Self and Other in the Explanation of Behavior: 30 Years Later

Joshua Knobe and Bertram F. Malle
Princeton University University of Oregon

Published in *Psychological Belgica, 42*, 113-130.

Abstract

It has been hypothesized that actors tend to attribute behavior to the situation whereas observers tend to attribute behavior to the person (Jones & Nisbett 1972). The authors argue that this simple hypothesis fails to capture the complexity of actual actor-observer differences in people's behavioral explanations. A new framework is proposed in which reason explanations are distinguished from explanations that cite causes, especially stable traits. With this framework in place, it becomes possible to show that there are a number of distinct actor-observer asymmetries in explanation, each stemming from a distinct psychological process by which explanations are generated.

People can explain behavior from either of two basic perspectives. In *actor explanations*, a behavior is explained by the person who actually performed the behavior. In *observer explanations*, a behavior is explained by someone other than the person who performed the behavior.

Now, suppose we pose the question: "What psychological processes might lead actor explanations to differ from observer explanations?" A sensible answer would be: "All sorts of different processes. Actors and observers attend to different kinds of events, they have access to different kinds of information, and they are influenced by different kinds of motivations."

Thirty years ago, Jones and Nisbett (1972) advanced the hypothesis that all of these differences in psychological processes ultimately lead to a single difference in the behavior explanations that actors and observers provide. In other words, even though Jones and Nisbett endorsed the claim that the processes that guide actors' and observers' explanations differ in a wide variety of respects, they suggested that each of these process differences independently leads to the same behavioral effect. The purported effect was quite simple. Observers, it was claimed, tend to attribute a person's behavior to factors that lie within that person, whereas actors tend to attribute behavior to factors that lie in the external situation.

The Jones-Nisbett hypothesis has played an extremely important role in subsequent attribution research, being put to the test in numerous studies throughout the next decade (e.g., Arkin, & Duval, 1975; Goldberg, 1981; Herzberger & Clore 1979; Nisbett, Caputo, Legant, & Marecek, 1973; Lenauer, Sameth & Shaver, 1980; Regan, & Totten, 1975). By 1982, when Watson published an influential review, he felt it was safe to conclude that the hypothesis "now appears to be firmly established" (Watson 1982, p. 698). Future studies, he claimed, should focus not so much on whether the basic effect was really there (since that had already

been conclusively demonstrated) but on the various factors that might enhance, eliminate or reverse this effect (cf. Robins, Spranca, & Mendelsohn, 1996).

In this article, we want to re-examine this widely accepted conclusion. Jones and Nisbett pointed to a number of important psychological processes that lead actors and observers to provide different sorts of explanations. But — as we shall try to show — it was a mistake to suppose that all of these psychological processes have one single effect on actors' and observers' explanations of behavior. To do justice to the complexity of the phenomena at issue, we need to link each psychological process to its corresponding effects on the various aspects of the explanatory process. The result is a much more nuanced and multi-dimensional view of actor-observer asymmetries in explanation.

Traits and Reasons

We begin with an important conceptual distinction. Jones and Nisbett distinguished "situation attributions" from another type of attribution in which behavior is explained in terms of factors that lie within the agent who actually performed the behavior. This other type of attribution is often called a "person attribution," but it has also been referred to as a "dispositional attribution."

It is unclear, however, precisely what the term "dispositional" is supposed to mean in this context (Ross & Fletcher, 1985). The term might be used to refer to any factor that lies within the person (including emotions, traits, beliefs, sensations, and so forth; Heider, 1958), or it might be used to refer specifically to relatively stable person factors such as personality traits (Jones & Davis, 1965). But, of course, if the term is understood in this latter, more restrictive sense, it will not be possible to classify all explanations into the categories of

situation and disposition. (How would one classify an explanation like: “He ran away because he thought there was a bear in the room?”)

We therefore avoid the ambiguous term “dispositional” and adopt a more fine-grained analysis in which explanations that refer to personality *traits* are distinguished from explanations that refer to other factors that lie within the person, such as emotions, occurrent thoughts, bodily states, and so on.

Among these other person factors, one deserves special attention. People’s explanations of intentional behavior most commonly refer to *reasons* – i.e., to the beliefs, desires and valuings in light of which the agent[1] formed an intention to act. (Buss, 1978; Malle, 1999; Read, 1987; Schueler 2001). To see the difference between reasons and other types of explanations, consider two possible responses to the question “Why did Jessie punish her son?”

(1) “Because she thought that he was the one who broke the window.”

(2) “Because she is a tyrant.”

Notice that only the first of these explanations is an attempt to capture Jessie’s own decision-making. Jessie herself might have thought: “My son broke the window; therefore I should punish him.” But she certainly would not have thought: “I am a tyrant; therefore I should punish my son.” In fact, even if Jessie’s tyrannical personality somehow led her to punish her son, she was probably not aware of this trait at the time when she decided to perform her action (Malle, Knobe, O’Laughlin, Pearce & Nelson, 2000). In this sense, it can be said that explanation (1) gives Jessie’s *reason* for punishing her son whereas explanation (2) gives, not Jessie’s reason, but a factor that lies in the *causal history* of her reason. Causal

history explanations can refer to a variety of factors, including emotions, bodily states, or—as in this case—the agent's traits.

When the distinction between reasons and other person factors is properly heeded, doubt is cast on the view that there is a general tendency for observers to give more person attributions than actors do. In fact, we will argue that the whole concept of "person attributions" serves no useful function in the study of actor-observer asymmetries. Nothing is gained by lumping all explanations that refer to person factors into a single category, since (as we show below) there are different actor-observer asymmetries for different factors that lie within the person. Specifically, it appears that observers use more trait explanations but that actors use more reason explanations.

Psychological Processes Underlying Actor-Observer Asymmetries

We now develop in more detail our argument that there are several actor-observer asymmetries in behavior explanations rather than only one. We consider each of the specific psychological processes that Jones and Nisbett (1972) proposed to account for the purported tendency of observers to use more person attributions. These psychological processes, we believe, are real and important. However, one cannot derive from them the hypothesis that observers give more person attributions than actors do. In fact, we will argue that the different processes posited by Jones and Nisbett lead to a variety of different actor-observer asymmetry in explanations. This section develops the theoretical argument; the next section reviews supportive empirical evidence.

Attentional Processes

Jones and Nisbett (1972) suggest that the agent's behavior is especially salient to observers. In itself, this seems like a very plausible claim. It is widely agreed that observers attend more to the agent's behavior than they do to most aspects of the external situation (Heider 1958) and that observers attend more to the agent's behavior than actors do themselves (Malle & Pearce, 2001). After all, observers get direct visual information about the agent's behavior that simply isn't available to actors.

But Jones and Nisbett added another wrinkle to this basic claim: They argued that, since the observer attends more to the agent's behavior, the observer will be more likely to attribute that behavior to factors within the agent himself.

Later researchers tested this hypothesis by directly manipulating an agent's salience and then measuring the degree to which explainers tended to give more person attributions for that agent. The results were inconsistent at best. Storms (1973) found that explainers gave more person attributions for more salient agents, but subsequent studies either failed to replicate this effect (Taylor & Fiske, 1975; Uleman, Miller, Henken, Tsemberis, & Riley, 1981) or found that explainers actually gave *fewer* person attributions for more salient agents (McArthur & Post, 1977; Taylor, Fiske, Close, Anderson, & Ruderman, 1977).

There is, however, a deeper problem with the Jones-Nisbett account. If observers attend to the agent's behavior, why should they be especially likely to explain this behavior in terms of psychological factors that lie within that agent? Ultimately, it seems that an explainer would only be especially likely to explain the agent's behavior in terms of psychological factors if she attended to those psychological factors themselves. So the basic question here is not which sort of explainer — actor or observer — attends most to the agent's observable

behaviors; the question is which sort of explainer attends most to psychological factors that lie within the agent.

When we turn to an examination of these psychological factors, we need to distinguish carefully between traits and reasons. Though observers may be especially attentive to the agent's traits (as attribution researchers have argued), it does not seem plausible to suggest that observers attend more to reasons than actors do. The actor cannot fail to attend to her own reasons, since these were the reasons in light of which she chose to perform the action.

But if actors are especially attentive to their own reasons, they should immediately think of these reasons when they are called upon to explain their own actions. Actors should therefore offer more reason explanations than observers do, a prediction that flatly contradicts the traditional claim that actors are *less* likely than observers to explain behavior in terms of person factors.

Information Access

As Jones and Nisbett point out, actors often have detailed information that simply is not available to observers. This greater access to information, they claim, leads actors to make more situational attributions than observers do. Thus, Jones and Nisbett consider the case of a person who reacts to an insult with a seemingly extreme burst of rage. An observer might conclude (falsely) that the agent is a brash and irritable person. Only the actor herself is aware that she reacted so strongly to this insult because it was the latest in a long series, the "straw that broke the camel's back" (Jones & Nisbett, 1972, p. 84). More generally, Jones and Nisbett claim that, since actors have more information than observers do, they should be able to make complex situational attributions that observers would not be able to construct.

Experimental tests, however, show no evidence that greater knowledge leads to more situation attributions. Neither Nisbett et al. (1973) nor Taylor and Koivumaki (1976) found any correlation between knowledge of the agent and tendency to make situational attributions, and Hampson (1983) showed that the use of traits in describing others increases, rather than decreases, with familiarity. It is therefore necessary to revisit the question of how differences in information access lead actors and observers to offer different explanations.

First of all, it is essential to be clear about precisely what type of information observers are supposed to lack. We need to distinguish at least three major types — information about the agent's situation, information about the agent's traits, and information about the agent's reasons. Actors typically have more access to all three types of information, but the extent of this informational advantage differs from one information type to the next. In some types, actors have far more information than observers do; in others, actors have only slightly more information than observers do.

Actors often have only a slight advantage in situational information, at least when actor and observer are both participating in the same social situation. Actors have more trait information than observers in cases where the observer barely knows the agent, but when observers know the agent well they may actually have more (or more accurate) trait information than the agent herself (Funder, 1991). The actor's greatest informational advantage is in knowledge of the agent's reasons. Actors typically know (or believe they know) the reasons for their actions (Donellan, 1967), whereas observers are often completely unaware of the agent's reasons.

If information access advantages drive actor-observer asymmetries in explanation, we should expect actors to give more reason explanations than observers do. After all, when

observers have no idea what the agent's reasons are, they cannot use reason explanations and must resort to some other form explanation. Actors, by contrast, almost always believe they know their own reasons and can therefore give reason explanations for almost any action they wish to explain.

Motivational Mechanisms

Jones and Nisbett make two claims about motivational mechanisms: (a) that actors show something akin to "reactance" toward explaining their own behavior in terms of traits and (b) that although actors and observers have different attitudes toward the agent, this difference in attitudes should not alter the basic actor-observer asymmetry.

Reactance. Jones and Nisbett draw on Brehm's (1966) reactance theory, according to which people prefer to see themselves as free agents unfettered by confining forces. They then claim that reactance leads people to avoid explaining their own behavior in terms of traits: "The perception of freedom is probably best maintained by simultaneously ascribing traits to others and denying them in oneself" (Jones & Nisbett, 1972, p. 92). The idea seems to be that, insofar as an agent attributes her own behavior to stable traits, she admits that she was not free to do whatever she chose.

Ironically, however, this psychological mechanism should also discourage actors from explaining their own behavior in terms of the *situation.* Surely, if people prefer not to see themselves as controlled by their own traits, they will be even more reluctant to see themselves as controlled by factors in the external situation. (Indeed, that was precisely the lesson of Brehm's original theory.) It therefore seems highly unlikely that reactance would lead actors to choose situational explanations over trait explanations.

To take a simple example, consider three possible ways in which an actor might explain her decision to clean her desk:

- [*Situational causal history*] "Because, when I was young, my parents always made sure that I kept everything organized."
- [*Trait causal history*] "Because I am compulsively neat."
- [*Reason explanation*] "Because I thought it would help me study more effectively."

Here it appears that the situation explanation would generate the most reactance, the trait explanation slightly less, and the reason explanation none at all. By emphasizing her own goals and beliefs, the agent clearly does not make herself seem like a captive; instead, she makes herself seem all the more free and in control (Bergmann 1977; Miller & Norman, 1975). In general, then, if actors chose the form of explanation most consistent with a view of themselves as free agents, it seems that they would explain their own behavior using reasons.[2]

The explainer's attitude toward the agent. People tend to believe that their own beliefs are true, that their own traits are good and that their own actions are justified. But observers often have no such bias. They are perfectly willing to conclude that the agent holds false beliefs, has undesirable traits, or performs unjustified actions. Might this difference in attitude play a role in people's use of explanations?

It does seem possible that certain modes of explanation tend to pin the full responsibility for the behavior on the agent whereas other modes of explanation tend to separate the agent from the behavior, making it appear that the agent does not deserve much credit or blame for the behavior she performed. But, as Jones and Nisbett point out, this process cuts both ways. The actor would tend to assume full responsibility for positive

behaviors but would try to avoid responsibility for negative behaviors. Hence, on their view, attitude toward the agent does not produce some general effect whereby actors tend to prefer a particular type of explanation. Rather, it leads actors to prefer one type of explanation for positive behaviors and another for negative behaviors, canceling out any effect across all behaviors.

This conclusion seems to us to be fundamentally sound. Attitude toward the agent does not, in general, lead actors or observers to prefer one particular type of explanation. However, attitude toward the agent does affect the way actors and observers express their explanations in language. And, as we show below, this difference in linguistic expression may mislead researchers into thinking that actors and observers differ along the traditional person-situation dimension.

Self-Other Asymmetries at Three Levels of Analysis

Jones and Nisbett pointed to a number of important psychological processes that lead actors and observers to offer different types of behavioral explanations. But when we re-examined each of these processes in turn, we found that none of them seemed likely to produce an asymmetry along the traditional person-situation dimension. In fact, a number of these processes seemed specifically to prevent the traditional asymmetry from arising.

We therefore propose a more complex account. Instead of claiming that all of the psychological processes posited by Jones and Nisbett lead to a single effect, we suggest that these processes lead to a variety of different effects in a number of different aspects of the explanatory process. Indeed, we tie each of the processes to a different effect:

- Asymmetries in attention lead to an asymmetry in which behaviors people explain.

- Asymmetries in information access lead to an asymmetry in people's judgments about why behaviors occurred.
- Asymmetries in attitude toward the agent lead to an asymmetry in the way people express their explanations in language.

This analysis still leaves open the question of why previous researchers found the person-situation differences postulated by Jones and Nisbett. An additional goal is therefore to account for these reported effects in terms of methodological and linguistic artifacts.

Level 1: Event Selection

Jones and Nisbett suggested that actors and observers typically attend to different types of events. This suggestion has been confirmed by recent research. In a series of studies, Malle and Pearce (2001) found that participants in a social interaction were especially likely to attend as observers to the other person's *actions* (i.e., to observable, intentional behavioral events) but as actors, to their own *experiences* (i.e., their unobservable, unintentional behavioral events). Still, it should not be presumed that this asymmetry in attention directly leads to an asymmetry in the ways people explain behavior. Even though observers attend carefully to the agent's actions, they rarely explain actions in terms of other actions, and even though actors attend carefully to their own experiences, they rarely explain those experiences in terms of other experiences. The key asymmetry here is a difference that lies, not in the explanations people provide, but in the kinds of behaviors they select to explain.

In a typical attribution experiment, subjects are instructed to provide a causal attribution for an event selected by the experimenter (e.g., "What caused him to be so nervous?"). But in ordinary life people do not try to explain arbitrarily selected events. Rather,

they usually try to explain the most salient events that they do not understand, and these events are of a different kind for actors and observers. Malle and Knobe (1997) showed that actors tend to explain their own experiences whereas observers tend to explain the agent's actions.

In attribution experiments, then, actors and observers may focus on different behavioral events even when they are specifically instructed to provide attributions for the same behavior. Consider an experiment in which subjects are asked to explain first why they themselves were nervous and then why their conversation partners were nervous. When asked about themselves, they may focus more on the *experience* of nervousness; when asked about the conversation partner, they may focus more on nervous *actions* (fidgeting, avoiding certain conversational topics, etc.). Thus, although actors and observers appear to be providing attributions for precisely the same type of behavior, they may in fact be concerned with behaviors of two very different types. Of course, this analysis does not imply that all previous actor-observer findings can be accounted for by differences in the kinds of behaviors people explain. It does imply, however, that actor-observer asymmetries in the selection of behaviors explained must be clearly distinguished from actor-observer asymmetries involving different explanations for the same type of behavior.

Level 2: Explanatory Judgments

Actors are more likely to know their reasons than observers are. Moreover, since the agent's reasons actually figured in that agent's decision making, these reasons should be especially salient and available to the agent himself. One might therefore predict that, compared to observers, actors will be especially likely to explain their behavior using reasons (Buss, 1978; Locke & Pennington, 1982).

Reasons, however, are not the only mode of explanation for intentional behavior. Explainers may also choose to describe the "causal histories of reasons" – those factors that provide the background and origin of the agent's reasons. To return to an earlier example, if we ask Jessie's son why his mother punished him, he might say something like "Because she's a tyrant." This explanation does not offer Jessie's reason for punishing her son; it gives a potential causal history of Jessie's reasons. Since actors have an especially great advantage in knowledge of reasons, they are unlikely to explain their behavior using causal histories (except when there are specific demands not to use reasons; Malle, 2001; O'Laughlin & Malle, 2002). Observers, however, are sometimes unable to infer the agent's reasons. One might therefore predict that, compared to actors, observers will more often resort to causal history explanations.

To test these predictions, we (Malle, Knobe, & Nelson, 2002) examined actors' and observers' explanations in a variety of contexts, including memory protocols, natural interactions, and structured interviews. Some studies asked people to recall why-questions and their corresponding explanations, others identified spontaneous explanations in conversation. Some studies let people choose the behaviors they explained, others preselected those behaviors. All in all, we conducted six studies with a total of over 600 participants and over 5000 explanations (Malle et al., 2002; see also Malle, 2002).

The predictions were unequivocally supported in each of the six studies. Overall, actors offered about one and a half times as many reasons as observers did, whereas observers offered about twice as many causal histories as actors did.

Further analyses were conducted on the types of reason explanations and the types of causal history explanations people offered. Each reason explanation was classified as either a

belief reason, a desire reason, or a valuing (though this last category was rather infrequent). The results showed that actors offered far more belief reasons than desire reasons, whereas observers offered slightly more desire reasons than belief reasons. Like the reason/causal history asymmetry, this asymmetry might be due to knowledge differences. Belief reasons typically represent idiosyncratic information such as perceived circumstances, anticipated outcomes, and considered alternatives, which are rarely known by observers. Desire reasons, by contrast, can often be inferred from general social rules and practices (Bruner, 1990; Bartsch & Wellman, 1989).

In addition, each causal history explanation was classified as either a situation factor, a stable trait, or a nontrait person factor (e.g., emotions, bodily states, past behaviors). The results showed that observers did give a higher number of trait explanations. But this effect emerged only because observers offered more causal histories of *all* types. There was no tendency for observers to specifically give trait causal histories rather than other types of causal histories (Malle et al., 2002).

Level 3: Linguistic Expression

When people formulate behavior explanations in language, they not only express their explanatory judgments but also indicate their evaluative attitude toward the agent, his action, and his reasons. Compare (a) “He did it because he wanted x” with (b) “He did it because *for some reason* he wanted x.” In both cases, the behavior is explained in terms of a desire for x, but in the second case, the explainer tries to indicate that she believes the desire for x to be somehow foolish or unreasonable.

In a similar way, the explainer can choose between explanations like (a) “She is taking an umbrella because it’s going to rain” and (b) “She is taking an umbrella because *she thinks*

it's going to rain." In both cases, the explainer accounts for the agent's behavior in terms of her belief that it is going to rain.[3] But in the first case, the explainer endorses the agent's belief — he implies that he too believes it is going to rain — whereas in the second, the explainer distances himself from the agent's belief. By explicitly stating that the agent believes it is going to rain, the explainer suggests that perhaps he has some doubt as to the truth of the agent's belief.

We refer to expressions like "he believed," "they thought," and "she assumed" as *mental state markers* (Malle, 1999). These expressions explicitly mark the subsequent clause (e.g., "…that it is going to rain") as the content of the agent's mental state. By using mental state markers, explainers can highlight the agent's own reasoning process and thereby mark the difference between the agent's beliefs and their own. Thus, in the example above, the expression "she thought" serves to mark the clause about the rain as the content of the agent's belief and, at the same time, makes it clear that the explainer herself may not share that belief.

This distinction between marked and unmarked belief reasons forms the basis of another actor-observer asymmetry, this time at the level of linguistic expression. Compared with observers, actors more often wish to portray their actions in a rational light. And how better to portray oneself in a rational light than to give unmarked belief reasons, thereby stating a "fact" that supports one's action? Indeed, research has shown that when people are specifically instructed to portray themselves as rational, they more often use unmarked belief reasons (Malle et al., 2000).

One might therefore predict that, compared to observers, actors should more often use unmarked belief reasons. We tested this prediction in the six studies referred to above and

found that, on average, actors and observers used the same number of marked beliefs but actors used twice as many unmarked beliefs as observers did (Malle et al., 2002).

This asymmetry in the use of mental state markers is not only an interesting phenomenon in its own right; it also provides a new way of accounting for results that seemed to support the classic person-situation hypothesis. Consider again the agent who decides to bring her umbrella because she thinks it's going to rain. If asked to explain her behavior, she may say "…because it's going to rain," whereas an observer (who does not necessarily share the agent's belief) may say instead "…because *she thinks* it's going to rain." In cases like these, the actor and the observer may hold exactly the same hypothesis about the causal process that led up to the action; the only difference is in the way they choose to express that hypothesis in language. And yet, if researchers code the resulting explanations into person and situation categories by simply looking at the words that appear in the explanation, they will code the former as a "situation attribution" (since it only mentions the rain) and the latter as a "person attribution" (since it explicitly mentions the agent's belief). The result will be an apparent actor-observer asymmetry in person vs. situation attributions that, in fact, reflects a difference in the linguistic expression of explanations (Antaki, 1994; Malle et al., 2000; Ross, 1977).

Interim Conclusion

In his attempt to explain the appeal of the classic actor-observer hypothesis, David Watson wrote: "The interest generated by the Jones-Nisbett hypothesis is undoubtedly due, in part, to its simple and bold formulation" (Watson, 1982, p. 683). We agree; the appeal of the Jones-Nisbett hypothesis undoubtedly stems, in large part, from its simplicity. But when the phenomena themselves are extremely complex, there is no point in trying to account for them

with a simple hypothesis. The differences between actors and observers are manifold — including differences in events explained, differences in beliefs about the explanation of those events, and differences in the linguistic expression of these explanations — and there is simply no way to capture them all in one simple formula.

The Epistemology of Self and Other

Thus far, we have focussed on differences in attention, information access, and attitudes toward the agent — all processes that are related with but not identical to the process of constructing an explanation. However, even if all these processes were well-understood, we still would not have answered the question of how people actually go about constructing a behavior explanation. Moreover, to return to the central concern of this special issue, we still would not know whether actors and observers construct explanations using the same basic process or whether they use two very different processes.

In their investigation of this question, psychologists have often proceeded on the assumption that people's main goal in constructing explanations is to identify sources of variance (Fiedler, Walther, & Nickel, 1999; Försterling, 1992; Kelly, 1967; Van Overwalle, 1997). The question then becomes whether actors and observers account for variance using the same basic process (e.g., Nisbett & Wilson, 1977) or whether actors often use a process that simply isn't available to observers (e.g., White, 1980).

We do not intend to contribute to this debate here. Rather, we want to emphasize once again that people employ a number of distinct modes of explanation and that these modes of explanation differ from each other in truly fundamental ways. It would therefore be a mistake to assume that *all* explanation is a matter of searching for sources of variance in behavior. The

more accurate view would be that different modes of explanation are based on different kinds of information and are constructed using different processes.

In particular, it seems clear that reason explanations should not be construed as attempts to identify sources of variance. Thus, when Margaret says “I’m taking an umbrella because it’s going to rain,” she does not mean to assert that the future rain explains the variance in her present umbrella-taking behavior. What she means to assert is simply that her reason for taking an umbrella is that it is going to rain.

Similarly, suppose that we ask a person “Why are you rummaging through that cupboard?” and she answers “Because I want to find the oregano.” The agent is not thereby asserting that her desire to find the oregano explains the variance in her cupboard-rummaging behavior. In fact, she might be well aware that, when she doesn’t need oregano, she often ends up rummaging through the cupboard to find some other herb. So she might readily acknowledge that her quest for oregano explains none of the variance in her general cupboard-rummaging behavior. But that has no bearing on her original explanation: In this instance, the reason she rummaged was because she wanted to find the oregano.

Now, if we asked an agent to identify sources of variance in her behavior, it might turn out that she would proceed by drawing on shared implicit theories (as Nisbett & Wilson, 1977, claim), or it might turn out that she would draw on information that isn’t available to observers (as White, 1980, claims). However, it is important to understand that, in ordinary circumstances, when agents explain their behavior using reason explanations, they are not identifying sources of variance at all. So the various competing theories about how actors and observers identify sources of variance cannot tell us how actors and observers generate reason explanations.

Rather, the question as to how actors and observers generate reason explanations should be regarded as an open and intriguing question, worthy of investigation in its own right. Do actors and observers both generate reason explanations by drawing on implicit theories (Gopnik, 1993)? Or do actors and observers both make use of a specialized mental module, the Theory of Mind Mechanism (Scholl & Leslie, 1999)? Or could it be that actors draw on their own memories whereas observers use a process of simulation (Gordon, 1986; Goldman, 1989, 2001)? Not even a rudimentary understanding of these complex phenomena can be achieved, however, until we carefully distinguish reason explanations from explanations that give the causal history of an agent's reasons and from explanations that refer to causes of unintentional behavior.

Conclusion

Traditional attribution theory holds that the principal aim of the explainer is to identify sources of variance in behavior. On this view, the explainer is faced with the question as to whether the variance in a given behavior should be attributed to the person or to the situation. Within this general framework for understanding explanations, it was easy to understand the appeal of Jones and Nisbett's (1972) thesis: that actors tend to attribute behavior to the situation whereas observers tend to attribute behavior to the person.

We have tried to sketch a very different framework for understanding folk explanations of behavior. This new framework gives a central role to the distinctions among various types of explanations. Above all, it distinguishes between *reason explanations* and other modes of explanations that involve causes alone.

When we approach the issue of actor-observer asymmetries within this new framework, there is no longer any meaningful question as to whether actors and observers

generally attribute variance to the person or to the situation. A different set of questions becomes meaningful: How do actors and observers choose which events to explain? How do they choose whether to give reason explanations or causal history explanations? When do they choose to offer beliefs and when desires? And how do they choose whether or not to linguistically mark belief reasons? Through an investigation of these questions, we can gain insight into the principal differences between actors' and observers' behavior explanations and, hence, the psychological processes underlying the perception of self and other.

References

Antaki, C. (1994). *Explaining and arguing: The social organization of accounts.* Thousand Oaks, CA: Sage.

Arkin, R.M., & Duval, S. (1975). Focus of attention and causal attributions of actors and observers. *Journal of Experimental Social Psychology, 11,* 427--438.

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

Bergmann, F. (1977). *On being free*. Notre Dame: University of Notre Dame Press.

Brehm, J. (1966). *A theory of psychological reactance.* New York: Academic Press.

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

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

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

Fiedler, K., Walther, E, & Nickel, S. (1999). Covariation-based attribution: On the ability to assess multiple covariates of an effect. *Personality and Social Psychology Bulletin, 25,* 607-622.

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

Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science, 2,* 31-39.

Goldberg, L. R. (1981). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality and Social Psychology*, *41*, 517-552.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1-14.

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

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

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language, 1*, 158-71.

Hampson, S. E. (1983). Trait ascription and depth of acquaintance: The preference for traits in personality descriptions and its relation to target familiarity. *Journal of Research in Personality, 17*, 398-411.

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

Herzberger, S. D., & Clore, G. L. (1979). Actor-observer attributions in a multitrait-multimethod matrix. *Journal of Research in Personality, 13*, 1-15.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 371-388). Hillsdale, NJ: Erlbaum.

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

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

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

Lenauer, M., Sameth, L. & Shaver, P. (1976). Looking back at oneself in time: Another approach to the actor-observer phenomenon. *Perceptual and Motor Skills*, *43*, 1283-1287.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*, 21-43.

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

Malle, B. F. (2002). The social self and the social other. Actor-observer asymmetries in making sense of behavior. In J. P Forgas & K. D. Williams (Eds.), *The social self: Cognitive, interpersonal, and intergroup perspectives* (pp. 189-204). Philadelphia, PA: Psychology Press.

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

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

Malle, B. F., Knobe, J., & Nelson, S. E. (2002). *Actor-observer asymmetries in folk explanations of behavior: New answers to an old question.* Unpublished Manuscript, University of Oregon, Eugene.

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

McArthur, L. Z., & Post, D. L. (1977). Figural emphasis and person perception. *Journal of Experimental Social Psychology, 13,* 520-535.

Miller, D. T, & Norman, S. A. (1975). Actor-observer differences in perceptions of effective control. *Journal of Personality and Social Psychology, 31,* 503-515.

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

Nisbett, R. E., & Wilson, T.D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84,* 231--259.

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

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

Regan, D. T., & Totten, J. (1975). Empathy and attribution: Turning observers into actors. *Journal of Personality and Social Psychology, 32,* 850-856

Robins, R. W., Spranca, M. D., & Mendelsohn, G. A. (1996). The actor-observer effect revisited: Effects of individual differences and repeated social interactions on actor and observer attributions. *Journal of Personality and Social Psychology, 71,* 375-389.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 174-221). New York: Academic Press.

Ross, M., & Fletcher, G. J. O. (1985). Attribution and social perception. In G. Lindsey & E. Aronson (Eds.), *The Handbook of Social Psychology* (Vol. 2, pp. 73-114). New York: Random House

Scholl, B. J, & Leslie, A. M. (1999). Modularity, development and 'theory of mind.' *Mind & Language, 14,* 131-153.

Schueler, G. F. (2001) Action explanations: Causes and purposes. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.). *Intentions and intentionality: Foundations of social cognition* (pp. 251-264). Cambridge, MA: MIT Press.

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

Taylor, S. E., & Fiske, S. T. (1975). Point-of-view and perceptions of causality. *Journal of Personality and Social Psychology, 32,* 439-445.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249-288). New York: Academic Press.

Taylor, S. E., & Koivumaki, J. H. (1976). The perception of self and others: Acquaintanceship, affect, and actor-observer differences. *Journal of Personality and Social Psychology, 33,* 403-408.

Van Overwalle, F. (1997). A test of the joint model of causal attribution. *European Journal of Social Psychology, 27,* 221-236.

Taylor, S. E., Fiske, S.T., Close, M., Anderson, C. & Ruderman, A. (1977). *Solo status as a psychological variable: The power of being distinctive.* Unpublished manuscript. Harvard University. (As cited in Taylor & Fiske, 1978.)

Uleman, J. S., Miller, F. D., Henken, V., Riley, E., & Tsemberis, S. (1981). Visual perspective or social perspective?: Two failures to replicate Storms' rehearsal, and support for Monson and Snyder on actor–observer divergence. *Replications in Social Psychology, 1*, 54-58.

Watson, D. (1982). The actor and the observer: How are their perceptions of causality divergent? *Psychological Bulletin, 92*, 682-700.

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

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review, 87*, 105-112.

Author Notes

Joshua Knobe, Department of Philosophy, Princeton University; Bertram F. Malle, Department of Psychology, University of Oregon.

Parts of this article were written while both authors were guests at the University of Canterbury, Christchurch, New Zealand. Many thanks to the deparment and, in particular, to Garth Fletcher for hosting us. We are also grateful to Susan Fiske, Sarah Nelson, Dan Rothschild, and Bas van Fraasen for discussions or comments on an earlier draft.

Correspondence should be addressed to Joshua Knobe, Department of Philosophy, 1879 Hall, Princeton University, Princeton, New Jersey 08544-1006, E-mail: jknobe@princeton.edu, or to Bertram F. Malle, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403-1227, E-mail: bfmalle@darkwing.uoregon.edu.

Endnotes

[1] Throughout this paper, we have adopted the following terminological conventions: The *agent* is the person who performs the action; the *explainer* is the person who explains the action. If the agent and the explainer are the same person, this person is called the *actor*. If the agent and the explainer are two different people, the person who offers the explanation is called the *observer*. In addition, to avoid pronoun confusion, we always refer to the agent as *she* and, in observer explanations, to the explainer as *he*.

[2] Jones (1978) later seemed to arrive at a similar conclusion about the role of freedom, when he says:

> If you ask people, did the situation make you do this, they'll say no. But if you ask them if they wanted to do this because of the nature of the situation, they'll say yes, because they have control then. (Jones, 1978, p. 378 or 379)

Jones makes a distinction here between reasons on one hand ("I wanted to do this because of the situation") and situation causes or traits on the other hand (cf. Buss, 1978). Unfortunately, neither Jones nor the subsequent attribution literature reconciled this distinction with the simplified situation-disposition dichotomy. The introduction of reasons alone should have led to revised predictions about actor-observer differences. But there are even more revisions necessary, as we argue in the next section.

[3] At first blush, one might think that the first explanation ("because it's going to rain") simply cites a situation cause. But that would be a mistake: No explainer would assume that the rain in the future could somehow retroactively cause the agent to bring her umbrella.