18

Moral, Cognitive, and Social *The Nature of Blame*

BERTRAM F. MALLE, STEVE GUGLIELMO, and ANDREW E. MONROE

Brown University

Blame is a moral judgment that has a cognitive and a social nature. We first focus on the cognitive side and introduce a theoretical model of blame that integrates insights and evidence from extant research. Within this model, we demonstrate the critical role of such concepts as *agent*, *intentionality*, *and obligation*—all of which are grounded in people's theory of mind. We then contrast two views on the ordering of blame and theory of mind based inferences: blame-late models, which claim that blame follows mental state inferences; and blame-early models, which claim that the opposite order holds. After integrating these two views within our model, we turn to two eminently social topics of moral judgment: blaming as a social act; and blaming of group agents. We suggest that our model of cognitive blame provides a fruitful framework for both of these topics, thus highlighting the intimate connection between blame as a cognitive phenomenon and blame as a social phenomenon.

Humans blame, and perhaps only humans do. But what is blame? Blame is grounded in the capacity to have a "theory of mind"[1]—a system of concepts and processes that aid a human social perceiver in inferring mental states from behavior. To blame an agent people must know a set of behavior-guiding norms, observe an agent's norm-violating behavior, and infer a manifold of mental states that underlie the behavior. Without a theory of mind, an organism may still be able to punish, but it would not be able to blame the way humans do.

A second unique feature of blame is that it has not only a cognitive side (i.e., processes that lead up to a *judgment* of blame), but it also has an interpersonal side (i.e., observable *acts* of social blaming). The latter requires language,

Y127289.indb 313 12/6/11 11:54 AM

communication, and the ability to consider other people's responses and once more relies on a theory of mind. The social roots of both theory of mind and moral judgment are deep. The major function of theory of mind is to enable individuals to coordinate social interaction (e.g., Goody, 1995). Likewise, the major function of moral judgment is to regulate social behavior, especially in more complex groups where deviations from group norms can be costly. Blaming deals precisely with such norm deviations. And if people did not blame socially, they would do little to regulate each other's behavior or to pass on their culture's values (Kashima, this volume). But how people *arrive* at acts of blame is a question largely about cognition.

A MODEL OF BLAME

Humans do not make moral judgments about earthquakes or hurricanes. Judgments are moral if they are directed at *agents* who are presumed to be capable of following socially shared norms of conduct. Thus, the initial components in the emergence of blame are as follows (Figure 18.1):

1. Detecting that some event or outcome[2] deviated from a norm
2. Assessing that an *agent* was involved and *caused* this event
3. Deciding whether the agent brought about the event *intentionally*

Once this latter decision has been made, two very different tracks lead to blame. Along the left track in Figure 18.1, if the agent is believed to have acted intentionally:

4a. Perceivers consider the agent's *reasons* for acting. Blame then is graded depending on the justification these reasons provide—minimal blame if the agent was justified in acting this way; maximal blame if the agent was not justified.

Along the right track in Figure 18.1, if the agent is believed to have acted unintentionally:

4b. Perceivers consider whether the agent should have prevented the norm-deviating event (*obligation*).
5. Perceivers consider whether the agent could have prevented the event (*capacity*).

We now discuss in detail each of these hypothesized components. We have called this a "step model of blame" (Guglielmo, Monroe, & Malle, 2009) because several information processing components build on each other (e.g., intentionality is irrelevant if no agent caused the event) and will be temporally ordered (e.g., assessment of reasons must follow assessment of intentionality; Malle, 1999, 2004). As with all complex information processes, however, there may be room for default values, omitted steps, and motivated reasoning (cf. Johnson & Carpinella;

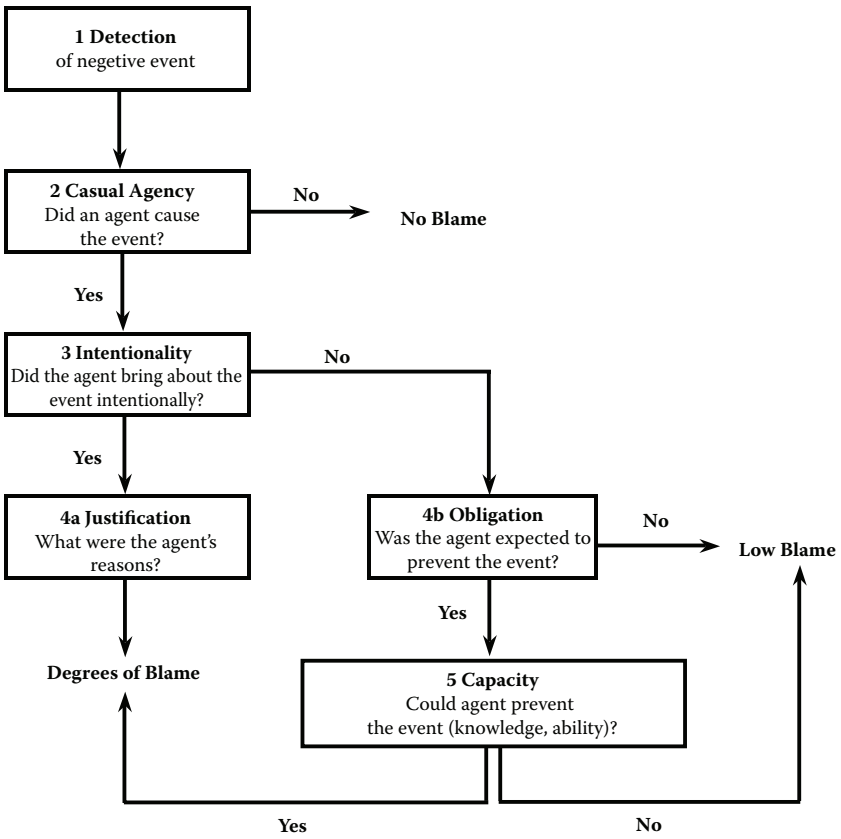

Figure 18.1 Step model of ordinary assessments of blame.

von Hippel, this volume), and research will have to establish both frequency and impact of such complications.

Detection

En route to blame, perceivers must first detect a negative event—an event that deviates from a norm (e.g., norms to prevent harm or uphold fairness; Graham, Haidt, & Nosek, 2009).

People are highly sensitive to norm deviations. Such events trigger rapid evaluative responses (Van Berkum, Holleman, Nieuwland, Otten, & Murre, 2009), and, compared with positive or neutral events, negative events command more attentional resources, are more widely represented in language, and exert a stronger impact on both self-perception and interpersonal behavior (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Taylor, 1991).

Detection of norm-deviating events may rely on "moral intuitions" or rules of "moral grammar" (Haidt, 2001; Mikhail, 2007). These intuitions may activate the moral judgment machinery by flagging events that are potentially worthy of moral

judgment. People may also have immediate negative affect following the detection of such badness; and one function of affect may be to make norm deviations salient. But whatever affect people feel at this early point is directed at the negative *event*. Something bad happened, and there is no information yet about why it happened and who, if anyone, is to blame (Pomerantz, 1978). The mere detection of a norm-deviating event does not yet constitute a blame judgment.

Blame stems in part from assigning meaning to an event. Finding meaning answers a why question, resolves the tension of uncertainty by filling a gap in understanding and restoring coherence and control. People experience nagging why questions for a variety of events, but particularly for negative ones (Malle & Knobe, 1997b; Wong & Weiner, 1981). Explaining the origin of negative events is essential, but in principle such explaining could be done without blaming. Genuine blame requires meaning of a particular kind—one that involves agents who caused the negative event.

Agent Causality

Social perceivers blame people, not physics. So for mere outcomes (e.g., a dead person, a broken window) to lead to genuine blame, the perceiver must first establish whether an agent caused the outcome (Shaver, 1985; Sloman, Fernbach, & Ewing, 2009). Numerous studies have demonstrated the crucial impact of agent causality in assigning blame (Cushman, 2008; Lagnado & Channon, 2008), for social perceivers from age 5 on (Shultz, Wright, & Schleifer, 1986).

When the norm-deviating event is an observed action rather than simply an outcome, the recognition of agent causality comes for free, because only agents perform actions. Moreover, some negative outcomes (such as a gunshot wound) strongly imply the causal involvement of a human agent, whereas many others require further investigation (such as a broken arm or a bloody lip).

What do people categorize as an "agent"? The agency concept, emerging early in infancy, relies on features such as self-propelledness and contingent action (Johnson, 2000; Premack, 1990). That is not enough, however, to qualify as a *morally eligible* agent. The exact criteria that make an agent morally eligible deserve a more detailed analysis elsewhere. However, it seems clear that moral eligibility requires the ability to understand and remember norms and the ability to modify one's behavior in accordance with these norms—to *choose* to follow the norms (Guglielmo et al., 2009). If such abilities are absent (e.g., in infancy or in certain mental or physical illnesses), blame either will not be assigned or will be decisively mitigated.

Our theory postulates that an agent's causal involvement falls into two fundamentally different categories—intentional and unintentional (Heider, 1958; Malle, 2004; White, 1995). Thus, judgments of intentionality make up the critical third step in our model of blame.

Intentionality

The capacity to recognize a behavior as intentional is a central component of human social cognition. The origins of the intentionality concept lie in infants' ability to

recognize some motion as goal directed (Wellman & Phillips, 2001; Woodward, 1998) and to segment the motion stream into units that correspond to intentional actions (Baldwin, Baird, Saylor, & Clark, 2001). By the age of 2, children acquire the concept of desire, recognize that another person can have desires different from their own (Repacholi & Gopnik, 1997) and infer an agent's desires even from incomplete action attempts (Meltzoff, 1995). Over the next few years, children acquire the concept of belief, grasp the purely mentalistic nature of false belief, and later, not before the age of 6, differentiate intentions from desires (Astington, 2001; Baird & Moses, 2001). This differentiation eventually culminates in an adult concept of intentionality that encompasses five components—desire, belief, intention, skill, and awareness (Malle & Knobe, 1997a).

Even though the adult concept of intentionality consists of five components and people are sensitive to the presence or absence of each of these components (Guglielmo & Malle, 2010a, 2010b; Malle & Knobe, 1997a, 2001), we should not expect people to deliberate about these components each time they judge a behavior intentional. Instead, people quickly perceive intentionality in everyday situations (Barrett, Todd, Miller, & Blythe, 2005) but consider carefully each of the components if uncertainty or the weight of the judgment demands it.

Intentionality judgments regulate attention in social interaction. As actors, people attend more to their own unintentional (both behavioral and mental) events; as observers they attend more to the other person's intentional events (Malle & Pearce, 2001). Intentionality judgments also guide explanations and predictions of behavior (Heider, 1958). Most importantly, to account for intentional and unintentional behavior people use distinct modes of explanation, which differ in conceptual, cognitive, and linguistics properties (Malle, 2004, 2011).

Of primary interest here is the role that intentionality plays in moral judgment. Children begin to incorporate intentionality into their moral judgment by the age of 5 (Shultz et al., 1986). Though they are considerably influenced by outcome severity, they understand that doing something bad intentionally is worse than doing it unintentionally (Darley, Klosson, & Zanna, 1978). For adults, the intentionality distinction in moral judgment requires little cognitive effort (Solan, 2003) because perception of most behaviors already comes with a judgment of intentionality, often perceptually (Scholl & Tremoulet, 2000) or as part of scripts (Schank & Abelson, 1977).

In whatever way the perceiver arrives at a judgment of intentionality, plenty of evidence shows that people blame intentional norm violations more severely than unintentional ones (e.g., Darley & Shultz, 1990; Ohtsubo, 2007). Our model therefore asserts that intentionality amplifies blame, but it also importantly holds that intentionality judgments bifurcate the perceiver's processing of norm-violating events. Thus, even if the negativity of the event has already been evaluated, proper blame cannot be assigned to the *agent* until the question of intentionality has been answered. People then search for and are sensitive to rather different information when encountering intentional as opposed to unintentional negative events. The two paths 4a and 4b in Figure 18.1 delineate this differential information processing, which we now describe.

Reasons and Justification

When people judge a behavior intentional (left path in Figure 18.1), they consider the agent's particular reasons for acting. This is something people do with ease, and they find it painful *not* to know the reasons of someone's action (Malle, 2004). Often the reasons for undesirable actions will themselves be undesirable (e.g., monetary gain, or a specific goal to injure the victim), and selfish or vengeful reasons make moral evaluations particularly severe (Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002). But sometimes the agent's reasons provide justification for the action and thereby cast it in a less blameworthy light (e.g., Howe, 1991). A schoolboy who hurts another may do it because he defends his sister against a bully (a justified reason) or because he tries to provoke a fight (an unjustified reason), and he will be blamed more in the latter case.

Unintentional Events

People's explanations for unintentional behavior are far simpler than those for intentional behavior. In interpreting intentional behavior, people use three distinct modes of explanation (Malle, 1999, 2004, 2011). One mode requires consideration of the agent's subjective reasoning (reason explanations); another focuses on the causal background of those reasons, such as personality, culture, and context (causal history of reason explanations); and a third specifies objective enabling factors that allowed the agent to successfully complete the intended action (enabling factor explanations). Unintentional behaviors are explained by a single mode: causes, which more or less mechanically bring about the behavior without any involvement of reasoning, intentions, or the like. However, when people consider *blaming* unintentional behaviors, their processing becomes quite complex. They go beyond the backward-looking explanations of the behavior and enter forward-looking considerations of the behavior's potential recurrence and any prospects of its prevention. Though instances of blame have backward-looking (e.g., retributive) elements, the overall function of blame, and especially its social expression, is primarily forward looking (reformative) because it is one of the community's tools to regulate behavior.

Thus, when people regard the agent as having unintentionally brought about a negative outcome, they face two central questions: whether the agent *should have* prevented the outcome (obligation) and *could have* prevented the outcome (capacity). Both of these conditions are grounded in the intentionality concept. A social community imposes obligations on its members to prevent negative outcomes under the assumption that they are able to intentionally act (or at least intend to act) in accordance with these obligations. This ability is a moral eligibility condition, which a 1-year old will fail but a 5-year old will pass. If moral eligibility holds, a failure to follow one's obligation to prevent a negative outcome will trigger substantial blame. However, at times an outcome is not preventable—even if the person tried, she could not intentionally avert it. Such a capacity limitation can be a principled one (e.g., lack of physical strength) or a local one (e.g., in the particular context, the events unfolded too quickly).

Evidence for the Impact of Obligation

Most studies of moral judgment hold obligation constant—they typically contain stories in which the agents unquestionably have an obligation to prevent negative outcomes. As a result, little direct evidence is available for obligations' impact on blame. When they have been examined, however, obligations have shown considerable influence. Hamilton (1986) reported that people in higher positions of a social hierarchy are subject to stronger obligations for preventing negative outcomes and are blamed more for those outcomes when they occur. Such role position effects were also found when causality was ambiguous (Gibson & Schroeder, 2003) or when it was indirect, as in cases of vicarious responsibility (Shultz, Jaggi, & Schleifer, 1987).

Evidence for the Impact of Capacity

Shultz et al. (1987) showed that people who have control over others (e.g., parents over their children) are held responsible for the wrongdoings of those whom they control—because they have not only the obligation but the capacity to prevent the harmdoing. Prevention capacity comes in two shades: the *cognitive* capacity to foresee the negative outcome; and the *physical* capacity to actually avert the outcome. Both are necessary. The impact of foreseeability has been demonstrated in adults as well as children from age 4 on (e.g., Nelson-Le Gall, 1985; Shaw & Sulzer, 1964), and Weiner (1995) reviewed numerous studies in which the agent's ability to control an outcome is a strong predictor of blame. These capacities are unlikely to be judged all-or-none. An agent may have foreseen some parts of a negative outcome but not its full scope, and an agent may have had the capacity to take some preventive steps, but perhaps not the most effective ones. As a result, variation in assumed capacity will translate into degrees of blame (Figure 18.1).

Vicarious Blame

Pet owners are sometimes blamed for damage caused by their pets; parents for damage caused by their children; and company management for accidents in the workplace. Such vicarious blame applies only when—following the unintentional path—obligation and capacity to prevent are plausible. But vicarious blame seemingly violates the causality requirement (step 2 in our model), because the one who is blamed (e.g., the pet owner) did not actually cause the negative event (e.g., the dog biting a child in the park). However, people accept causation by omission and thus consider the pet owner blameworthy for *allowing* his pit bull to roam around the park, which led to the child being bitten by the dog. Within counterfactual theories of causation, this is not a surprising claim: If only the owner had put the dog on a leash, it wouldn't have bitten the child (Dowe, 2001).

COMPETING MODELS

The literature on moral judgment contains numerous models of the antecedents, psychological processes, and consequences of such judgments (Guglielmo, 2011). These models can be roughly categorized into two groups, which are divided by one key disagreement (Guglielmo & Malle, 2011)—whether blame judgments *follow* mental state judgments (which we label blame-late models) or *precede*

mental state judgments (which we label blame-early models). Our step model is a blame-late model, and we begin with a discussion of other models of this kind.

Blame-Late Models

Blame-late models propose that judgments of blame critically rely on prior assessments of an agent's causal involvement and mental states. For example, "entailment models" posit that certain early judgments serve as necessary conditions for subsequent judgments, the last of which is that of blame (Fincham & Jaspars, 1980; Shaver, 1985; Weiner, 1995). Although these models offer somewhat different accounts of the precise judgments that precede blame, they generally agree that blame critically depends on prior assessments about the extent to which an agent caused the negative event in question, did so intentionally, had certain characteristics mental states, or had the ability to produce a different outcome. In the absence of these assessments, according to blame-late models, it makes no sense to consider the blameworthiness of an agent's behavior.

Our step model of blame has affinities with these entailment models but also has some notable differences. Entailment models have an intervening concept of responsibility—one that follows causality assessments and precedes blame assessments—whereas we believe that such a step is unnecessary. The primary reason is that responsibility is a hopelessly equivocal concept (Fincham & Jaspars, 1980; Hamilton & Sanders, 1981; Hart, 1968). It collapses distinct phenomena under a single label and is often confounded with other phenomena. A recent study shows that at least four distinct constructs are subsumed under, or comeasured with, responsibility: causality, foreknowledge, intentionality, and wrongfulness (Gailey & Falk, 2008). In addition, the term *responsibility* has been used to refer to an agent's obligation (Hamilton, 1986), to a person's eligibility for moral judgment (Oshana, 2001), and simply to blame (e.g., Shaw & Sulzer, 1964; Shultz, Schleifer, & Altman, 1981). Conversely, some studies purport to measure blame by asking participants about responsibility. But because responsibility is such a vague concept, responsibility measures are less sensitive than blame measures to a variety of determinants of moral judgment, including intention, justification, and foreseeability (e.g., Critchlow, 1985; McGraw, 1987). For all these reasons we omit responsibility from our model.

Studies that have explicitly measured blame judgments paint a consistent picture of the influence of mental state assessments on these judgments. Darley and Shultz (1990) show that, whereas agents receive some blame when they foresee but fail to prevent harm (e.g., through negligence or recklessness), they receive much more blame when they intentionally bring about the harm. Recent studies demonstrate that people assign far more blame to an agent for causing a negative outcome intentionally (e.g., cutting off a pedestrian, burning a stranger's hand) than for causing the identical outcome unintentionally (e.g., Cushman, 2008; Lagnado & Channon, 2008; Ohtsubo, 2007). Furthermore, to the extent that people perceive the agent's reasons to be selfish or vengeful, they make more severe moral judgments about the agent (Reeder et al., 2002; Woolfolk, Doris, & Darley, 2006). Mental states are also critical for the mitigation of blame—for example, blame is

reduced when the agent lacks knowledge (Nelson-Le Gall, 1985) or had desirable goals (Howe, 1991).

Cushman's (2008) model highlights the importance of causal and mental inferences preceding moral judgments, but the model also offers an important distinction between two kinds of moral judgments. On the one hand, people judge the *wrongness* of an agent's behavior, and these wrongness judgments are driven entirely by assessments of the agent's mental states (primarily beliefs and desires). Thus, people deem an agent's behavior especially wrong when the agent believes his behavior will bring about a negative outcome and wants this outcome to occur (regardless of whether the negative outcome actually occurs). On the other hand, people assess an agent's *blameworthiness* by also taking into account the actual consequences of the agent's behavior—whether a negative outcome in fact occurred. In this way, an agent receives more blame for an action that happens to have bad consequences than for one that does not, holding constant the agent's mental states (Mazzocco, Alicke, & Davis, 2004; Robbennolt, 2000). However, mental state judgments still critically guide blame: intentionally bringing about a negative outcome (i.e., when both belief and desire are present) is blamed much more than unintentionally causing that outcome (i.e., when the agent lacks both belief and or desire).

Cushman's (2008) model and our step model share important features, but they do differ in ways that will require reconciliation. First, wrongness is a cousin of blame but takes a different object of judgment. It is the *intentions* behind an action that are judged as more or less wrong; it is *agents* who are blamed for what they cause in the world. The step model of blame may thus recast wrongness judgments as *blame for intentions* (whether turned into action or not). This would predict equivalence between people's judgments of how *wrong* an agent's intention was and how much *blame* the agent deserves for adopting that intention. Currently no evidence is available on whether such an equivalence holds.

Second, Cushman's (2008) model does not specify exactly how people deal with unintentional behaviors. His model predicts different degrees of blame as a function of mental state inferences, but the information processing that operates on those inferences is not spelled out. Our step model maps out two very different paths of information processing—consideration of beliefs and desires as reasons for intentional actions (like Cushman's model describes) and considerations of obligations and capacities in evaluating unintentional behavior (factors that Cushman's model does not identify).

Blame-Early Models

A second class of moral judgment models proposes that blame occurs prior to (and can therefore influence) assessments of causality and mental states. Haidt (2001) suggests that people have immediate moral intuitions upon considering negative behaviors: "One feels a quick flash of revulsion at the thought of incest and one knows intuitively that *something is wrong*" (p. 814, emphasis added). People also make moral *judgments*, which are "evaluations (good vs. bad) of the actions or

character *of a person*" (p. 817, emphasis added), and, in turn, "moral judgment is caused by quick moral intuitions" (p. 817).

Alicke (2000) offers an even more explicit claim about the impact of blame on mental state judgments. "People use outcome information as a basis for ascribing blame and they then justify their attributions by altering their judgments of the a priori criteria" (Alicke, Davis, & Pezzo, 1994, pp. 283–284). These a priori criteria include the critical components of blame-late models, such as assessments of an agent's causal role, intentions, foresight, and motives. Thus, people engage in a process of "blame validation," whereby their initial blame judgments serve to guide their subsequent assessments about the content of the agent's mental states.

Knobe's (2010) moral pervasiveness model makes a somewhat different claim. On this model, judgments about causality and mental states still guide blame judgments, as they do for the blame-late models previously discussed. However, an "initial moral judgment" (Phillips & Knobe, 2009) precedes and directs this causal and mental analysis. Consequently, "moral judgment is pervasive, playing a role in the application of *every* concept that involves holding or displaying a positive attitude toward an outcome" (Pettit & Knobe, 2009, p. 593). Thus, Knobe's model is more properly conceptualized as a moral–judgment–early model, rather than a blame-early model, but we will group it under the latter heading because it still posits that moral judgments precede mental state judgments.

The evidence for these blame-early models comes in many forms. Haidt has shown that people have early evaluative reactions when considering negative behaviors and that they are often "dumbfounded" when attempting to verbally justify their moral evaluations (Haidt & Hersh, 2001; cf. Dijksterhuis, this volume). Alicke has shown that the negativity of an agent or an outcome can influence people's judgments about the agent's causal role or negligence in producing the outcome. In one study, Alicke (1992) found that a character who was speeding to hide cocaine was judged more blameworthy and more causally responsible for his ensuing car accident than was a character who was speeding to hide a gift for his parents. The early evaluation of the "extra-evidential" motive for speeding influences people's blame judgments, even though it should be irrelevant for assessing the agent's blameworthiness for the accident. Finally, studies by Knobe and others suggest that, compared with positive or neutral actions, people judge negative actions as more intentional (Knobe, 2003), caused (Knobe & Fraser, 2008), and foreseen (Beebe & Buckwalter, 2010).

There are several reasons to doubt, however, that these findings support a blame-early model. Haidt typically measures what we have labeled detections of norm deviation ("this is morally wrong"), not judgments of *blame* (e.g., Haidt & Hersh, 2001; Wheatley & Haidt, 2005). More important, Haidt's studies do not vary information about causality, intentionality, or mental states, thus making it impossible to examine whether or not moral judgments precede and influence those nonmoral judgments.

The results of Alicke's (1992; Alicke et al., 1994) studies may arise from an informational impact of negativity rather than a motivational one, because negative information provides particularly diagnostic evidence of a person's dispositions

(Reeder & Brewer, 1979; Skowronski & Carlston, 1989). People may rightly infer that agents who produce particularly negative outcomes do so with more causal involvement or more negative mental states. For example, it may be reasonable to infer that the drug-hiding agent from Alicke's (1992) study is in fact more reckless or careless than the gift-hiding agent, and that inference may mediate the further inference of a greater causal role in bringing about the accident. But such inferences of care, recklessness, or other mental states are not measured either in Alicke's studies or numerous other studies that investigated effects of outcome severity on blame judgments (for a review, see Robbenholt, 2000). In one exception (Fincham, 1982), outcome severity in fact predicted mental state inferences (desire to damage), and these inferences in turn predicted blame judgments—consistent with a blame-late model.

Finally, follow-up research on Knobe's intriguing findings of valence effects on intentionality judgments has shown identical patterns for behaviors that lack moral content (Machery, 2008; Uttich & Lombrozo, 2010), suggesting that Knobe's results can be explained by people's sensitivity to norm deviations more generally, not moral violation in particular. Knobe's effects of valence on intentionality judgments also disappear once several confounds (e.g., the agent's desire, the action's difficulty) are properly controlled (Guglielmo & Malle, 2010a, 2010b). And just as studies of Alicke's model have not empirically assessed spontaneous evaluations, studies of Knobe's model also have not empirically assessed the "initial moral judgments" that are claimed to influence mental state inferences. Consequently, the key claim of these models is, at present, not well supported.

Resolving the Debate

Can the distinct claims and findings of the two classes of blame models be integrated? Such an integration is indeed possible, and it relies on the distinction between early event-focused *reactions* and later person-focused *judgments* (Guglielmo & Malle, 2011; Monin, Pizarro, & Beer, 2007). People often experience negative affect upon detecting negative events—death, environmental damage, and so on—but this affect turns into a moral judgment (e.g., blame) only after people analyze the causal and mental-state structure of the event. This causal and mental analysis serves as an emotion appraisal (Lazarus, 1984) and gives meaning to the perceiver's early affective response, thereby transforming evaluations of an event into moral judgments of an agent. This would mean that blame-early models are reinterpreted as event evaluation models, and those event evaluations lead—via causal and mental state inferences—to genuine blame.

Any test of such an integrative model requires the distinct assessment of event reactions and agent blame, which have never been measured separately in the extant literature. We would also need to assess the time course and speed of each of these processes. Information about causality and mental states must then be made available in time-sequenced steps to demonstrate that event evaluations come online before any of this specific information is understood but that agent blame comes online only after the specific information is understood.

BLAMING AS A SOCIAL ACT

Moving Beyond Purely Cognitive Models

Extant psychological models of blame focus almost exclusively on the intrapersonal processes of arriving at blame judgments. But people do more than blame others in their own heads. Blame is expressed in face, body, and language; it is doled out, countered, and negotiated. Without such social expression, blame would do little to achieve one of its major functions: to regulate social behavior. A comprehensive theory of blame must be able to delineate the antecedents and consequences of such social acts of blaming.

One social aspect of blaming is featured in Haidt's (2001) "social intuitionist" model, according to which people who express a moral judgment exert direct influence on other people's moral intuitions. "If your friend is telling you how Robert mistreated her, there is little need for you to think systematically about the good reasons Robert might have had. The mere fact that your friend has made a judgment affects your own intuitions directly" (p. 820). However, according to Haidt, people rarely have access to the emergence of their moral judgments (they are "dumbfounded" by their intuitions), so it is not apparent what they can say during their social expression of blaming beyond, "This is wrong, he is bad." If people cannot consciously retrieve grounds for their judgments, they would not be able to argue about, negotiate, and justify moral judgments—which, we believe it is apparent, they do.

Steps Toward Social Blame

Our model of blame specifies the kind of information that can be used in negotiation and justification. We assume that people normally have access to the contents of several judgments: the negativity of the outcome, the agent's suspected causal involvement, intentionality, obligations, and various inferred mental states (of, e.g., intention, reasons, knowledge). This information is available for justifying, contesting, and negotiating a public moral claim. We see this process most clearly in the courtroom, where causality, intentionality, obligation, and knowledge have to be explicitly "proven" for a verdict to ensue.

But social acts of blaming are addressed not only to other observers but also to the perpetrator, especially by the victim of a transgression. In Duff's (1986) idealized version of blame, the blamer engages the perpetrator in a moral deliberation, with the ultimate goal to change the perpetrator's behavior on the basis of remorse, insight, and recommitment to the very values that he had violated. Even in a less ideal world, perpetrator and blamer communicate about the basis of the blame (Pearce, 2003), debating the very components that are specified in the step model of blame: Did I cause it? Did you do it intentionally? Should you have prevented it? Could I have prevented it? The step model thus provides a useful initial framework to examine some of the informational and conceptual components in social acts of blaming—directed at either the perpetrator or other observers.

A theory of social blaming must address under what conditions people express blame. On the one hand, people may be less likely to express blame in the presence of the perpetrator, because blaming may be costly (the perpetrator may lash out) or prohibited (e.g., by norms of role, status or relationship). On the other hand, blaming the perpetrator directly may actually reform the person's behavior. Reform is more likely to occur when there is a preexisting relationship between blamer and perpetrator and when the blamer not only condemns the other person's behavior but also appeals to shared values that have been violated; thus, the blamer attempts to make the person recognize the wrongness of his actions (Duff, 1986). Such "persuasive" blaming presumes that the perpetrator endorses the shared values, and it shows respect for the perpetrator's rationality to change his behavior accordingly. Persuasive blaming may even signal the blamer's willingness to listen to the perpetrator's own perspective and to consider possible justifications for the offending behavior. Persuasive blame thus has the power to repair the strained relationship between blamer and perpetrator (Bennett, 2002), a power that is harnessed in restorative justice procedures (Kuo, Longmire, & Cuvelier, 2010).

These virtues of persuasive blaming are necessarily absent in third-person blaming, which is addressed to other observers. With little chance of reforming the perpetrator, it serves primarily to express the blamer's emotions, to reassert the violated norms, and to seek validation for those norms (Duff, 1986; Pearce, 2003). Third-person blaming can arise from an inability to reform (e.g., because the perpetrator is too high in status to be directly addressed) or from the blamer's refusal to even attempt any reform. In the latter case, the act of blaming may represent the first step toward socially excluding the perpetrator (Kurzban & Leary, 2001). Blaming a suitable target, especially an outsider, can in fact increase the coherence of a group and aid in the collective endeavor of making sense of negative events (Treichler, 1999). One of the cruelest examples is the Nazi propaganda to blame Jews for the economic crisis and cultural "ills" of Germany in the 1930s. This propaganda led both to increased group coherence (nationalism and wide support for the Nazi party) and to the brutal escalation of legalized social exclusion all the way to genocide. Importantly, the propaganda claimed specific causal, even intentional, contributions of Jews to the society's woes. It was not just an irrational lashing out of negative affect; it was a systematic "argument" that adhered to the informational and conceptual components of blame.

No doubt, people sometimes blame irrationally and unfairly. They dispense of all reform and argument and instead express community-sanctioned hatred, as in the shocking practice of lynching (Dray, 2002). Whether such acts of hate should count as "blame" is debatable. But people consider these acts as deeply unjust precisely because they wholly ignore the foundational questions of blame: Was the agent causally involved? Did he act intentionally? Could he have prevented the outcome?

BLAMING GROUPS

People treat not only individuals as moral agents; they also treat groups that way if the group has the abilities of forming reasons and acting intentionally in light of

these reasons. Our model of blame should therefore generalize to the evaluation of group actions, which would mean that each step in the model can be applied to groups (Malle, 2010). Indeed, people easily detect norm-violating group behaviors (Malle, 2004, Chapter 8); they have no trouble making intentionality judgments about group behavior (Dillon & Malle, 2011); and they ascribe reasons to group agents (O'Laughlin & Malle, 2002). Along the path of blame for unintentional behavior (e.g., accidents due to a corporation's negligence), groups are certainly subject to obligations, and questions of capacity (both knowledge and ability to act) figure prominently in legal and public discussions of corporate liability. Thus, without making contentious assumptions, it seems plausible that people blame group agents using the same cognitive apparatus they use when blaming individuals.

Social Blaming of Groups

This correspondence between blaming individuals and groups at the cognitive level extends only partially to the social level. Interesting problems arise with the social blaming of groups. Here is the first: How well can blame for group agents be expressed? Social perceivers do not actually encounter nations, governments, or corporations; even teams or committees are rarely seen face to face. In modern life, people can write letters to a group agent, sue them, or publicly denounce them. But these expressions will be infrequent, limited in scope, and come with little assurance that the addressee actually notices or cares about the blame.

The second problem is this: If blame is rarely expressed and even more rarely heard, regulation of group agents' behavior may run idle. A social perceiver can vote against a government or refuse to buy from a company, but here she alters her own actions more than the group agent's actions. Only when individual social perceivers aggregate or join together can social blame become an effective regulator. It often takes a group agent to fight a group agent.

A third problem is that group agents lack (or at least are perceived to lack) most affective mental states (Knobe & Prinz, 2008), so they will also be unlikely to feel guilt, regret, or remorse. As a result, groups will have fewer moral scruples, which further blocks social regulation as well as deterrence. If groups are rational, solely cognitive agents, potential blame or punishment becomes part of the utility calculation for their actions; anticipated guilt or regret lies outside these calculations.

CONCLUSIONS

The nature of blame is both cognitive and social. On the cognitive side, people progress through considerations of causality, intentionality, and mental states to arrive at judgments of blame, which are thus fundamentally grounded in a theory of mind. Importantly, people blame intentional and unintentional events in distinct ways, taking into account justifying reasons for intentional events but prevention obligation and prevention capacity for unintentional events. Reconciling the conflicting positions of blame-early and blame-late models, we maintained that early evaluations in response to a norm-deviating event should not count as blame

because they are event directed whereas genuine blame is agent directed. Such blame can then be socially expressed either to the agent or to other social observers, but research has only begun to examine the conditions under which such social blame occurs. Finally, people appear to blame groups with the same cognitive apparatus as they blame individuals, but the social expression of group blame faces distinct obstacles that also await future research.

ENDNOTES

1. For the present audience, the term *theory of mind* is the most commonly used, but near synonyms are *folk psychology* or *common-sense psychology.* For a discussion see Malle (2005, 2008).
2. Events are time-extended processes (e.g., a car skidding on ice), whereas outcomes are the results of events (e.g., the car having crashed into a tree). For our present purposes this distinction is not important.

REFERENCES

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368–378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

Alicke, M. D., Davis, T. L., & Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition*, *12*, 281–308.

Astington, J. W. (2001). The paradox of intention: Assessing children's metarepresentational understanding. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 85–103). Cambridge, MA: MIT Press.

Baird, J. A., & Moses, L. J. (2001). Do preschoolers appreciate that identical actions may be motivated by different intentions? *Journal of Cognition & Development*, *2*, 413–448.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*, 708–717.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*, 313–331.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323–370.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind and Language*, *25*, 474–498.

Bennett, C. (2002). The varieties of retributive experience. *Philosophical Quarterly*, *52*, 145–163.

Critchlow, B. (1985). The blame in the bottle. *Personality and Social Psychology Bulletin*, *11*, 258–274.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*, 525–556.

Darley, J. M., Klosson, E. C., & Zanna, M. P. (1978). Intentions and their contexts in the moral judgments of children and adults. *Child Development*, *49*, 66–74.

Dillon, K. D., & Malle, B. F. (2011). *A robust hierarchy of social inferences across individual and group agents.* Paper presented at the annual meeting of the Society of Philosophy and Psychology, Montreal, Canada.

Dowe, P. (2001). A counterfactual theory of prevention and "causation" by omission. *Australasian Journal of Philosophy*, *79*, 216–226.

Dray, P. (2002). *At the hands of persons unknown: The lynching of Black America*. New York: Random House.

Duff, A. (1986). *Trials and punishments*. Cambridge, England: Cambridge University Press.

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of responsibility: From man the scientist to man as lawyer. *Advances in Experimental Social Psychology*, *13*, 81–138.

Fincham, F. D. (1982). Moral judgment and the development of causal schemes. *European Journal of Social Psychology*, *12*, 47–61.

Gailey, J. A., & Falk, R. F. (2008). Attribution of responsibility as a multidimensional concept. *Sociological Spectrum*, *28*, 659–680.

Gibson, D. E., & Schroeder, S. J. (2003). Who ought to be blamed? The effect of organizational roles on blame and credit attributions. *International Journal of Conflict Management*, *14*, 95–117.

Goody, E. N. (Ed.). (1995). *Social intelligence and interaction: Expressions and implications of the social bias in human intelligence.* Cambridge, UK: Cambridge University Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

Guglielmo, S. (2011). *Moral judgment: An integrative review*. Manuscript under review.

Guglielmo, S., & Malle, B. F. (2010a). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition*, *117*, 139–150.

Guglielmo, S., & Malle, B. F. (2010b). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*, 1635–1647.

Guglielmo, S., & Malle, B. F. (2011). *Mind over morality: Mental-state inferences (still) guide moral judgments.* Manuscript under review.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry: An Interdisciplinary Journal of Philosophy*, *52*, 449–466.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31*, 191–221.

Hamilton, V. L. (1986). Chains of command: Responsibility attribution in hierarchies. *Journal of Applied Social Psychology*, *16*, 118–138.

Hamilton, V. L., & Sanders, J. (1981). The effect of roles and deeds on responsibility judgments: The normative structure of wrongdoing. *Social Psychology Quarterly*, *44*, 237–254.

Hart, H. L. A. (1968). *Punishment and responsibility*. New York: Oxford University Press.

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

Howe, E. S. (1991). Integration of mitigation, intention, and outcome damage information, by students and circuit court judges. *Journal of Applied Social Psychology*, *21*, 875–895.

Johnson, S. C. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences*, *4*, 22–28.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190 –194.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–329.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology (Vol. 2): The cognitive science of morality: intuition and diversity* (Vol. 2, pp. 441–447). Cambridge, MA: MIT Press.

Knobe, J., & Prinz, J. J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, *7*, 67–83.

Kuo, S.-Y., Longmire, D., & Cuvelier, S. J. (2010). An empirical assessment of the process of restorative justice. *Journal of Criminal Justice*, *38*, 318–328.

Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin*, *127*, 187–208.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*, 754–770.

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*, *39*, 124–129.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, *23*, 165–189.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*, 23–48.

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

Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 225–255). New York: Oxford University Press.

Malle, B. F. (2008). The fundamental tools, and possibly universals of social cognition. In R. M. Sorrentino & S. Yamaguchi (Eds.), *Handbook of motivation and cognition across cultures* (pp. 267–296). New York: Elsevier/Academic Press.

Malle, B. F. (2010). The social and moral cognition of group agents. *Journal of Law and Policy*, *20*, 95–136.

Malle, B. F. (2011). Time to give up the dogmas of attribution: A new theory of behavior explanation. In M. P. Zanna & J. M. Olson (Eds.), *Advances of Experimental Social Psychology* (Vol. 44, pp. 297–352). San Diego, CA: Academic Press.

Malle, B. F., & Knobe, J. (1997a). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

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

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 45–67). Cambridge, MA: MIT Press.

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

Mazzocco, P. J., Alicke, M. D., & Davis, T. L. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology*, *26*, 131–146.

McGraw, K. M. (1987). Guilt following transgression: An attribution of responsibility approach. *Journal of Personality and Social Psychology*, *53*, 247–256.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*, 838–850.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*, 143–152.

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology*, *11*, 99–111.

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, *1*, 211–224.

Nelson-Le Gall, S. A. (1985). Motive-outcome matching and outcome foreseeability: Effects on attribution of intentionality and moral judgments. *Developmental Psychology*, *21*, 323–337.

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.

Ohtsubo, Y. (2007). Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect. *Japanese Psychological Research*, *49*, 100–110.

Oshana, M. (2001). Responsibility: Philosophical aspects. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13,279–13,283). Oxford: Pergamon.

Pearce, G. E. (2003). *The everyday psychology of blame*. Doctoral dissertation, Department of Psychology, University of Oregon.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, *24*, 586–604.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry, 20*, 30–36.

Pomerantz, A. (1978). Attributions of responsibility: Blamings. *Sociology*, *12*, 115 –121.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, *36*, 1–16.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*, 61–79.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology*, *83*, 789–803.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*, 12–21.

Robbennolt, J. K. (2000). Outcome severity and judgments of "responsibility": A meta-analytic review. *Journal of Applied Social Psychology*, *30*, 2575–2609.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*, 299–310.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer Verlag.

Shaw, M. E., & Sulzer, J. L. (1964). An empirical test of Heider's levels in attribution of responsibility. *Journal of Abnormal and Social Psychology*, *69*, 39–46.

Shultz, T. R., Jaggi, C., & Schleifer, M. (1987). Assigning vicarious responsibility. *European Journal of Social Psychology*, *17*, 377–380.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *13*, 238–253.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, *57*, 177–184.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*, 131–142.

Sloman, S. A., Fernbach, P., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In D. Bartels, C. Bauman, L. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making* (pp. 1–26). Boston, MA: Academic Press.

Solan, L. M. (2003). Cognitive foundations of the impulse to blame. *Brooklyn Law Review*, *68*, 1003–1029.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, *110*, 67–85.

Treichler, P. A. (1999). *How to have theory in an epidemic: Cultural chronicles of AIDS*. Durham, NC: Duke University Press.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*, 87–100.

Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological Science*, *20*, 1092–1099.

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

Wellman, H. M., & Phillips, A. T. (2001). Developing intentional understandings. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 125–148). Cambridge, MA: The MIT Press.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*, 780–784.

White, P. A. (1995). *The understanding of causation and the production of action: From infancy to adulthood. Essays in developmental psychology.* Hillsdale, NJ: Erlbaum.

Wong, P. T., & Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, *40*, 650–663.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, 1–34.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*, 283–301.