Are Intentionality Judgments Fundamentally Moral?

Bertram F. Malle Steve Guglielmo
Brown University

People's capacity to recognize a behavior as intentional is a central component of human social cognition. This capacity has evolved for its adaptive value in social interaction, and it develops rapidly in the early years of life (Malle, Moses, & Baldwin, 2001; Zelazo, Astington, & Olson, 1999). Furthermore, the intentionality concept is part of folk psychology, the larger conceptual and cognitive system that allows people to make sense of human behavior in terms of mental states. In this system, intentionality plays a pivotal role because it directly connects behavior with mind, classifying actions as intentional when they are caused by certain characteristic mental states such as belief, intention, and awareness.

Much of the research on intentionality judgments has focused on their conceptual structure (Kashima, McKintyre, & Clifford, 1998; Malle & Knobe, 1997), their underlying cognitive and neural processes (Baldwin & Baird, 2001; Davis, 2005; Saxe Xiao, Kovacs, Perrett, & Kanwisher, 2004), and their essential role in behavior explanations (Malle, 2004). Recently, however, a flurry of research and vigorous debates have drawn attention to the relationship between judgments of intentionality and moral judgments, especially blame.

Intentionality and Moral Judgment

We can distinguish two opposing models of the relationship between intentionality and morality. The first assumes that intentionality judgments are one of the important inputs to perceptions of moral valence. According to this model, a social perceiver first assesses an observed behavior's intentionality (a cognitive judgment) and then, in light of this judgment, assigns blame or praise to the agent of that behavior. Schematically, the model claims *intentionality* → *blame*/*praise*. This model has dominated the literature (Darley & Shultz, 1990; Fincham & Jaspars, 1980; Ohtsubo, 2007; Shaver, 1985; Weiner, 1995) and has received considerable empirical support. We call it the standard model.

The second model assumes that social perceivers have immediate moral intuitions in response to a negative outcome or behavior and that these intuitions influence and direct subsequent judgments of the behavior's intentionality. As a general account of how moral judgments arise from intuitions and emotions, such models have been promoted for some time (Alicke, 2000; Greene et al., 2001; Haidt, 2001; for a review, see Haidt & Kesebir, 2010). Recently, however, a specific account of the purported influence of moral valence on intentionality judgments in particular has emerged (Knobe, 2003a, b). In contrast to the standard model, this model reverses the schematic relationship between intentionality and moral perception: *blame*/*praise* → *intentionality*.[1] We call it the challenger model.

In this paper we will briefly review the evidence for each model and then introduce our own studies that have pitted the two models against each other.

---

[1] More recently, Pettit and Knobe (2009) argued that people's intentionality judgments are influenced by their assessments of badness/goodness rather than by their blame/praise. [1] Regardless of which particular moral sentiment is implicated by Knobe's model, our discussion is relevant to the general claim that morality influences judgments of intentionality.

To set up the comparison between the two models, we highlight two related claims that distinguish the models. The first claim concerns information processing. The standard model assumes that whatever information is processed in making intentionality judgments, the moral valence of the behavior is not critically considered; the challenger model assumes exactly such consideration. What is at issue here is a direct influence of moral valence by virtue of its morality (e.g., by a simple rule "if bad, then likely intentional"), not by virtue of factors concomitant with moral valence (e.g., rarity or difficulty of the action; see Guglielmo & Malle, under review-b). The challenger model assumes that even if all concomitant factors were controlled for, moral valence would still influence intentionality.

The second claim concerns the timing of judgments. The standard model assumes that intentionality judgments are made before genuine moral judgments whereas the challenger model assumes that moral judgments are made before intentionality judgments. For example, Nadelhoffer (2006) has argued that "Our judgments about the blameworthiness of an action may come before our determination of whether the action was performed intentionally" (p. 583).[2]

The Standard Model: Evidence for Intentionality Guiding Blame

To understand exactly how intentionality relates to moral judgments, we must first clarify what we mean by *intentionality*. Malle and Knobe (1997) examined the ordinary conception of this term and found that people require the presence of five components to deem an action intentional: the agent's *desire* for an outcome, *beliefs* about the action leading to the outcome, the *intention* to perform the action, *awareness* of the action while performing it, and a sufficient degree of *skill* to reliably perform the action. Only when all five conditions are met do people call an action intentional (Malle & Knobe, 1997).

A great deal of evidence demonstrates that variations in a behavior's intentionality produce substantial variations in people's moral judgments about the behavior.[3] In his model of responsibility and blame, Weiner (1995) argued that blame is maximal when an agent could have done otherwise but nonetheless intentionally performs a negative behavior. Similarly, Darley and Shultz (1990) reviewed evidence demonstrating that agents receive some blame when they foresee but fail to prevent harm (e.g., through negligence or recklessness) but much more blame when they intentionally bring about the harm. More recently, Cushman (2008), Lagnado and Channon (2008), and Ohtsubo (2007) have shown that a given negative behavior (e.g., cutting off a pedestrian, burning a stranger's hand) elicits substantially more blame when performed intentionally than when performed unintentionally. Mikhail (2007, 2008) has proposed that assessments of intentionality—along with those of causality and physical harm—may constitute an essential part of people's "moral grammar." On this model, the fundamental representational structure that people use when judging the morality of behavior contains a "node" that tracks the intentions of the agent(s) in question. Solan (2003, 2006), too, assigns considerations of the agent's intentionality a fundamental role in moral blame; once causality, harm, and intentionality are taken into account, a judgment of blame comes essentially for free.

---

[2] The models disagree here about people's genuine moral judgments of the specific agent's action, not about people's assessment that something desirable or undesirable happened, which any model places early in the processing chain (see Fig. 1).

[3] Extant studies have manipulated intentionality in a variety of ways. Some studies manipulated the overall intentionality of the behavior (e.g., by telling participants that the behavior was performed "on purpose" or "by mistake"). Others manipulated specific features of intentionality (e.g., by telling participants that the agent knew about and/or wanted the bad outcome to occur). Regardless of the experimental approach, results have consistently shown that stronger evidence for a negative behavior to be intentional leads to stronger perceived blameworthiness.

Integrating the extant research on intentionality and blame, our "step model" of blame (Guglielmo, Monroe, & Malle, 2009) maps the major cognitive antecedents to blame judgments and specifically localizes the role of intentionality judgments. We argue that once social perceivers decide that an agent caused a negative event, they assess intentionality. People will tend to strongly blame the agent if they regard the negative behavior as intentional (though they may reduce that blame if the agent can offer a justification for the behavior; Quigley & Tedeschi, 1996). However, even an unintentional behavior warrants substantial blame if two conditions are met: the agent had an obligation to prevent the negative event (*should have* prevented it) and the agent had the capacity to prevent the event (*could have* prevented it). Thus, judgments of intentionality guide people's path of arriving at blame and are normally causal and temporal antecedents to blame.

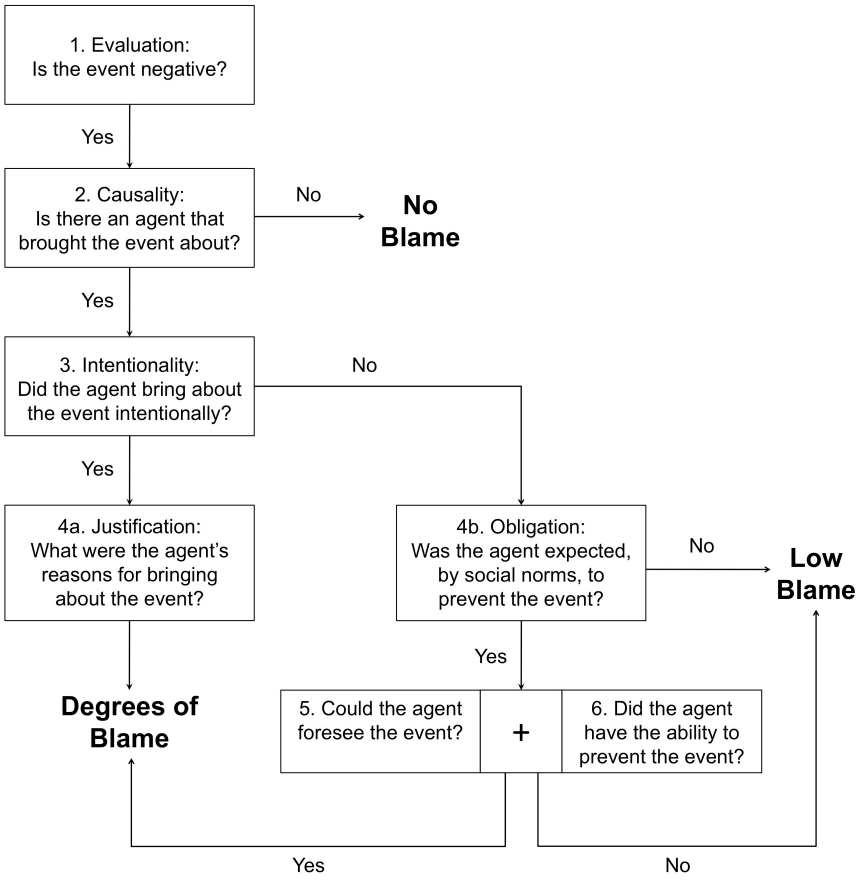

Figure 1. A step model of blame and the central role of intentionality

Reprinted with permission from Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52,* 449-466. Published Jan 10, 2009, Taylor & Francis.

The Challenger Model: Evidence for Blame Guiding Intentionality

Despite the evidence in support of the standard model of intentionality and blame, a number of findings appear to suggest the opposite model, according to which blame precedes and influences intentionality. Alicke (1992, 2000, 2008) has documented ways in which "spontaneous evaluations" (e.g., regarding an agent's negative motives or an outcome's undesirability) may influence a variety of judgments leading up to the agent's culpability. Alicke proposes that "observers may engage in a biased information search to support a desired blame attribution" (2000,

p. 567), and, even more strongly, that "everyday blamers are capable of violating virtually every rational prescription that moral philosophers, legal scholars, and rational decision theorists hold dear" (2008, p. 179).

In one study, Alicke (1992) demonstrated that a character who was speeding to hide a vial of cocaine was judged more blameworthy and more causally responsible for his ensuing car accident than was a character who was speeding to hide an anniversary gift for his parents. Alicke, Davis, and Pezzo (1994) showed that a character who shot and killed an intruder in an act of self-defense was seen as both more blameworthy and more negligent when the victim happened to be innocent than when the victim happened to be a dangerous criminal. In general, Alicke's findings show that spontaneous evaluations either directly increase blame or indirectly increase blame by influencing one of the steps in our blame model (see Figure 1), such as assuming a stronger causal link (step 2), an unjustified motive (step 4a), or the ability to prevent the bad outcome (step 6).

Thus, although Alicke's findings suggest that spontaneous evaluations may influence certain steps of our blame model they do not appear to threaten the basic structure of the model. In addition, Alicke's studies have not, at least so far, examined whether people's judgments of intentionality (step 3 in our blame model) are biased by spontaneous evaluations.

Recent work by Knobe (2003a, 2003b), however, has claimed just that: judgments of intentionality are often guided by moral evaluations. Knobe found that, under some circumstances, people appear to judge negative actions intentional but corresponding neutral or positive actions unintentional. In particular, even if an agent (i) does not intend to perform a particular action (Knobe, 2003a) or (ii) does not have the requisite skill to perform the action (Knobe, 2003b), people deem the action intentional so long as it is negative. We refer to the first pattern as the "side-effect effect" (Leslie, Knobe, and Cohen, 2006), as it suggests that people view negative side effects as intentional. We refer to the second pattern as the "skill effect," as it suggests that people view negative unskilled actions as intentional.

These findings pose a challenge for two reasons. First, they suggest that certain criteria of the folk concept of intentionality (Malle & Knobe, 1997), such as intention and skill, may not be necessary conditions for the intentionality of negative actions. Second, they suggest that an action's immorality or blameworthiness may influence perceptions of its intentionality, a claim that contradicts extant models of intentionality as well as the legal process of assessing intent to decide on guilt and punishment (Malle, 2006).

New Evidence on the Competing Models

In a series of recent and ongoing studies, we have analyzed Knobe's findings to assess whether the standard model of the intentionality-blame relationship should be abandoned in favor of Knobe's challenger model. To do so we took the experimental conditions in Knobe's original studies as a starting point and manipulated a variety of critical elements. These were Knobe's (2003a) original conditions demonstrating the side-effect effect (with the relevant differences indicated by italics):

> HARM: The vice-president of a company went to the chairman of the board[4] 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*.

---

[4] In all our replications and variations of this scenario, we introduced, for brevity, a "CEO" instead of a "chairman of the board." We will therefore refer to the protagonist of the side-effect studies as a CEO from here on out.

Did the chairman intentionally *harm* the environment? (Yes or No)

HELP: 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, and it will also *help* the environment." The chairman of the board answered, "I don't care at all about *helping* 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 *helped*.

Did the chairman intentionally *help* the environment? (Yes or No)

People's intentionality judgments varied greatly depending on the moral valence of the outcome: Whereas only 23% said the helping was intentional, 82% said the harming was intentional (Knobe, 2003a). Moreover, blame ratings in HARM were higher than praise ratings in HELP, and these ratings were correlated with judgments of intentionality—the more blame people assigned, the more likely they were to view the harming as intentional.

The conditions demonstrating the original skill effect (Knobe, 2003b) were as follows:

AUNT: Jake desperately wants to *have more money.* He knows that he will *inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window.* He raises his rifle, gets *her* in the sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet *hits her directly in the heart. She dies instantly.*

Did Jake intentionally kill his aunt? (Yes or No)

CONTEST: Jake desperately wants to *win the rifle contest.* He knows that he will *only win the contest if he hits the bulls-eye.* He raises the rifle, gets the *bull's-eye* in the sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet *lands directly on the bull's-eye. Jake wins the contest.*

Did Jake intentionally hit the bull's-eye? (Yes or No)

Again, people's intentionality judgments varied greatly depending on moral valence: whereas only 28% said that Jake hit the bull's-eye intentionally, 76% said he killed his aunt intentionally (Knobe, 2003b).

The side-effect and skill results directly speak to the first of the two differences between the contrasting intentionality-blame models: information processing. They show that manipulating the moral valence of an action (negative vs. positive or neutral) influences intentionality judgments. The important question is, however, whether it is moral valence itself or concomitant nonmoral factors that drive the effect on intentionality. We now examine this question.

What Really Guides What?

The Side-Effect Effect

In a series of studies (Guglielmo & Malle, under review a, under review b), we tested the hypothesis that Knobe's scenarios differ not only with respect to moral valence but also with respect to other informational intentionality conditions. In the side-effect vignettes, the agent's desire (or pro-attitude; Davidson, 1963) toward the outcome varies systematically between HARM and HELP. People (and the law) expect others to foster positive outcomes and to prevent negative outcomes (Pizarro, Uhlmann, & Salovey, 2003). The protagonists in HARM and HELP both defy this expectation, but with different implications. The helping CEO fails to welcome the benefit ("I don't care at all about helping the environment") and thus displays no evidence of desire or pro-attitude towards the environment. The harming CEO fails to prevent the harm to the environment, which shows some degree of pro-attitude toward the harm—he may tolerate, embrace, or even welcome it. Therefore, the harming CEO seems to show greater pro-attitude toward the outcome than does the helping CEO, which may account for the difference in intentionality judgments between the two

conditions.

Our findings supported this claim. We found that pro-attitude judgments (i.e., "how much the CEO wanted to harm/help the environment") were significantly higher in HARM than in HELP and that these judgments strongly predicted intentionality: The more the CEO was seen as wanting the harmful or helpful outcome, the more likely the harming or helping was seen as intentional (Guglielmo & Malle, under review a). It was therefore more appropriate to judge the negative case intentional, not because it was negative but because it contained intentionality-supporting information (about desire) that was clearly absent in the positive case.

We further showed that reducing the evidence of the CEO's desire for the harm led to a reduction in intentionality judgments. In this study, the CEO gave a more socially normative response when learning about the harmfulness of the program: "It would be unfortunate if the environment got harmed. But my primary concern is to increase profits. Let's start the new program." Results showed that judgments of both pro-attitude and intentionality were lower for this regretful CEO than for the original uncaring CEO, and inferred pro-attitude again strongly predicted intentionality judgments. Finally, we manipulated the pro-attitude of the helping CEO, who said: "I'm thrilled about helping the environment! And it's crucial that we increase profits. Let's start the new program." In this condition, inferred pro-attitude toward environmental benefits significantly increased, and so did people's intentionality judgments.

According to our results, therefore, Knobe's side-effect effect may have arisen not from moral differences but rather from differences in desire/pro-attitude, which happened to be confounded with the moral differences in Knobe's study.

However, differences in pro-attitude only partially account for the side-effect findings. In the help condition, desire is clearly absent, so no intentionality judgment is made; in the harm condition, desire is present but *intention* remains absent – and most people recognize this absence (Knobe, 2004; McCann, 2005). So why do people say that the CEO intentionally harmed the environment even though he lacked an intention to harm the environment?

Our answer is that people don't actually want to say that the CEO intentionally harmed the environment. To show this, we have to take a step back and consider the typical method of measuring intentionality, the dichotomous forced choice between saying "yes, it was intentional" and "no, it was not intentional." If people have a very clear representation of whether a given behavior is intentional or not, such a forced choice is unproblematic. However, there is good reason to believe that people did not have such a clear representation when facing the typical side-effect vignettes. Overall information about the scenarios was sparse, and manipulation of intentionality components (e.g., desire, intention) created ambiguity by design. In such a situation we should mistrust responses to a two-option forced choice. This is especially true because in the HARM scenario the "intentional" option and the "unintentional" option were differentially attractive. The "intentional" option was *linguistically* attractive because of the frequent co-occurrence (and thus semantic association) of the English word *intentional* and morally negative events (Malle, 2006). Conversely, the "unintentional" option was *conversationally* unattractive because to say that somebody "unintentionally harmed the environment" would normally relieve the person of blame, which participants certainly did not want to imply (Adams & Steadman, 2004; Wright & Bengson, 2009). In this situation, therefore, choosing the "intentional" option might reflect either a true intentionality judgment, a semantic association, or the rejection of an unacceptable option. Importantly, if people had doubts about the intentionality of the harm, they could not properly express those doubts because only two very different options to characterize the situation were offered.

In our studies, we therefore offered multiple response options to capture more precisely how people represent the relevant scenarios and to determine how often they would freely characterize a negative side effect as intentional (Guglielmo & Malle, under review a).

Participants read the HARM scenario and considered a list of five descriptions, selecting which was the most accurate and which one was the second-most accurate description of the CEOs behavior. In one study, the options were as follows:

"The CEO intentionally harmed the environment."

"The CEO intentionally adopted an environment-harming program."

"The CEO intentionally adopted an environment-harming and profit-raising program."

"The CEO intentionally adopted a profit-raising program that he knew would harm the environment."

"The CEO intentionally adopted a profit-raising program."

After pondering these options, just 10% of participants selected the first statement as either most accurate or second-most accurate—even though 80% of those same participants had endorsed this description when given the usual two-option (Yes or No) forced-choice question. A strong majority (70%) favored the 4th statement as most accurate, in which the CEO "intentionally adopted a profit-raising program that he knew would harm the environment." People do not appear to think that the agent brought about the side effect "intentionally." Instead, they feel that the agent performed his primary action of adopting the program (intentionally, of course) while knowing that harm would result from the action. This is a choice and it's blameworthy; but when given a chance, people distinguish between *intentionally* doing something and *knowingly* doing something—making a distinction that many legal systems confound (Malle & Nelson, 2003).

To directly demonstrate that people make this distinction we conducted a study in which people indicated whether it was most accurate to say that the CEO "intentionally harmed," "knowingly harmed," "willingly harmed," or "purposefully harmed" the environment. Nearly everyone (86%) said the CEO knowingly harmed the environment. Just 2% said that the CEO either "intentionally" or "purposefully" harmed the environment (Guglielmo & Malle, under review a).

But perhaps our measurement method was restrictive in its own way, because people had to commit to only one (most accurate) or two (most and second-most accurate) behavior descriptions. Thus, in a follow-up study we asked people to endorse all descriptions they believed were correct. Even with this lenient endorsement criterion, just 37% said the "intentionally harmed" description was correct, whereas 89% said the "knew would harm" description was correct.

In this same study, we integrated the two factors we argue are responsible for the illusory side-effect effect: the influence of pro-attitude and the influence of response options. That is, besides offering multiple behavior descriptions, we also attempted to equate the HARM and HELP scenarios with respect to the agent's expressed pro-attitude toward the side effect. To do so, we compared the standard HARM scenario to a variant of HELP in which the CEO expresses a pro-attitude toward the side-effect. When learning about the benefit to the environment, he says: "That's fantastic! It always makes me happy when our programs benefit the environment. Still, my main concern is that the company earns profits. Since the new program will increase profits, let's start it." In this HELP condition, 18% of participants said it was correct that "The CEO intentionally helped the environment," still slightly less than the 37% in the HARM condition said it was correct that "The CEO intentionally harmed the environment." Is this difference between 18% and 37% the result of a valence effect?

We think not. Despite the ratcheting up of the helping CEO's pro-attitude from 1.5 in the standard scenario to 3.1 in the modified scenario (on a 0–6 scale), the perceived pro-attitude in the HARM scenario was still somewhat higher ($M = 3.6$). Our earlier studies had shown that pro-

attitude predicts intentionality, so intentionality judgments should have been more frequent in HARM than in HELP merely because of the remaining difference in pro-attitude between the conditions. To test this prediction we conducted one last study in which the CEO, upon learning about the beneficial effect on the environment, says: "I'm thrilled about helping the environment! And it's crucial that we increase profits. Let's start the new program." For this even stronger endorsement, the pro-attitude rating reached a 4.6, and 43% of participants endorsed as correct the description "The CEO intentionally helped the environment." In the same study, the standard HARM scenario elicited the familiar 3.5 rating of pro-attitude, and 22% of participants in the HARM condition endorsed as correct the description "The CEO intentionally harmed the environment." These data further support the contention that intentionality judgments track pro-attitude, not moral valence.

Together, our studies on the side-effect effect (Guglielmo & Malle, under review a) contradict the information processing claim of the challenger model of the intentionality-blame relationship. Our results showed that Knobe's original HARM and HELP scenarios differed in the agent's pro-attitude toward the respective side effects and that pro-attitude strongly predicts intentionality judgments. Once the pro-attitude was lowered in the HARM scenario, intentionality rates decreased; once pro-attitude was raised in the HELP scenario, intentionality rates increased. Thus, the side-effect effect was, to a considerable extent, a pro-attitude effect, not a moral-valence effect. In addition, we showed that, even leaving the agent's pro-attitude toward HARM in place, most people do not view the negative side effect as intentional once they are given multiple options of describing the scenario. They prefer to describe the negative side effects as being brought about "knowingly", not "intentionally." In sum, people's systematic sensitivity to variations in pro-attitude and their consistent distinction between subtle alternatives of behavior descriptions suggest a careful process of judging intentionality even when considering highly immoral actions.

The Skill Effect

The second packet of evidence for the challenger model of intentionality and blame consists of studies on the skill effect (Knobe, 2003b). In the absence of the agent's skill of performing a behavior, the standard model of intentionality predicts infrequent intentionality judgments (Malle & Knobe, 1997). Knobe (2003b) confirmed this prediction for a neutral action of hitting the bull's-eye intentionally (28%). However, the prediction was violated for an immoral act of killing another person (76%). Do we have here decisive evidence for an impact of moral valence on intentionality judgments?

Our analysis of this effect, too, casts doubt on such a possibility. The first nonmoral element that helps account for the skill effect is an important basic action the protagonist in Knobe's (2003b) original AUNT vignette performs: pulling the trigger. In the original scenario, Jake first pulls the trigger, then slips and the shot goes wild. Consider an amateur photographer who presses the shutter but then slides off and shakes the camera. Arguably, she intentionally *took the photo*. Consider further that her shaking actually caught a moving target so perfectly that the target was in focus. Arguably, she did not intentionally *take that shot*. Our reasoning was that, just as the pressing of a shutter is a sufficient basic action that counts as taking a picture, so is the pulling of a trigger a sufficient basic action that counts as killing. If this reasoning is correct, then people's intentionality judgments should decline once the intentionality of the basic action is called into question.

Our results supported this hypothesis. Whereas nearly everyone said the killing was intentional when Jake slipped after pulling the trigger (93%), fewer said it was intentional when he slipped before pulling the trigger (71%), and even fewer said it was intentional when he slipped but there was no mention of the trigger being pulled (42%).

But if pulling the trigger "counted" as killing—thus leading people in Knobe's original AUNT case to judge the act of killing intentional—why did it not also count as "hitting the bull's-eye intentionally" in the CONTEST case?

Here we need to consider whether the two action descriptions probed in Knobe's (2003b) study were truly of equal *scope*—which refers to the ease and number of ways in which the action can be accomplished (cf. Goldman, 1970; Wegner & Vallacher, 1986). And it appears they were not. One description referred to the general action of *killing* whereas the other referred to the specific action of *hitting the bull's-eye*. We hypothesized that wide-scope actions such as killing more easily get called intentional than narrow-scope actions such as hitting the bull's eye and that this confounding of moral valence and scope helps account for the skill effect. Here is how.

Wide-scope actions are so general that many variations allow for successful performance (e.g., there all kinds of ways of killing somebody). Narrow-scope actions, by contrast, are so specific that few variations allow for successful performance (e.g., there are only a few ways of hitting the bull's eye). Thus, an agent with a given level of skill has a greater chance of intentionally performing a wide-scope action than a narrow-scope action because there are many more paths of reaching the goal. And that is particularly important for a person with low levels of skill, for whom there may be some ways of reaching a wide-scope goal but few to none of reaching a narrow-scope goal.

We can illustrate this principle with an example from an entirely nonmoral domain—geometry. Consider someone who is told to draw a line that either *intersects* or *bifurcates* an existing line drawn on a sheet of paper. To accomplish the action of intersecting, the person may draw the second line anywhere it crosses the first. To accomplish the action of bifurcating, the person must draw the second line precisely at the midpoint of the first line on the paper. Thus, the verb *intersect* has wider scope than the verb *bifurcate*—that is, there are many ways to successfully *intersect*, but only one way to successfully *bifurcate*. Imagine now two people who have very limited fine-motor skills. One is told to draw a line that bifurcates the first line. He clumsily and erratically draws a line—and it exactly bifurcates the first. We expect that most people would say that he did not do that intentionally (it was luck). The second person is told to draw a line that intersects the first. He clumsily and erratically draws a line—and it intersects the first. We expect that most people would say he did that intentionally. Given a limited amount of skill, an action that can be accomplished in many ways is easier and therefore more likely to be performed intentionally than an action that can be accomplished in only very few specific ways.

Returning to Knobe's example, two predictions follow. First, if the act of killing is of wider scope than the act of hitting the bull's-eye, killing should be seen as easier than hitting the bull's-eye. Indeed, this is what we found: the two actions differed not only in moral valence but also in difficulty (Guglielmo & Malle, under review b). Second, if Jake's immoral action were to be described with a verb of narrower scope, people's intentionality judgments should become infrequent. Once more, the prediction was borne out. Whereas 98% of participants said that Jake intentionally *killed* his aunt, only 27% said he intentionally *hit his aunt's heart*, a percentage as low as the percentage of people who said that Jake intentionally *hit the bull's-eye*. So even though Knobe's (2003b) original vignettes attempted to hold skill constant across the conditions, it was not. Whereas Jake did not have enough skill to hit the bull's eye (and did not have enough skill to hit his aunt's heart), he did have enough skill to kill. And that is why people say that he killed intentionally.

The skill effect—people's apparent tendency to view unskilled immoral actions (but not positive or neutral ones) as intentional—can thus be explained by two factors. First, the agent in the original studies intentionally performed a basic action (pulling the trigger) that counts as the broader act of killing. Once we removed this basic action, intentionality judgments fell below 50%, even

though the highly immoral outcome remained constant. Second, the act of killing is of wider scope and is therefore easier to accomplish than the act of hitting the bull's-eye. Once we equated the immoral and the neutral action for scope—comparing hitting the aunt's heart with hitting the bull's-eye—there was no longer a difference in intentionality judgments, even though the actions continued to differ in valence. A further study combined these two factors into a single manipulation. Focusing on the puzzling AUNT case, (a) we told participants that the agent slipped but we did not mention that the trigger was pressed and (b) we asked participants to judge the intentionality of the specific action (hitting the aunt's heart). Under these conditions, just 10% of participants deemed the action intentional even though they assigned a great deal of blame to the agent.

Finally, we designed a study in which there was no separation between a basic action (such as pressing the trigger) and a focal action (such as killing) and in which the difficulty and scope of the focal action was equal across conditions. In the story, the protagonist and his sister were playing darts and he attempted a final shot in the game that was either mean-spirited (in order to beat his sister, who has already had a very rough time and would be even more unhappy if she lost) or benevolent (in order to let his sister win and thus make her happy). In addition to this valence factor, the actor's skill was manipulated in the action performance (high skill: "He sets up his shot and… the dart lands ..." vs. low skill: "As he sets up his shot, he loses his balance, the dart slips out of his hand, ..."). The results of this 2×2 design were decisive: People's intentionality judgments were highly sensitive to the skill manipulation (85% for high skill and 27% for low skill) but unaffected by valence.

The Evidence So Far

Initial tests of the challenger model of blame and intentionality followed this logic: Remove a component of intentionality (such as intention or skill) from a behavior that is either negative or neutral/positive and measure people's intentionality judgments for each behavior. Initial evidence suggested that, in the absence of those components, people still consider the negative behavior intentional but do not consider the neutral/positive behavior intentional. These studies, however suffered from serious problems.

First, critical pieces of information were not held constant across the two behavior conditions. In the side-effect studies, the agent's pro-attitude toward the negative side effect was stronger than the agent's pro-attitude toward the positive side effect. In the skill studies, the negative behavior's scope (*killing*) was wider, and its difficulty lower, than the neutral behavior's scope (*hitting the bull's eye*), making a low level of skill sufficient for intentionally performing the negative behavior but not the neutral behavior. Once pro-attitudes were equal in the side-effect studies, intentionality ascriptions dropped markedly; likewise, once scope was equal in the skill studies, intentionality ascriptions dropped markedly.

Second, people identified a core action in each case that was clearly intentional but when they were asked about a different action, they were inclined to mark this different action as intentional—rather like a proxy for the clearly intentional one. In the side-effect studies, the core action was adopting a program that the agent knew would harm the environment. That is, the agent intentionally decided not to take the knowledge about harmful consequences into account, flouting the norm of preventing harm. Both the more concrete action of adopting the program and the more abstract action of flouting a norm were undoubtedly intentional actions, making it difficult for people to deny that the agent's overall behavior was intentional when asked about harming the environment. In the skill studies, the core action was pulling the trigger, which counted as killing because it occurred before the bullet entered its wayward flight and thus constituted the agent's

performed action, followed by the world not operating quite as anticipated. When the act of pulling the trigger was removed in our revised studies, intentionality judgments dropped notably.

Third, people were forced to choose between two descriptions of the behavior in question, neither of which properly captured how people conceptualized those behaviors. Of the two options, saying that the agent "did not intentionally harm" the environment" may have connoted impunity, so most people went with the option that he "intentionally harmed" the environment." But they did so only in a two-option forced-choice assessment. Once a variety of descriptions was available, people least often chose a characterization of the (negative) side effect as intentional and most often chose a description of the agent knowingly bringing about the outcome.

It's Time for Timing

The findings we have reviewed thus far cast doubt on the challenger model of the relationship between blame and intentionality. In particular, these findings contradict the challenger model's first major claim, which concerns information processing. According to this claim, genuinely moral considerations exert a direct influence on the formation of intentionality judgments, and this influence should persist even if concomitant factors are controlled for. But when we controlled for various concomitant factors—such as the agent's pro-attitude toward a side effect or the scope/difficulty of an unskilled action—intentionality judgments no longer seemed influenced by moral valence.

We now turn to the second claim—the timing of blame and intentionality judgments. The standard model and the challenger model of the relationship between intentionality and blame make distinct predictions about two aspects of the timing of these judgments.[5] The first concerns latency. The challenger model entails that blame judgments precede intentionality judgments—that is, it should take people less time to assess blame than to assess intentionality. In contrast, the standard model entails that intentionality judgments precede blame judgments, so it should take people less time to assess intentionality than to assess blame.

The second aspect of timing concerns facilitation. The challenger model claims that blame should facilitate intentionality judgments because blame guides, directs, and informs intentionality judgments. Thus, people should be faster to assess intentionality if they have first assessed blame than if they have not first assessed blame. In contrast, the standard model claims that because intentionality guides, directs, and informs blame judgments, intentionality should facilitate blame. Thus, people should be faster to assess blame if they have first assessed intentionality than if they have not first assessed intentionality.

Our ongoing research examines these contrasting sets of predictions (Guglielmo & Malle, 2009). In one study, participants read a series of sentences, each describing a negative behavior that was performed either intentionally (e.g., shoving a stranger while in line at an ATM) or unintentionally (e.g., knocking over a vase, breaking it into pieces). Following each sentence, participants provided a Yes or No button-press response to one of several questions, which were indicated by a single-word question cue. For example, the cue INTENTIONAL? stood for "Did the

---

[5] The challenger model does not make clear predictions about the relationship between praise and intentionality. Alicke's (2000, 2008) model does not incorporate praise, and Knobe's proposals have been inconsistent regarding praise. Knobe (2003b) suggested that the skill effect holds for both blameworthy negative actions (killing) and praiseworthy positive actions (saving lives). At the same time, Knobe's (2003a) side-effect demonstration contrasted a negative action (harming) with a positive action (helping). Pettit and Knobe (2009) do not mention praise and support their general claim that moral considerations influence folk-psychological judgments only with respect to negative moral valence. Thus, we focus here on predictions about negative moral judgments and use the term *blame* as a shortcut for any judgment about the immorality or blameworthiness of an action or the agent performing the action.

main character intentionally perform the behavior?" and the cue TOBLAME? stood for "Does the main character deserve to be blamed for how [s]he behaved?"

As we described earlier, the challenger model predicts that blame judgments should be faster than intentionality judgments, whereas the standard model predicts the reverse pattern. Our results contradicted the predictions of the challenger model, as people were faster to judge intentionality than to judge blame. In fact, this latency difference was strongest for the behaviors that were the most blameworthy, namely the intentional ones.

In a second study we examined moral judgments that are more basic than blame. Rather than asking whether the main character deserved to be blamed, we simply asked people whether the main character's behavior was bad. Even these basic moral judgments were no faster than judgments about intentionality. In fact, badness judgments tended to be slower than intentionality judgments, but this pattern was not statistically significant in this initial sample.

Thus, our studies on the latency predictions contradict the challenger model and support the standard model, showing that people take longer to assess blame (and, to a lesser extent, badness) than to assess intentionality.

We have also completed one study that assessed the contrasting predictions regarding facilitation. According to the challenger model, assessing blame should speed up subsequent assessments of intentionality; according to the standard model, assessing intentionality should speed up subsequent assessments of blame. Participants read several side-effect scenarios, similar to those used in Knobe's (2003a) original study and its many replications. After reading each scenario, participants answered two questions—one about blame and one about intentionality, whereby the order was randomized for each trial. We examined facilitation by comparing the latency of each judgment in the first position (i.e., when it could guide the other judgment) with the latency of the same judgment in the second position (when it could be guided by the other judgment). For example, blame would facilitate intentionality if the intentionality response latency was faster in the second position (when guided by blame) than in the first position (when guiding blame).

Contradicting the challenger model, blame did not facilitate intentionality. People's intentionality judgments were actually slightly slower (although not significantly so) when made after a prior blame judgment than when made first. Supporting the standard model, intentionality facilitated blame. People's blame judgments were much faster (by an average of 800 ms) when made after a prior intentionality judgment than when made without a prior intentionality judgment.

In sum, our initial studies on the timing of blame and intentionality judgments contradict the predictions of the challenger model. We found that people are slower to judge blame (and, to a lesser extent, badness) than they are to judge intentionality. Moreover, we found that intentionality judgments facilitated subsequent blame judgments but that blame did not facilitate subsequent intentionality judgments. These results—in combination with those discussed earlier regarding information processing—are highly problematic for the challenger model. All in all, our findings consistently support the standard model, according to which intentionality judgments both precede and guide moral judgments.

Future Research

Studies on the timing of moral judgments and intentionality judgments are a first important step to study the actual psychological processes implied by recent claims that intentionality judgments are infused with moral considerations (Doris, Knobe, & Woolfolk, 2007; Knobe, 2004; Nadelhoffer, 2006). In our research so far, we have measured the access speed of various judgments—that is, the time it takes people to report a judgment. However, we do not know whether people actually make the judgment when they encounter question cue (e.g., INTENTIONAL?) or whether they have made the judgment before seeing the question cue and retrieve this judgment

when presented with the cue. In the case of sentences as stimuli, some people may engage heavily in semantic processing and not make the relevant judgment until they see the question cue. For these people, response latencies represent the time it takes to *make* the relevant judgment. However, for people who made the judgments even before the question cue appeared, the latencies represent the time it takes to *retrieve* the relevant judgment. We are currently planning studies that address this ambiguity by presenting video stimuli and asking people to stop the video as soon as they have made a particular judgment. For example, the participant is asked to determine whether the upcoming behavior is intentional, begins to watch the video, and stops it as soon as the intentionality of the behavior is apparent. This stopping latency may be a better indicator of the time it truly takes to make the relevant judgment.

In future research, we need to explore the conditions under which certain judgments are slow or fast, not only which judgments are in general faster than others. We can expect to slow down both moral and intentionality judgments by making the behavior (or its context) ambiguous. Also, holding people accountable for their judgments—especially their moral ones—is likely to induce a deliberative process that slows down final judgments, but it would be interesting to see whether potential early flashes of evaluation are influenced by accountability as well. One might also expect that inducing affect just before people observe the stimuli could influence subsequent judgments (Goldberg, Lerner, & Tetlock, 1999). An angry, impatient, vindictive state of mind might lower the threshold for blame, and the question is whether it also influences judgments of intentionality.

Additional methodologies are needed to resolve another ambiguity in extant studies. Moral judgments, and in particular blame, are sometimes considered full-blown, deliberated assessments and sometimes flashes of approving or disapproving affect. An important, albeit difficult, question is whether such early affective flashes respond only to outcome information or whether they already take into account information about the behavior's intentionality. Physiological measures will be too slow to investigate the timing of such flashes, but ERP measures may well be able to handle the tight timing windows. Previous work suggests that there may be ERP markers for fast negative affect (Tucker et al., 2003); it remains to be seen whether any such markers can be found for intentionality judgments.

Most empirical research on morality and intentionality has focused on judgments of blame or responsibility (Shaver, 1985; Weiner, 1995), whereas far fewer studies have looked at the logic of praise, which appears to be distinct, not the mirror image of the logic of blame (Guglielmo & Malle, under review a; Ohtsubo, 2007; Solan, 2010). Moreover, judgments of intentionality and related social inferences may relate in interesting ways to other, infrequently studied moral sentiments, such as resentment, indignation, pride, and forgiveness.

The broader context of all this work is a deeper understanding of what it means to be human, to be a participant in social communities. Emotions, social cognition, and morality—as well as the capacities for language and complex relationships—are intertwined in ways that we are only beginning to understand (e.g., Malle, 2002; Tomasello, 1998). In light of this network of interrelated capacities, it may seem a somewhat narrow issue to probe the primacy of blame over intentionality or intentionality over blame. But we must know whether people's judgments of mind and action are irrevocably moral or whether people can distinguish between descriptive and normative assessments of human behavior. For if they cannot, our trust in juries, our hope for fairness, and the confidence in our own judgments may be shattered. The research reported here encourages us to maintain that trust, that hope, and that confidence.

References

Adams, F. & Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis, 64*, 173-181.

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. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*, 179-186.

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

Alicke, M. D., Weigold, M. F., & Rogers, S. L. (1990). Inferring intentions and responsibility from motives and outcomes: Evidential and extra-evidential judgments. *Social Cognition, 8*, 286-305.

Baldwin, D. A., Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences, 5*(4), 171-178.

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.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy, 60*, 685-700.

Davis, M. H. (2005). A "constituent" approach to the study of perspective taking: what are its fundamental elements? In B. F. Malle & S. D. Hodges (eds.) *Other Minds* (pp. 44-55). New York: Guilford Press.

Doris, J. M., Knobe, J., & Woolfolk, R. L. (2007). Variantism about moral responsibility. *Philosophical Perspectives, 21,* 121–183.

Fincham, F.D. & Jaspars, J.M.F. (1980). Attribution of responsibility: from man the scientist to man as lawyer. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology, 13*, (pp.82-120). New York: Academic Press.

Goldberg, J. H., Lerner, J. S. & Tetlock, P. E. (1999). Rage and reason: the psychology of the intuitive prosecutor. *European Journal of Social Psychology, 29,* 781-795.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Guglielmo, S. & Malle, B. F. (2009). *The timing of blame and intentionality: Testing the moral bias hypothesis.* Poster presented at the annual meeting of the Society for Philosophy and Psychology, Bloomington, IN.

Guglielmo, S. & Malle, B. F. (under review a). Can unintended side effects be intentional? Solving a puzzle in people's judgments of intentionality and morality.

Guglielmo, S. & Malle, B. F. (under review b). Enough skill to kill: Intentionality judgments and the moral valence of action.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 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., & Kesebir, S. (2010). Morality. In S. T. Fiske and D. Gilbert (Eds.), *The Handbook of Social Psychology* (5th Edition, pp. 181-217). Boston, MA: McGraw-Hill.

Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1, 289-313.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63*, 190-193.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*, 309-324.

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64*, 181-187.

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

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science, 17*, 421-427.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language*. Amsterdam: John Benjamins.

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

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture, 6*, 61-86.

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

Malle, B. F. & Nelson, S. E. (2003). Judging *mens rea*: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law, 21*, 563-580.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition.* Cambridge, MA: MIT Press.

McCann, H. J. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology, 18*, 737-748.

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

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development.* Cambridge: MIT Press.

Nadelhoffer, T. (2006). On trying to save the simple view. *Mind and Language, 21*, 565-586.

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

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

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*, 267-272.

Quigley, B. M., & Tedeschi, J. T. (1996). Mediating effects of blame attributions on feelings of anger. *Personality and Social Psychology Bulletin*, *22*, 1280-1288.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, *42*, 1435-1446.

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

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

Solan, L. (2006). Where does blaming come from? *Brooklyn Law Review, 71,* 941-945.

Solan, L. (2010). Blame, praise, and the structure of legal rules. *Brooklyn Law Review, 75,* xx-xx.

Tomasello, M. (1998). Social cognition and the evolution of culture. In J. Langer & M. Killen (Eds.), *Piaget, evolution, and development* (pp. 221-245). Mahwah, NJ: Erlbaum.

Tucker, D., Luu, P., Desmond, R., Hartry-Speiser, A., Davey, C., & Flaisch, T. (2003). Corticolimbic mechanisms in emotional decisions. *Emotion, 3,* 127-149.

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

Wright, J. C. & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind and Language, 24*, 24-50.

Zelazo, P. D., Astington, J. W., & Olson, D. R. (Eds.). (1999). *Developing theories of intention: Social understanding and self-control*. Mahwah, NJ: Erlbaum.