Moral, Cognitive, and Social: The Nature of Blame

Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe
Brown University

Humans blame; and perhaps only humans do. But what is blame? And what makes it uniquely human?

For one thing, 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 norms, observe an agent's norm-violating behavior, and infer a manifold of mental states that underlie the behavior. Without the latter, an organism may still be able to punish; but the organism would not be able to blame.

A second unique feature of blame is that it has not only a cognitive side—processes that lead up to a *judgment* of blame; but it also has a social side—observable *acts* of blaming. The latter requires language, communication, and the ability to anticipate other people's responses, which once more relies on a theory of mind.

In this chapter we will focus on the cognitive side of blame (as has the entire literature) and introduce a theoretical model that integrates insights and evidence from interdisciplinary work on blame. In particular, we will demonstrate the critical role of such concepts as *agent*, *intentionality*, *reasons*, and *choice*—all of which lie at the core of a theory of mind. In addition, we will analyze a recent debate on the relationship between the process of assigning blame and the processes of judging intentionality and mental states—a debate that will also touch on the broader question of the role of affect in moral judgment. We close with some first steps of exploring social acts of blame, suggesting that our cognitive model provides a useful framework in this exploration.

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 first steps in the emergence of blame are (see Figure 1):

(1) Detecting that some negative outcome or event deviated from a shared norm.

(2) Assessing that an agent caused this outcome or event.

But humans are not satisfied with establishing causality; they take a further step:
(3) Deciding whether the agent brought about the event intentionally.

Once this decision has been made, two very different tracks lead to blame. Along the left track in Fig. 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 that 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 Fig. 1, if the agent is believed to have acted unintentionally,
(4b) Perceivers consider whether the agent *should* have prevented the norm-violating event (obligation) and (5) whether the agent *could* have prevented the event (capacity).

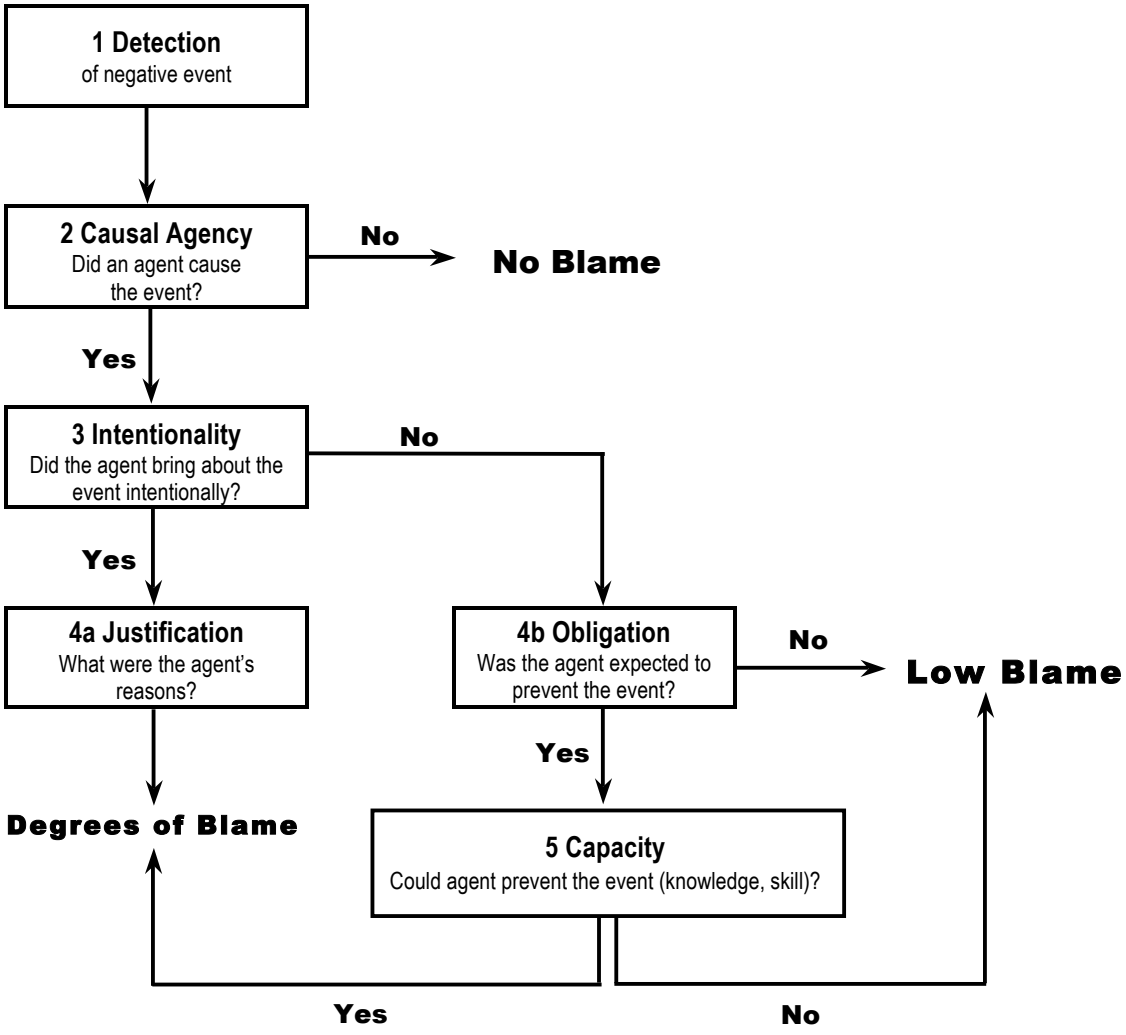

Figure 1. Step model of ordinary assessments of blame.

We discuss now in detail each of these hypothesized components or steps. 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 causality has been established) and will often be temporally ordered (e.g., assessing reasons must follow assessment of intentionality; Malle, 2004). As with all complex information processes, however, there may be room for backward loops, premature processing, and omissions, and research will have to establish both frequency and impact of such deviations.

Detection

En route to blame, perceivers first must detect a norm violation. That is, negative outcomes or events2 are recognized or interpreted as damage (e.g., a scratched car door) or harm (an injured dog), or as something bad, uncomfortable, or disgusting (Felstiner, Abel, & Sarat, 1980). Detection of such a norm-violating event does not yet constitute a moral judgment. People may have immediate negative affect, but whatever affect they feel at this stage is outcome-directed. *Something* bad happened, but there is no information yet about why it happened and who, if anyone, is responsible (Pomerantz, 1978). The negative affect may co-occur with the detection of a norm deviation (in fact, this is possibly one function of affect—to make norm deviations salient); but such affect neither is a moral judgment nor does it by itself generate a moral judgment.

People are highly sensitive to norm deviations. Such negative events trigger rapid evaluative responses (Ito, Larsen, Smith, & Cacioppo, 1998; Van Berkum, Holleman, Nieuwland, Otten, & Murre, 2009). Furthermore, a host of work on "negativity effects" shows that, compared to positive or neutral events, negative events command more attentional resources, are more widely represented in language, and exert a stronger impact on both interpersonal- and self-perception (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Taylor, 1991). This responsiveness is not unique to moral violations. People experience a nagging *why* question for all kinds of puzzling events, and particularly for negative ones (Malle & Knobe, 1997a; Roese, 1997; Wong & Weiner, 1981). Blaming grows in part out of an activity of assigning meaning to an event. Finding meaning resolves tension of uncertainty, a gap in understanding, restores coherence and control. For negative events, explaining the origin is important, but in principle it could be done without blaming. So blaming must have an additional function beyond establishing meaning. It seeks meaning of a particular

kind—the involvement of agents, because agents can be influenced and reformed, which is critical for upholding community norms.

Agent Causality

Social perceivers blame people, not physics. So for mere outcomes (e.g., a broken window, a dead person) to lead to blame, the perceiver must establish whether an agent caused the outcome (Shaver, 1985; Sloman, Fernbach, & Ewing, 2009). By contrast, observing norm-violating *behaviors* already comes with the recognition of agent causality, as behaviors are, by definition, caused by agents. The same is true for "nonbehaviors" such as omissions or intentions; letting someone die or planning to hurt someone may not be physical movements, but they can be described, interpreted, and morally criticized. What counts as a norm-violating event or behavior may rely in part on the operation of moral "intuitions" and "moral grammar rules." These intuitions may activate the moral judgment machinery by flagging the types of events that are potentially worthy of moral judgment. For example, Graham, Haidt, and Nosek (2009) suggest that moral judgments arise in response to five broad domains of norm deviations: those concerning harm, fairness, ingroup, authority, and purity. Thus, people may have intuitions that harmful behavior is "bad," particularly when it stems from physical contact (Cushman, Young, & Hauser, 2006; Greene et al., 2009; Mikhail, 2007). However, very little is achieved at this point in the way of moral judgment. The targets of moral judgments are not merely behaviors and outcomes; they are people. During the detection phase, social perceivers evaluate the deviance ("badness") of some event or outcome; during the agent causality phase, they shift toward a moral analysis. Because only agents (with certain capacities) can be blamed for what they cause, do, or don't do, the identification of an agent causing the negative outcome opens the door to genuine blame judgments.

What do people categorize as an agent? The agency concept growing out of 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 are that make an agent morally eligible have spawned a rich literature in philosophy. From a folk-conceptual perspective, this issue deserves a more detailed analysis elsewhere. However, it seems clear that one necessary feature for moral eligibility is the agent's ability to understand and remember norms. A second central feature, we have proposed, is the capacity for choice, with its implied ability to reason from beliefs and desires to intentions (Guglielmo et al., 2009;

Monroe & Malle, 2009). A clarification of agent eligibility is also relevant to an account of blame mitigation because failure to be eligible as a moral agent (e.g., because of a mental illness that dismantles the choice capacity) will lead to the most decisive level of mitigation.

Once an agent has been detected, the perceiver will assess the causal involvement of this agent in the norm deviation. Numerous studies demonstrate the crucial impact of causal involvement in assigning blame (e.g., Cushman, 2008; Lagnado & Channon, 2008), for social perceivers from age 5 on (Shultz, Wright, & Schleifer, 1986). But causal involvement falls into two fundamentally different categories—intentional and unintentional (Heider, 1958; Malle, 2004).

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 behavior stream into units that correspond to intentional actions (Baldwin, Baird, Saylor, & Clark, 2001). By the second year, 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 6 or 7 years, 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, 1997b).

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, 1997b, 2001), we should not expect people to deliberate about these five components each time they judge a behavior as intentional. Instead, we can expect people to use a more efficient path to assess intentionality in everyday situations but to consider carefully each of the components if uncertainty or the weight of the judgment demands it.

Intentionality judgments regulate attention in social interaction (Malle & Pearce, 2001). 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. Intentionality judgments also

guide explanations and predictions of behavior (Malle, 2004). Most important, to account for intentional and unintentional behavior people use very different modes of explanation, which differ in conceptual, cognitive, and linguistics properties (Malle, 2004, forthcoming).

Of primary interest here is the role that intentionality judgments play in moral judgment. Children begin to incorporate intentionality into their moral judgment at least as early as age 5 (Shultz et al., 1986; Surber, 1977). 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). The intentionality distinction in moral judgment comes for free cognitively (Solan, 2003) because an understanding of intentional action is already available before children have an understanding of moral norms and moral judgments. As they acquire (and have to obey) these kinds of norms, adults' differentiation between intentional and unintentional wrongdoing can increasingly rely on scripts and schemata. Sometimes, however, there will be uncertainty about intentionality, and the perceiver either searches for further information or takes a guess, which may be vulnerable to motivational influences (Alicke, 2000).

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 (Darley & Shultz, 1990; Lagnado & Channon, 2008; Ohtsubo, 2007; Young & Saxe, 2009; see also Dahourou & Mullet, 1999, for a replication with a non-Western sample). But exactly what does intentionality do in the moral judgment domain? According to our model, intentionality bifurcates the perceiver's processing of norm violations. People search for and process rather different information when encountering intentional as opposed to unintentional wrongs (the paths 4a and 4b in Figure 1). But before we describe these paths in more detail, we would like to travel on a brief excursion to address a simple but challenging question: why is intentional behavior blamed more than unintentional behavior?

A social community will have success to the extent that it keeps its members cooperating and staying within (implicit or explicit) norms (e.g., Sunstein, 1996; Wilson, 2002). Permitting too many free riders and deviants lowers group coherence and the possibility of collective action. But members of the community must recognize the limits of an individual's control over reality. Imagine, just for illustration, that Homo erectus had a distinction between negative outcomes due to a person's controlled behavior and negative outcomes due to a person's uncontrolled behavior.

The first is predictive of more of the same behavior in the future; the second is not. To understand that, these early humans need not have a sophisticated theory about the kinds of mental states that underlie this distinction (such as desire, reasoning, intentions); all they need to recognize is that an uncontrolled negative event is an *exception* to normal control paths. But exceptions are unlikely to repeat, so the community can worry less about such uncontrolled negative outcomes.

Compare to that a community of hominids who only consider outcome severity—in the spirit of pure consequentialism. These creatures cannot distinguish between a one-time deviation and systematic deviations; they will punish indiscriminately. But if uncontrolled harm is an *exception*, it won't happen again, in all likelihood, and there is less need for punishment as the community's behavior regulation. Some amount of punishment is still functional—as a warning to that person (and others) that they need to make every effort to not allow those exceptions (Hart, 1968).

This logic also helps clarify the fact that people typically don't praise unintentional positive behaviors (Shultz & Wright, 1985). They are accidents, exceptions and therefore not predictive of future repetition. Rewarding them with praise is likely to have no beneficial effects on the community. Intentional positive behaviors should be encouraged through praise, however, and not just consequentially as in operant conditioning, but by encouraging the generalizable sentiments of wanting to please others, being generous, caring, and helpful. Beyond reinforcing specific behaviors, praise builds an attitude, a value that has a chance for a broader impact.

We now follow the two tracks of blame formation that succeed the perceiver's intentionality judgment.

Reasons and Justification

In the case of intentional behavior (the left path in Fig. 4), the perceiver considers the agent's particular reasons for acting. This again comes for free because it is something people do with ease, and they actually find it painful *not* to know the reasons of someone's action (Malle, 2004). As people consider these reasons, some of them may actually provide justification for the norm-violating action at hand. An agent who hurt someone intentionally will be blamed less or not at all if he had justified reasons (e.g., a schoolboy defending his sister against bullies) than if he had unjustified reasons (e.g., a schoolboy trying to provoke a fight).

What reasons are justified is of course a manner of community or legal norms (Alexander, 2009, chap. 4).

The well-known trolley scenario serves as an example of the power of justification. A train station attendant notices a runaway train barreling down the tracks towards five unsuspecting workmen. He also notices a large man on the bridge and pushes him off, thereby using his body to stop the train. When asked why he did that, he answered that if he had done nothing, the train would have run along the track and killed five workers. Most people in most cultures do *not*, as it turns out, consider this an acceptable justification for pushing the large man off the bridge and killing him. Now consider a similar case in which the train is carrying a dangerous virus (Nichols & Mallon, 2006). The attendant, watching the train from a footbridge, notices a bomb on the tracks ahead. If the train passes over the bomb, it will explode and the virus will be released killing billions. He also notices a large man on the bridge and pushes him off, using his body to stop the train. Again, when asked why he pushed the man, the attendant might reply that if he had done nothing billions of people would have been killed. Most people consider this an acceptable justification for an otherwise serious violation: causing a person (the large man) to die. Numerous analyses of this case have been offered; all we want to point to here is the eminent role that the person's reasons for action play. The second attendant has a justified reason for sacrificing one person's life in exchange for saving billions; but he has no justification for shoving a person to his certain death in exchange for saving only five lives.

Reasons for action come in two main forms: desires and beliefs. Preliminary research in our lab indicates that beliefs may ultimately be stronger in justifying norm-violating actions and thereby reducing blame (Malle & S. E. Nelson, 2006). An example from the legal literature highlights the centrality of belief information in justifications. In *People v. Goetz*, Goetz was charged with attempted murder and first-degree assault for shooting four teenagers on the New York subway. Goetz reported that while riding the subway, he was approached by four black teenagers, one of whom asked Goetz for money. Having recently been a victim of a mugging, he feared the four teenagers were planning on beating and robbing him. Consequently, he drew a concealed pistol and fired five shots at the youths, injuring all four. At the trial, Goetz related his previous mugging experience and his belief that the teenagers intended to harm him. He was acquitted of both attempted murder and first-degree assault, though he was convicted of criminal

possession of a weapon. Apparently the jury agreed that Goetz's decision to shoot the four teenagers was justified by his belief that he was in danger.

Interestingly, the features of reasons that make them justified have not been investigated. We can suspect that people will be sensitive to at least the following features: for desires, how desirable the agent's goal was, how consensual the desirability is, and whether the agent had alternative means to fulfill the goal; for beliefs, how widespread the belief is in the community; how strong the evidence was at the time for the agent to hold the belief; and how self-serving it was to hold the belief.

Unintentional Events

Our model is designed to cover a broad range of unintentional events—unintended side effect, failed attempts at intentional actions, and involuntary behaviors—including those that are based on a defective choice mechanism (Guglielmo et al., 2009). Note, however, that actions under duress (e.g., committing a crime under threat to one's life) do not fall under the unintentional category because the agent acted intentionally—albeit under severely constrained options and therefore with justification.

People's explanations of unintentional behavior are far simpler than those for intentional behavior. The latter involves three possible modes of explanation that can singularly or jointly be recruited to explain the behavior. One requires consideration of the agent's subjective reasoning, another focuses on the causal background of those reasons (e.g., personality, culture, and context), and a third considers objective enabling factors that allowed the agent to successfully complete the intended action. Unintentional behaviors are explained by one mode—causes, which more or less mechanically bring about the behavior without any involvement of reasoning, intentions, or enabling conditions.

By contrast, when people analyze unintentional behaviors for possible blame, the considerations are quite complex. They go beyond the backward-looking explanations of the behavior's causes and reach into the forward-looking considerations of potential repeats of the event and prospects of their prevention. Instances of blame are often backward looking (retributive), especially when we consider blame as cognitive process; but the overall function of blame, and especially its social expression, is also forward looking (reformative) because it is one of the community's tools to regulate behavior.

When people regard the agent as having unintentionally brought about a negative outcome, they examine whether the agent *should have* prevented the outcome (obligation) and *could have* prevented it (capacity). Both of these considerations are grounded in the intentionality concept. Social communities impose obligations on one another because they know they can *intentionally* act (or at least intend to act) in accordance with these obligations and can therefore intentionally prevent negative outcomes. If an agent lacks the requisite abilities (such as reasoning, planning, and choice) to grasp obligations and to intentionally meet them, little to no blame is (or should be) doled out (Bratman, 1997). By contrast, if the agent has the necessary abilities and is subject to the obligation, then a failure to prevent the negative outcome will trigger substantial blame. A one-year old can neither understand what obligations are nor act in accordance with them; a five-year old can. The former will not be blamed, the latter will be (and the latter will probably also feel guilt as a form of self-blame).

Note that an intentional refusal to meet an obligation to prevent constitutes a norm violation that will be evaluated along the left (the intentional) path of the step model. The perceiver now considers the agent's justified or unjustified reasons for the violation. Perceivers may thus initially form a moral judgment about an agent's unintentional causing of a negative outcome and end with a moral judgment about the agent's intentional failure to prevent the outcome. This is one feature that made Knobe's (2003) chairman of the board story so interesting:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment." The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was harmed.

Knowing that the environment would be harmed and deciding not to take any steps of preventing such harm elicited strong blame assignments (M = 4.8 on a 0-6 scale). And that held even though most people did not think that the chairman *intentionally harmed* the environment.3

Evidence for the impact of obligation. Most studies of moral judgment hold obligation constant—they typically contain stories in which the agents clearly do have an obligation to prevent negative outcomes. As a result, there is little direct evidence for obligation's impact on blame. When it has been examined, obligation has shown considerable influence. Hamilton

(1986) reported that people with higher positions in a social hierarchy are subject to stronger obligations for preventing negative outcomes and are held more strongly responsible for those outcomes when they do occur. Similarly, Shultz, Jaggi, and Schleifer (1987) showed that even vicarious responsibility is greater for those further up in a social hierarchy.4 Haidt and Baron (1996) reported that people judged it morally worse for an agent to withhold information from a friend than from a neighbor, which itself was worse than withholding from a stranger. Thus friendship intimacy increases obligation to prevent negative outcomes. In this case, the wrongdoing was an omission, so it could be argued that we are on the left (the intentional) path of the step model because the accused intentionally withheld information. But if we consider the event to be "target person acts on false information," then we can say that the accused did not intentionally bring that about; it was an unintentional side-effect of his decision to not reveal information.

Evidence for the impact of capacity. Shultz et al. (1987) also 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 the capacity to prevent the harmdoing. (Weiner (1995) reviews numerous studies in which the agent's ability to control an outcome is a strong predictor of blame. For Weiner, controllability attaches to the *causes of the behavior or outcome.* For example, if a person's obesity is caused by a medical condition, the person cannot control that condition and people don't consider the person responsible for the obesity and therefore not blameworthy. If another person's obesity is caused by indulgent overeating, the person is (assumed to be) able to control that behavior and people consider the person responsible, hence blameworthy. However, the critical question is whether the person can *prevent* the outcome from occurring. Even if the medication condition is the cause of the obesity and cannot be controlled, the person might undergo surgery to shrink the stomach and, even in the presence of the medical condition, prevent obesity. General properties of controllability that attach to causes are less important than the options the agent has *and has not yet* explored to prevent the negative outcome.

Preventability of unintentional harm not only influences cognitive blame but also has direct social-emotional consequences. Jones and Kelly (2010) showed that people lose trust in harmdoers and like them less if they were able to prevent the harm, and Weiner showed

numerous times that people have less sympathy with those who have control over negative outcomes.

Currently, there is no evidence that clarifies the order in which obligation and capacity are assessed. Our model suggests that obligation is checked earlier, and that hypothesis rests on differential processing efficiencies. First, easier judgments are likely to be made earlier. There is far less information to consider when checking obligation than when checking capacity, and it would be quite inefficient to first go through potentially difficult assessments of whether the agent could have prevented the negative outcome and then realize that the agent had no obligation to prevent it. Some obligations may even be accessible instantly at the time of event detection, namely, when the violated norm prohibiting X and the obligation to prevent occurrences of X are part of the same knowledge structure. Second, obligation is a more context-general expectation, based on roles, relationships, and the nature of the outcome—all category-based information. By contrast, capacity is a context-specific assessment, based on the details of the event at hand. Processing such details normally takes more time than the activation of category-based information.

Competing Models

The literature on blame and moral judgment has featured at least 10 models of the antecedents, psychological processes, and consequences of such judgments (see Guglielmo, under review, for a detailed discussion). These models can, however, be divided into two main groups, and we focus on the central disagreement between these groups—whether blame judgments *follow* mental state judgments (which we label blame-late models) or *precede* mental state judgments (which we label blame-early models).

Blame-Late Models

Blame late models propose that judgments of blame critically rely on prior assessments about an agent's mental states (among other things). For example, "entailment models" posit that certain early judgments serve as necessary and sufficient conditions for subsequent judgments, the last of which is that of blame or punishment (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, or had the control to produce a different outcome. In the absence of these assessments, according to blame late models, it makes no sense to ask about the blameworthiness of an agent's behavior.

Our step model of blame of course has clear affinities with these entailment models but 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. In both ordinary language and in the psychological literature, the concept of responsibility is typically defined either as identical to causality (Harvey & Rule, 1978), as identical to obligation (Hamilton, 1986) or as identical to blame (Shultz, Schleifer, & Altman, 1981). Models that include the responsibility concept also often pack a number of other important distinctions into this components that our model treats separately: intentionality (which is a primary input to blame), coercion (a possible justification for an intentional action), and capacity (which moderates the extent to which unintentional behaviors incur blame).

The key question is whether blame judgments in fact follow assessments of these features. A host of evidence suggests that they do. Blame judgments respond to variations in the agent's causal role in a negative event (Cushman, 2008; Sloman et al., 2008), and the extent to which the agent was coerced in acting (Woolfolk, Doris, & Darley, 2006) or had control over an unintentional behavior (Weiner, 1995). Most importantly, blame differs markedly as a function of the agent's mental states. Intentional negative actions receive greater blame than unintentional ones (Cushman, 2008; Lagnado & Channon, 2008; Ohtsubo, 2007; Young & Saxe, 2009), and even children's moral judgments differentiate between desired and undesired harm, and between purely unintentional and foreseen yet unintended actions (Darley & Shultz, 1990; Nelson-LeGall, 1985).

Blame-Early Models

A second class of 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 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 "moral judgment is caused by quick moral intuitions" (p. 817).

Alicke (2000) offers a more explicit claim about the impact of blame on mental state judgments. "People use outcome information as a basis for ascribing blame and that 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 his model, judgments about causality and mental states still guide blame judgments, as they do for the "blame late" models discussed above. However, an "initial moral judgment" (Pettit & 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 affective reactions when considering negative behaviors and that they are often "dumbfounded" when attempting to verbally justify their moral evaluations (Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993). 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 (Alicke, 1992; Mazzocco, Alicke, & Davis, 2004). Finally, studies by Knobe and others suggest that, compared to 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 finding support a blame-early model. Haidt's studies never vary information about causality, intentionality, or mental states, thus making it impossible to examine whether moral judgments precede and influence those nonmoral judgments. Moreover, Haidt typically measures detections of norm violation ("this is

wrong"), not actual judgments of blame; and the relation between intuitions about something being wrong and full-blown moral judgments of the person remains unclear.

The results of Alicke's studies may arise from an informational impact of negativity, rather than a motivational one, because we know that negative characteristics provide better diagnostic evidence of a person's underlying dispositions and motives than do positive characteristics (Reeder & Brewer, 1979; Skowronski & Carlston, 1989). Finally, follow-up research on Knobe's intriguing findings has shown identical patterns for behaviors that lack moral content (Machery, 2008; Uttich & Lombrozo, 2010), suggesting that Knobe's findings can be explained by norm violation more generally, not moral violation in particular. Most importantly, however, tests of blame-early models have not included measures of the critical early moral judgments that are said to guide mental state judgments, such as Alicke's spontaneous evaluations or Knobe's initial moral judgments. Consequently, the key claim of these models is, at present, not well supported.

Resolving the Blame-Early vs. Blame-Late Debate

Intentionality and moral judgment. One important issue that has been featured in the debate is clarifying the connection between intentionality judgments and blame judgments. Knobe's (2003) findings showed that people view certain negative actions as more intentional than similarly structured positive or neutral ones, suggesting that people's moral assessments precede their mental assessments. Our recent work, however, challenges this interpretation (Guglielmo & Malle, 2010a, 2010b; Malle & Guglielmo, 2011). We have shown that aside from varying the moral valence of the actions in question, Knobe's scenarios also varied other critical information, such as the agent's desire or skill. Once this information was manipulated or properly controlled, there were no longer any differences in people's intentionality judgments as a function of valence (Guglielmo & Malle, 2010a, 2010b).

Moreover, even when considering Knobe's original chairman story, hardly anyone characterized the outcome as intentional once they were allowed to more freely express their conceptualization of the story (Guglielmo & Malle, 2010a). In that case, a strong majority of people indicated that the chairman *knowingly* brought about the outcome, a pattern that was true regardless of whether the outcome was negative (environmental harm) or positive (environmental benefit). People continued to give the (harming) chairman substantial blame,

because of course he had an obligation to prevent the environmental harm and possessed the relevant capacity to do so, as our model predicts.

The Role of Timing. The fundamental difference between the competing models is that they make distinct claims about the sequence of people's blame judgments and mental state judgments. Thus, directly examining the timing of different judgments offers a promising approach for adjudicating between the models. For example, for the sequence of blame and intentionality, blame-late models, including our step model, predict that people will be slower to assess blame than to assess intentionality. In contrast, blame-early models predict the opposite—namely, that people will be faster to assess blame than intentionality. We are currently conducting studies to test these competing hypotheses.

An Integrated View. Is it possible to integrate the distinct claims and findings of the two classes of blame models? An integrated perspective is indeed possible, and it relies on a distinction between early affective responses and later genuine judgments (Guglielmo, under review). People surely experience negative affect upon detecting negative events—death, environmental damage, and so on—but this affect turns into a moral judgment (e.g., of blame) only after people interpret this affect through a lens that analyzes the causal and mental-state structure of the event. This conceptual framework—one that provides causal and mental analysis of the event at hand—gives meaning to one's early affective response and thereby transforms evaluations of outcomes into moral judgments of a person.

Applying the Model I: Blaming Groups

There is broad agreement in the literature that a group's capacity for intentional action is a prerequisite for the group's status as a moral agent. As Isaacs put it, "showing that collectives are capable of intentional action is necessary for showing that they are appropriate objects of praise and blame" (Isaacs, 2006, p. 62). The set of features of intentionality, mental states, intentions, and reason-based choice (rationality) is also what French postulates as central in rendering a corporation a "moral agent" (French, 1979, 1996) He argues that corporations are moral agents *because* they are capable of intentional action. These are claims about the metaphysics of corporations; however, they are in accordance with ordinary social perception. If, as we have seen, the ability to act intentionally and have reasons for acting is critical for moral agency, then groups will similarly be seen as moral agents to the extent that they possess these qualities.

But the status of corporations and other groups as intentional agents makes them only *eligible* for moral evaluation. How does such evaluation work in detail? Is it the same as that for individuals? We need not automatically assume that collective moral judgment operates the same way, but if there is no evidence to the contrary, we may continue to accept it as a working hypothesis. A basic theoretical argument also strengthens this equal-operation hypothesis. If people's powerful folk psychology is unflappably applied to group agents, and if moral judgment deeply draws upon folk psychology, then moral judgment, too, should be applied to group agents (Malle, 2011b).

To test the equal-operation hypothesis, we use our step model of blame to examine whether judgments of group action could be shuttled through a cognitive apparatus akin to that used for individual agents.

A brief look into any newspaper reveals that people easily and often detect norm-violating group behaviors—performed, for example, by teams, gangs, corporations, political parties, governments, or nations. Norms for groups may differ from those for individuals, but for the norms that do apply to group agents, moral breaches are certainly recognized.

People also have no trouble distinguishing between intentional and unintentional group behavior. Unintentional collective behaviors may be less frequent than intentional ones (O'Laughlin & Malle, 2002), but acts of negligence (by definition unintentional) are commonplace in accusations of objectionable corporate behavior.

Following the left path in the step model of blame, we know that people ascribe reasons to group agents (O'Laughlin & Malle, 2002), so we can expect people to consider reasons as possible blame moderators for norm-violating actions. A corporation or government will certainly offer such justifying reasons for its own actions in order to mitigate potential blame.

Following the right path of arriving at blame, the presence of norms for group agents implies that there are obligations in place as well, for being subject to a norm means being obligated to conform to it, and if there is a norm of prevention (especially of harm), the obligation to prevent will figure prominently in people's judgments.

Furthermore, groups arguably vary in their capacities to prevent possible negative outcomes. It should be uncontroversial that they can vary in their knowledge of certain facts and also that they can have skills, resources, and opportunities to either bring about or prevent outcomes.

Thus we arrive, without making contentious assumptions, at a picture according to which group agents can be blamed through operation of the same cognitive apparatus through which individuals are blamed. We have no direct evidence that the formation of group blame follows only these steps, but there are at least no apparent obstacles for the social perceiver to do so.

Applying the Model II: Blaming as a Social Act

Problems for Purely Cognitive Models

Extant models of blame focus almost exclusively on the intrapersonal processes of arriving at blame judgments. But there is no doubt that people do more than blame others in their own heads. Blame is expressed in face, body, and language; it is doled out, countered, negotiated. A comprehensive theory of blame must be able to delineate the antecedents and consequences of such social acts of blaming.

In Haidt's (2001) "social intuitionist" model, 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 do not have any access to the emergence of their moral judgments (they are "dumbfounded" by their intuitions), so there is really nothing to say during the social expression of blaming except "This is wrong, he is bad." If people cannot consciously retrieve any grounds for their judgments, how should they be able to argue about, negotiate, and justify moral judgments?

Steps Toward Social Blame

Our model of blame offers some answers. We assume that people 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 intention, reasons, knowledge, etc.). People may not know how all these information components "fit together" to produce a blame judgment, but the information itself is available to them 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 "proven" for a verdict to ensue.

But social acts of blaming aren't only addressed to other observers. They are also addressed to the perpetrator, especially by the victim of the transgression. In Duff's (1990) 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 you cause it? Did I 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 other observers as well as to the perpetrator.

A full theory of social blaming must address two further questions: Under what circumstances do people express their blame, and what consequences does blaming have? Here, the two audiences—the perpetrator and other observers—are likely to lead to different answers.

When to blame. In the presence of the perpetrator, the threshold to express blame may generally be higher because expressing it (a) may be prohibited by social norms of role, status, or relationship; (b) may lead to a hostile response from the perpetrator; (c) or may endanger the relationship with the perpetrator.

As part of its higher threshold, social blaming must also obey (more strongly than private sentiments) the fundamental condition of "responsible agency": somebody who cannot *respond* to blame also should not be blamed. Fischer and Ravizza (1998) call this requirement "reasons-responsiveness." A 2-year old cannot properly respond to arguments (reasons) and criticism (blame) by correcting her actions. The limitations may be partially cognitive (understanding the binding nature of norms and other people's demands) but also lie in self-control. Either way, we do not blame the 2-year old the same way as we would blame the 5-year old for the same behavior.5

Some people, such as psychopaths, may not be responsive to reasons and interpersonal criticism even though they probably have all the cognitive and agentic capacities. Or do they? There is debate on this issue, but whatever the outcome of that debate, we don't have to require that blame be *successful*. People sometimes are not responsive to criticism, arguments, demands, blame—but we nonetheless know they have the capacities from other cases in which they are responsive. Perhaps the same can be said for the psychopath, who may be reasons-responsive in

cases of self-serving outcomes but not in the cases of other-serving outcomes. The problem may be motivational, not cognitive.

Functions of blaming. The function of blaming is likely to differ as well by audience. Directly blaming the perpetrator normally offers a better chance of actually reforming the person's behavior, especially if there is a preexisting relationship between blamer and perpetrator. In the ideal case, the blamer not only condemns the other person's behavior but appeals to the person's values, to community standards, in an attempt to make the person recognize the wrongness of his actions and encourage different behavior in the future. Such an act respects the other person's rationality, responsiveness, ability to understand and change, but it also reflects a willingness on the part of the blamer to listen to the person's own perspective and consider possible justifications for the behavior. In less ideal cases, people blame irrationally, unfairly, and without respect or argument, which may be a sign of defective relationships (Bradbury & Fincham, 1990; Fincham, Beach, & G. Nelson, 1987).

Third-person blaming—addressed to other observers—has no chance of reforming the perpetrator; instead, it serves primarily to express the blamer's values and to seek social validation for those values (Duff, 1990; Pearce, 2003). Third-person blaming can arise out of an inability to reform (e.g., because the perpetrator is too high in status to be directly addressed) or out of the blamer's refusal to even attempt any reform. In the latter case, the act of blaming may represent the first step toward social exclusion of 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 seriously negative events. Cultural studies have recently documented this process in various African nations' grappling with the HIV epidemic (Rödlach, 2006; Stadler, 2003; Treichler, 1999). One of the most cruel examples, however, 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. In this and similar cases, the propaganda very much claimed 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.

At other times, third-person blaming is indeed irrational and affectively driven, dispensing of all demands on argument, response, or reform and instead offering community-sanctioned opportunities to express hatred, as in the shocking practice of lynching (Dray, 2002). Whether such acts of hate should count as "blame" is unclear, but people consider such acts as deeply unjust precisely because they refuse to grant the accused any response and wholly ignore the foundational questions of blame: Was the agent causally involved? Did he act intentionally? Could he have prevented the outcome?

Is the "Blame Game" a Bad Thing?

Some time in the 20th century, the expression "(playing the) blame game" emerged (according to the OED, in 1958). At its core it describes the activity of assigning blame, finding fault after a negative event has been discovered. But it clearly implies something undesirable: "a phrase insinuating our well-established agreement that the game itself is blameworthy" (Robbins, 2007, p. 140). It often involves multiple people blaming each other—"pointing fingers" at multiple candidate targets. But what is bad about it? In most cases, blamers still provide arguments on the basis of causal analysis, propose hypotheses of intent and knowledge, and explicitly ascribe obligations and capacities to prevent. What makes players of the blame game undesirable is that they consistently accuse others of wrongdoing while deflecting or denying their own wrongdoing. Detached observers, who condemn the players, want one or more of those involved to "take responsibility." Neither the detached observers, however, nor the players of the blame game operate without reflection, willy-nilly picking targets of blame. They *argue* for their accusations and defenses (the players most likely in self-serving and distorting ways), and once more, we expect, the standard components or steps of blame serve as their guideposts.

Social Blaming of Groups

Earlier we argued that people direct blame at group agents and do so with essentially the same psychological apparatus that they apply to individual agents. This was an argument about the cognitive side of blame. Interesting problems arise, however, 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 rare, 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; Malle, 2009), 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.

Summary

We examined the role of fundamental social-cognitive processes in blame judgments and proposed a model of blame that characterizes blame as both a cognitive phenomenon and a social phenomenon. We put our model in the context of two large classes of blame models—those that postulate blame to be an earlier process and others that postulate blame to be a later process. We provided theoretical reasons and reported empirical findings that favor the blame-late models. No doubt, people often have an early affective evaluation when they detect a "bad outcome," but blame as an actual judgment about the agent typically requires a number of additional conceptual and cognitive steps. Appling this model of blame, we explored to what degree blame for group agents resembles blame for individual agents and sketched what we know and need to learn about blame as a social act.

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). However, we will use these terms interchangeably because for our present purposes this distinction is not important.

3 Knobe's original claim was that people do think the chairman intentionally harmed the environment. However, in a series of studies we have demonstrated that this claim is false (Guglielmo & Malle, 2010a). See later section entitled "Resolving the Blame-Early..."

4 Vicarious blame (owners are blamed for damage caused by their pets; parents, for damage caused by their children) is rare, and it operates only when obligation and capacity are established. In one sense these cases violate the causality requirement; but there is a concept of "allowing causes" in the philosophical literature on causation, and people may have something like this in mind when they consider, say, the pet owner blameworthy because she should have and could have controlled her pet. Within counterfactual theories of causation, this is not a surprising claim: if only the owner had not taken his eyes of his dog, it wouldn't have bitten the child.

5 There are limitations to this claim. First, some people fail to recognize the limitations of a 2-year old and blame, even punish the child as if she were much older. But these people may fail to recognize the capacity limitation; they presuppose the capacity, which still supports the claim that the capacity assumption is significant.

References

Alexander, L. (2009). *Crime and culpability: A theory of criminal law*. New York: Cambridge University Press.

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 US: The 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.

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.

Bradbury, T. N., & Fincham, F. D. (1990). Attributions in marriage: Review and critique. *Psychological Bulletin*, *107*, 3-33.

Bratman, M. E. (1997). Responsibility and planning. *The Journal of Ethics*, *1*, 27-43.

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

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science*, *17*, 1082 -1089.

Dahourou, D., & Mullet, E. (1999). The relationships among intent, consequences, and blame in Burkina Faso adolescents and young adults. *IFE Psychologia: An International Journal*, *7*, 32-45.

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.

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

Dray, P. (2002). *At the hands of persons unknown: The lynching of black America* (1st ed.). New York: Random House.

Duff, R. A. (1990). *Intention, agency and criminal liability*. Oxford: Basil Blackwell.

Felstiner, W. L., Abel, R. L., & Sarat, A. (1980). The emergence and transformation of disputes: Naming, blaming, claiming. *Law & Society Review*, *15*, 631-654.

Fincham, F. D., Beach, S. R., & Nelson, G. (1987). Attribution processes in distressed and nondistressed couples: III. Causal and responsibility attributions for spouse behavior. *Cognitive Therapy and Research*, *11*, 71-86.

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.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge studies in philosophy and law. Cambridge: Cambridge University Press.

French, P. A. (1979). The corporation as a moral person. *American Philosophical Quarterly*, *16*, 207-215.

French, P. A. (1996). Integrity, intentions, and corporations. *American Business Law Journal*, *34*, 141.

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.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364-371.

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

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

Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, *26*, 201-218.

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

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*, 613-628.

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

Hart, H. L. A. (1968). Intention and punishment. In *Punishment and responsibility* (pp. 113 - 135). Oxford, UK: Clarendon Press.

Harvey, M. D., & Rule, B. G. (1978). Moral evaluations and judgments of responsibility. *Personality and Social Psychology Bulletin*, *4*, 583-588.

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

Isaacs, T. (2006). Collective moral responsibility and collective intention. *Midwest Studies In Philosophy*, *30*, 59-73.

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, *75*, 887-900.

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

Jones, E. E., & Kelly, J. R. (2010). "Why am I out of the loop?" Attributions influence responses to information exclusion. *Personality and Social Psychology Bulletin*, *36*, 1186-1201.

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. In *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. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, *7*, 67-83.

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.

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

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

Malle, B. F. (2009). Folk theories of consciousness. In W. P. Banks (Ed.), *Encyclopedia of consciousness* (Vol. 1, pp. 251-263). Oxford, England: Elsevier/Academic Press.

Malle, B. F. (2011a). Time to give up the dogmas of attribution: A new theory of behavior explanation. *Advances of Experimental Social Psychology*, *44*.

Malle, B. F. (2011b). The social and moral cognition of group agents. *Journal of Law and Policy*.

Malle, B. F., & Guglielmo, S. (2011). Are intentionality judgments fundamentally moral? In C. Mackenzie & R. Langdon (Eds.), *Emotions and moral reasoning*, Macquarie Monographs in Cognitive Science. Psychology Press.

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

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

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

Malle, B. F., & Nelson, S. E. (2006). How bad is it? The role of explanations and intentionality in evaluations of
objectionable behavior. Presented at the Society of Personality and Social Psychology Annual Conference, Palm Spring, California.

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.

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.

Monroe, A. E., & Malle, B. F. (2009). 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*.

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

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*, 530-542.

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 effect1. *Japanese Psychological Research*, *49*, 100-110.

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

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

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.

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

Robbins, B. (2007). Comparative national blaming: W. G. Sebald on the bombing of Germany. In A. Sarat & N. Hussain (Eds.), *Forgiveness, mercy, and clemency*. Stanford, CA: Stanford University Press.

Rödlach, A. (2006). *Witches, Westerners, and HIV: AIDS & cultures of blame in Africa*. Walnut Creek, CA: Left Coast Press.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*, 133-148.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296-320.

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

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. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *17*, 97-108.

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*, Psychology of Learning and Motivation (1st ed., pp. 1-26). Academic Press.

Solan, L. M. (2003). Cognitive Foundations of the Impulse to Blame. *Brooklyn Law Review*, *68*,

1003-1029.
Stadler, J. (2003). Rumor, gossip and blame: implications for HIV/AIDS prevention in the South African lowveld. *AIDS Education and Prevention*, *15*, 357-368.
Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, *96*, 903-968.
Surber, C. F. (1977). Development processes in social inference: Averaging of intentions and consequences in moral judgment. *Developmental Psychology*, *13*, 654-665.
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: 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., 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, NY: 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.
Wilson, D. S. (2002). *Darwin's cathedral: Evolution, religion, and the nature of society*. Chicago: University of Chicago Press.
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.
Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*, 2065-2072.