistic tool, they communicate to their audience how reasonable or justified they feel the agent's action was (Malle et al. 2000). Third, when desire reasons are expressed without mental state

markers, they still indicate their nature as subjective reasons. Grammatically, unmarked desires are always expressed in the form of "(in order) to," "so (that) . . . ," or "for . . . (sake)," each citing the agent's purpose for acting. Thus, even in their unmarked form they express the reason's subjectivity. The marked form, however, can highlight the agent's deficiency (e.g., "He went to the store because he *needed* more milk") or self-centeredness (e.g., "We had to stop because he *wanted* to have some coffee").

Reason Contents

Reasons, as representational states, always have a content: the object, action, or state of affairs that is desired, valued, or believed. The content of reasons is what the agent considers in forming an intention (e.g., "They say it will rain tomorrow [belief content]; having the party in the rain wouldn't be fun [valuing content]; I'd better cancel it [intention]"). Thus, the content of reasons is what renders actions intelligible. Whereas the general folk model of reason explanation specifies that actions are to be explained by *some* beliefs, desires, and valuings, the actual explanation of a concrete action requires one to know which particular reasons the agent had, and this particularity is given by the content of reasons.

Even though reason contents are essential for explanations of intentional action, their psychological study is marred by difficulties. For one thing, a given reason content can be represented in various ways on the linguistic surface. For example, I might say "I wish I were rich" or "I want to be rich" to express exactly the same desire content of being richer; however, the first expression mentions the agent (*I were rich*), whereas the second does not (*to be rich*), following a simple grammatical operation of "equi-subject deletion" (Givon 1993). Research that classifies behavior explanations according to their mentioning of the agent versus the situation (e.g., McGill 1989; Nisbett, Caputo, Legant, and Marecek 1973) is therefore oversensitive to surface variations and grammatical rules rather than to the actual mental content represented in the reason.

But even refined classifications of reason content that take grammatical rules into account (and, for example, code "I want to be rich" as referring to agent content) quickly run into difficulties. For example, how should one classify "I cried because I received a farewell letter from her"? Is the receiv-

ing agent in the foreground, or should we code for the implicit sender of the letter? My co-workers and I have tried to be inclusive and permit a variety of interaction codes for these complex cases (e.g., interactions between the agent and the situation, the agent and other people, other people and the situation, etc.; see Malle 1998). However, even with coding that is sensitive to grammar and to complex contents, we have not yet found reason content to be reliably predictive of other psychological variables. For example, the classic thesis that people explain their own actions more with reference to the situation and others' actions more with reference to them as agents (Jones and Nisbett 1972) does not hold up to scrutiny (Malle 1999; Malle et al., in preparation).

It is conceivable that reason content is so action specific that it defies general psychological regularities. Because reason content provides the rational connection among beliefs and desires in leading up to an intention, it may not have to serve any further psychological function. But before we accept this conclusion, other classification schemes might be tested. Instead of classifying reason content into agent, situation, and various interactions, perhaps we should code the content for its social desirability. We might examine whether reasons with desirable content are more accepted by audiences than those with undesirable content, whether explainers are more likely to lie when the true content of their reasons is undesirable in the eyes of a given audience, and whether desirable content is sought when excusing an agent and undesirable content is sought when accusing an agent.

Clearly, much research is needed on the psychological functions of the three parameters of reasons (and on other parameters not discovered so far). Another topic of research, which already has some findings to report, concerns the question "Under what conditions are reason explanations used in the first place?" The answer is simple: In about 80 percent of cases, people explain intentional actions by means of reasons, because reasons answer directly what a why-question inquires about (namely, what motivated the agent to perform the given action—what was the *reason*-ing behind it). The more complex question is "When and why do people *not* use reason explanations and instead use causal history of reason explanations or enabling factor explanations?" I now turn to the conditions under which people do use these alternative modes of explanation.

When People Use Causal History of Reason Explanations

One alternative to providing reasons when explaining intentional behavior is to cite factors that lay in the causal history of the agent's reasons (Malle 1994; Malle 1999; see also Hirschberg 1978; Locke and Pennington 1982). Consider the following examples:

Anne invited Ben for dinner because she is friendly.

Carey watered the plants because she stayed at home in the morning.

Even though these explanations clarify intentional behavior, they do not mention what reasons the agent considered when forming her intention; rather, they mention the causal history of those reasons. Causal history of reason (CHR) explanations describe the context, background, or origin of reasons, so they are not constrained by the agent-subjectivity or rationality rules that apply to reasons themselves. Anne did not consider "I am friendly, I better invite Ben for dinner." Rather, her friendly disposition triggered some of her reasons, such as a desire for talking to Ben or doing something nice for him. Similarly, Carey probably did not consider "I stayed at home in the morning; therefore I will water my plants." More likely, her being at home triggered her desire to care for her plants or made her realize that they needed water.

CHR explanations account for only about 20 percent of intentional-behavior explanations; thus, when explainers offer CHR explanations, they deviate from the standard of giving reasons. Under what conditions do explainers do that? O'Laughlin and Malle (2000) suggest two conditions; I add a third.

Knowledge

If the explainer does not know the exact reason for an intentional behavior but nevertheless wants to offer an explanation, he may offer a causal history explanation. Consider the following transcript: "Then why would Tanya come up and talk to us out of her own free will?"—"Well . . . weird people do these kinds of things." As can be inferred from the explainer's hedging, he does not actually know the reason why Tanya decided to talk

to them. Instead of admitting his ignorance, however, he offers an explanation citing a personality trait that presumably caused whatever specific reason Tanya had for her action.

The knowledge condition predicts an actor-observer asymmetry of using CHR explanations (relative to reason explanations). Actors typically know the reasons for their own actions and should therefore offer mostly reason explanations, whereas observers often do not know others' reasons and should therefore offer relatively more CHR explanations. Indeed, we found this asymmetry across a wide variety of contexts, including conversations, memory protocols, questionnaires, and interviews (Malle et al., in preparation).

Conversational Relevance

Explanations in conversation are subject to rules of relevance (Grice 1975; Hilton 1990; Sperber and Wilson 1986; Turnbull 1986). Reasons typically provide the most relevant and informative answers to why-questions; sometimes, however, they are cumbersome or obvious, and in these cases explainers may prefer to use CHR explanations. I will describe these cases in turn.

When a series or trend of intentional behaviors is explained, the reasons for each specific action may vary, while the "historic" determinants of this set of reasons may be constant. Thus, a causal history explanation can offer a parsimonious account of the whole class of possible reasons that the actor may have for each respective action. For example, in "I go to the supermarket almost every day because I have three kids," the fact of having three kids is not the agent's conscious reason for going to the supermarket. Rather, on each separate occasion, the fact of having three kids brings about a reason to go to the supermarket: one time a child is sick and needs cough drops, another time a child stains the carpet and there is no stain remover in the house, and so on. Across these occasions (and their corresponding reasons), the single fact of having three kids explains why the actor repeatedly goes to the supermarket, and it does so more parsimoniously than a string of individual reasons would. The same logic of parsimony applies to the explanation of aggregate behaviors, which describe behavior trends across people rather than behavior trends of one person across time. When a whole group of people act in similar ways but each individual has different reasons to so act, a CHR explanation is a parsimonious account of the

entire group's behavior, setting aside the variety of individual reasons (O'Laughlin and Malle 2000).

A second case in which conversational relevance favors CHR explanations is when an agent's reasons are obvious and the explainer seeks to provide an answer that goes beyond the obvious: "Why is she planning to get pregnant?" "I guess her biological clock is ticking." In this example, the explainer probably assumed that the questioner already knew the woman's reason for getting pregnant (her desire for a child). What the questioner may wonder is why she has this desire in the first place. A CHR explanation that clarifies the origin of her desire provides an informative answer to this question.

Strategic Presentation

Finally, causal history explanations can be used strategically to downplay the agent's reasoning process (which is normally highlighted by offering reason explanations, especially belief reasons). For example, a suitor may want to downplay the degree of deliberation behind his actions: "Why did you come all the way to bring me the book?" "Oh, because I was in the area, and I happened to have the book with me." Moreover, CHR explanations are occasionally used to downplay intentionality and responsibility and thus mitigate blame or punishment (Wilson 1997). For example, Nelson and Malle (2000) found that people's use of causal histories is greater when explaining undesirable actions than when explaining desirable actions.

To summarize: People use causal history factors to complement reason explanations or to substitute them for reasons that are not known or would not achieve a particular communicative goal. Because some CHR explanations refer to traits, the concept of causal history explanations also helps clarify the relationship between traits and mental states in explanations of intentional behavior (Rosati, Knowles, Kalish, Gopnik, Ames, and Morris, this volume). Traits can be used instead of mental state explanations when the agent's reasons are unknown or do not fulfill the speaker's conversational goals. Moreover, traits can elucidate the background and origin of the agent's specific reasons. Finally, because traits are temporally stable, they aide in predicting the agent's future behavior in different contexts regardless of the agent's context-specific reasons.

When People Use Enabling Factor Explanations

The second alternative to using reasons when explaining intentional behavior is to cite enabling factors. Enabling factor explanations do not answer a motivational question (as do reason and CHR explanations); rather, they answer a performance question. Consider these examples:

How come John aced the exam?—He's a stats whiz.

Phoebe got all her work done because she had a lot of coffee.

Enabling factor explanations exist because of the imperfect link between intention and action. An agent might have reasons to act a certain way and so might form an intention. But whether this intention is turned into a successfully performed action often depends on factors beyond the agent's intention and reasons—factors that *enable* the action. Because enabling factor explanations clarify performance rather than motivation, they should increase in response to the question "How was this possible?" relative to the motivational question "Why?" or "What for?" Indeed we found that enabling factor explanations occurred 4–12 times more frequently in response to a "How possible?" question than in response to any other explanatory question (Malle et al. 2000).

A second condition under which enabling factor explanations increased in frequency was when the behavior in question was difficult (as is the case with artistic, athletic, or complex actions). In such cases it may often seem surprising that the behavior was successfully performed, and surprise demands explanation. Accordingly, we found that enabling factors occurred 7–8 times more frequently with difficult behaviors than with easy behaviors. (See also McClure and Hilton 1997.)

Competing Models of Folk Explanation

Two models of folk explanation have received generous support and attention in the psychological literature: social psychology's attribution theory and developmental psychology's study of children's explanations within their theory of mind.

Attribution Theory

Attribution theory has been the favored psychological model of folk explanations of behavior for more than 40 years. Its history began when Heider (1958) offered insights into folk explanations by considering them part of people's "naive theory of action." However, the models developed later by Jones and Davis (1965) and Kelley (1967) left the central aspect of this naive theory—the concepts of intention and intentionality—behind. Heider repeatedly emphasized the importance of intentionality, but he used the terms *personal causality* and *impersonal causality* to refer to folk conceptions of intentional versus unintentional behavior (Heider 1958, pp. 100–101). This choice of terms and the occasional ambiguities in Heider's writing (Malle and Ickes 2000) led to a major recasting (and misunderstanding) of his distinction into one between "person causes" and "situation causes." Soon after Kelley's (1967) landmark paper, attribution researchers classified all folk explanations of behavior into those that cite person causes and those that cite situation causes, irrespective of the behavior's intentionality.

The person/situation dichotomy may appear simple, elegant, and predictively useful. Its major flaw is, however, that people don't think about behavior solely in terms of person causes and situation causes. In fact, no direct evidence has ever been provided that people's theory of behavior assigns a significant role to the person/situation dichotomy. All evidence comes from reactive measures that forced people to express their explanations in terms of person/situation ratings or from content codings that classified only the linguistic surface of explanations (Malle et al. 2000). In contrast, there is evidence that people alter their explanations depending on the behavior's intentionality, and that they explain intentional behavior primarily with the agent's reasons whereas they explain unintentional behavior with mere causes (Malle 1999). Attribution theory provides a good account of how people explain such unintentional behavior using "cause explanations." However, a causal attribution model is insufficient as an account of how people explain behavior in general and intentional behavior in particular.

For example, attributional analyses of reason explanations have ignored the complex grammar of reasons and drawn misleading conclusions from

reasons' linguistic surface (Malle 1999; Malle et al. 2000). In particular, when reasons lack a mental state marker and their content mentions something about the situation, researchers have mistakenly classified them as "situation causes" (e.g., "She didn't go because her ex was there"). But in such unmarked reasons explainers are not referring to situational causes that somehow made the agent act; rather, they are expressing the content of a belief that the agent considered before acting. The mistake of treating reason contents as causes is most obvious when that content speaks about future or hypothetical states, as in "He doesn't let his daughters go out after midnight because something could happen." No doubt the explainer is referring here to hypothetical dangers that function not as mere causes but as contents of the agent's belief.

Developmental Work on Explanations

The last 10 years have seen a surge of interest in children's theory of mind. Even though explanations are considered a hallmark of this developing theory (Gopnik 1998), researchers have only recently turned to examining children's explanatory reasoning in detail (Bartsch and Wellman 1995b; Kalish 1998; Schult and Wellman 1997; Wellman, Hickling, and Schult 1997). Two assumptions have guided much of this research. The first is that children develop three distinct explanatory models: a folk psychology, a folk physics, and a folk biology. The second is that folk-psychological explanations construe human behavior in terms of internal states, particularly beliefs and desires. Unfortunately, this classification system blurs the critical distinction between reason explanations and other modes of behavior explanation by treating beliefs and desires broadly as mental states and not functionally as either reasons, causes, or causal history factors.

For example, Schult and Wellman (1997) group under "psychological explanations" those statements that refer to the agent's mental states, including desires and beliefs, but also moods and lack of knowledge. (See also Bartsch and Wellman 1995b, chapter 6.) The examples cited by Schult and Wellman show that the class of "psychological explanations" includes both reason explanations of intentional behavior ("Why did Jimmy pour milk in his cereal bowl?" "Because he likes it") and cause explanations of unintentional behavior (e.g., "Why did Sarah squeeze

ketchup on her ice cream?" "Because she didn't know it was ketchup.").[5] Moreover, Hickling and Wellman (2000) seem to classify some reason explanations as nonpsychological explanations when the content of the reason refers to biological or physical states—for example, "The reason I ask for so much juice is because I get thirsty."

Wellman and colleagues have demonstrated that children as young as 3 years systematically use mental state explanations for human behavior, but these findings leave open the question whether children differentiate between mental states as reasons and mental states as mere causes. Perhaps children first apply mental state explanations broadly to human behaviors and learn to distinguish between reasons and other mental causes only after acquiring the concept of intentionality, around the age of 5 years (Shultz and Wells 1985). Command over this concept would probably involve an understanding of the scope and limits of choice, also acquired around age 5 (Kalish 1998). One test for whether children understand the reason-cause distinction might involve asking them to differentiate belief/desire explanations that function as reasons from belief/desire explanations that function as causes (or as causal histories of reasons, in which case the explanations are equated in terms of accounting for intentional behaviors).

Explanations as Cognitive Process: Theoretical Inference, or Simulation?

Much research and thinking regarding explanations has been devoted to the question of what cognitive processes underlie the forming of explanations. The dominant answer within social psychology has been that all explaining—regardless of the object of explanation—relies on domain-general (tacit) cognitive mechanisms, such as covariance analysis, evidence updating, or connectionist nets. (See, e.g., Cheng and Novick 1990; Kelley 1967; Kruglanski 1989; Read and Marcus-Newhall 1993; van Overwalle 1998.) However, it is hard to see how these models would account for the distinction between reasons and other modes of explanation. After all, explanatory reasoning would only consist of correlating causes and effects, no matter whether those causes are reasons and the effects are intentional actions. What makes a reason explanation plausible and acceptable, however, is not just the assumption that the hypothesized mental state probably caused the action in question but also the assumption that it was a *rational reason* for acting that way, and such considerations involve

domain-specific semantic analyses rather than domain-general syntactic analyses.

The literature offers two main candidates for such domain-specific structures. One is theory-theory, which postulates that people use distinct concepts (e.g., intention) and rules (e.g., that intentional behavior always has a point or purpose) when explaining human behavior, and these concepts and rules together make up a folk theory of mind and behavior. An alternate candidate is simulation theory, which postulates that perceivers simulate others' putative mind states, using their own faculties of perceiving, reasoning, and feeling as models that deliver (off-line) predictions or explanations (Gordon 1986; Goldman 1989).

Without reconsidering the entire debate between simulation theory and theory-theory (Carruthers and Smith 1996; Davies and Stone 1995), I would like to explore briefly how simulation and theory-theory might account for the four modes of folk explanation of behavior identified earlier: reasons, causal histories of reasons, enabling factors, and mere causes. I am not assuming that simulation theorists would necessarily claim to account for all four modes of explanation (as theory-theorists would). In fact, most simulationists might claim to account only for explanations involving mental state ascriptions, others perhaps only for reason explanations. Even so, we must ask what theoretical apparatus researchers have to adopt in order to account for all four modes of explanation.

Enabling factor explanations present the most clear-cut case. In describing what made it possible that a given action was accomplished, these explanations concern the process that allowed an intention to "come into the world," which is something that perceivers cannot simulate experientially but can only think through theoretically. Over time, they will acquire generalizations about the kinds of factors that enable certain types of actions. Thus, theory-theory provides an adequate account for enabling factor explanations, but simulation theory does not.

When discussing cause explanations of unintentional behavior, we must distinguish between two major types of unintentional behavioral events. The first includes events that are "biological" or "physical" in nature, such as sweating after exercise or tripping over a root. An adult perceiver does not simulate exercising to know why someone in athletic clothes looks sweaty and exhausted, and it is unlikely that children would first have to exercise and sweat themselves before they can learn the empirical generalization in

question. Cause explanations of physical or biological behavior thus seem to be based on learned knowledge structures, hence on some version of a theory.

The second type of unintentional behavior includes "psychological" events, such as sadness after losing a game or pain from an injection. These phenomena are clearly open to simulation. In fact, some psychological behaviors automatically trigger empathic simulation (which is why people cry in movie theaters and look away when somebody receives an injection). Empirical research on emotional contagion and empathy (see, e.g., Levenson and Ruef 1997) supports the claim that feelings and emotions are suitable and frequent objects of imitation and simulation. Occasional knowledge-based prediction or explanation is not ruled out, but the prevalence of emotional imitation and contagion in children suggests that the simulative approach may come first in this domain (Goldman, this volume). Thus, a full account of cause explanations requires both theory-theory (to account for biological and physical events) and simulation theory (to account for psychological events).

In the case of reason explanations, the challenge is to account for the normative and rational "glue" that connects reasons to intentions and actions. Theory-theory would postulate that social perceivers assume a rationality principle (e.g., "People act to get what they want, given what they believe about how to get it" (Ripstein 1987, p. 468)), and a version of this rule is indeed part of the adult concept of intentionality (Malle and Knobe 1997a). However, it seems highly implausible that 3-year-olds have such an abstract insight into rationality. At the same time, 3-year-olds seem to be quite adept at offering reason explanations (Bartsch and Wellman 1995b). The simulation approach does not postulate anything special in the child perceiver (even in one only 3 years old) that we don't already attribute to the child agent, namely a practical reasoning faculty that *itself* rationally progresses from beliefs and desires to intentions and actions. However, such simulations would be hard to get started without general rules that guide the search for possible explanations—rules such as "If the behavior looks intentional, search for beliefs and desires and plug them into the deliberation mechanism." On the other hand, it is not clear how perceivers who according to theory-theory follow a general rationality rule "fill in" the concrete beliefs and desires that might be the agent's reasons. Inferring on the basis of a rationality rule that the agent had *some* beliefs and desires is one thing;

inferring the specific ones in this context is quite another. Thus, perceivers may well project their own perceptions of the situation onto the agent and may occasionally complement them with beliefs and desires that they regard as idiosyncratic to the agent. This input could then be fed into the perceiver's own practical reasoning faculty or, once sufficiently developed, into a more abstract inference heuristic. For reason explanations, then, as for cause explanations, both simulation theory and theory-theory are needed.

CHR explanations, too, may require a mixed account. CHR explanations referring to abstract cultural and personality factors are difficult if not impossible to simulate experientially, whereas those that refer to specific situational or internal triggers invite an act of pretending to be in the agent's shoes. Once the perceiver simulates the agent's situation or internal state, it may become apparent why the agent acted in this particular way. As an example, Ripstein (1987) offers Orwell's explanation for why he did not shoot an enemy soldier: the man was running across the trenches, holding up his trousers. In this case, simulation rather than general rules seems to lead to explanation and understanding of the agent's action. The specific image of a man holding up his trousers seems to trigger in anybody the desire to not shoot him, and once one experiences that image and the consequent desire the decision not to shoot the person is wholly intelligible (and much more so than if one merely cited the obvious desire "he didn't want to shoot him"). Not surprisingly, good writers, filmmakers, and defense lawyers lead their audiences to simulate a character's situation in such detail that the agent's subsequent reasons and actions seem to follow with utter necessity.

Stich and Nichols (1995) argued that "it may well turn out that some of our folk-psychological skills are indeed subserved by simulation processes, while others are subserved by processes that exploit a tacit theory." Similarly, the full range of folk explanations of behavior appears to require both the capacity to simulate and the capacity to use generalized knowledge structures.

Summary

I have tried to demonstrate that the concept of intentionality is a key to understanding folk explanations of behavior. The structure of intentionality defines three domains of intentional action that people find worth

explaining: the factors that enable an action's successful performance, the reasons for acting, and the causal history of those reasons. In contrast, behaviors not considered intentional are straightforwardly explained by antecedent causes. Among these four modes of explanation, reason explanations are unique in that they must meet the constraints of agent subjectivity and rationality, thus capturing agents' own reasoning toward their intentions to act. Future research should focus on the social functions of behavior explanations and the cognitive processes, such as inference and simulation, that underlie folk explanations of behavior.

Acknowledgments

Preparation of this chapter was supported by NSF CAREER award SBR-9703315. I am grateful to Alvin Goldman, Joshua Knobe, Sarah Nelson, and Fred Schueler for their comments on an earlier draft.

Notes

1. Awareness is not actually an enabling factor; it only ensures that the agent monitors her action and executes it in such a way as to fulfill the intention she has. Awareness therefore has no explanatory function for intentional actions. "Why did he leave the room?" is not answered by "He was aware of doing it." However, lack of awareness can be used to explain why an intended action remained unsuccessful.

2. Throughout this chapter I use female pronouns for agents and male pronouns for explainers.

3. The question of how people select the reasons they cite from these many possibilities has not been studied in great detail (Hesslow 1988). General parameters include the explainer's knowledge (e.g., O'Laughlin and Malle 2000) and his assumptions about the audience's knowledge (e.g., Hilton 1990; Turnbull and Slugoski 1988). Another parameter is the motive to present an image of the agent as, say, rational or moral (Malle, Knobe, O'Laughlin, Pearce, and Nelson 2000).

4. For detailed coding rules, see Malle 1998.

5. Compare "Why did the gun accidentally go off?" "John was trying to clean it." (Wellman 1990, p. 99)