Intentionality, Morality, and Their Relationship in Human Judgment

Bertram F. Malle*

ABSTRACT

This article explores several entanglements between human judgments of intentionality and morality (blame and praise). After proposing a model of people's folk concept of intentionality I discuss three topics. First, considerations of a behavior's intentionality affect people's praise and blame of that behavior, but one study suggests that there may be an asymmetry such that blame is more affected than praise. Second, the concept of intentionality is constitutive of many legal judgments (e.g., of murder vs. manslaughter), and one study illustrates people's subtle considerations of intentionality in making those judgments. Third, controversial recent studies suggest that moral considerations can affect judgments of intentionality, and an asymmetry may exist such that blame affects those judgments more than praise. I report two new studies that may shed light on these recent findings, and I discuss several theoretical models that might account for the impact of moral considerations on intentionality judgments and for the relationship between the two more generally.

KEYWORDS

Theory of mind, moral psychology, psychology and law, social cognition, attribution

In this article I explore the relationships between judgments of intentionality and moral evaluation. I begin by introducing a model of the folk concept of intentionality – the concept that ordinary people use when making sense of behavior – and offer evidence supporting this model. Then I examine a possible asymmetry for moral evaluations such that negative evaluations might be more responsive than positive moral evaluations to information about a behavior's intentionality. Next I look at intentionality in the legal domain, but this exploration must remain brief, because there is very little empirical work available to flesh out

* Institute of Cognitive and Decision Sciences and Department of Psychology, University of Oregon. Email: bfmalle@darkwing.uoregon.edu.

the discussion. Finally, I turn to a possible asymmetry in intentionality judgments such that these judgments might differ substantially depending on whether they are made for positive or for negative human behavior. Specifically, recent studies suggest that people are more inclined to call a behavior intentional if it has negative, immoral consequences than if the same behavior has neutral or positive moral consequences. I examine the reliability and validity of these studies and then consider a number of theoretical models that try to account for the findings. I close with recommendations for future research and theory building.

The Folk Concept of Intentionality

Considerations of intentionality permeate human social life. In court and sports, in budding relationships and routine interactions, people observe and process behavior and judge it for intentionality. Such judgments are so deeply ingrained in human cognition that we might count intentionality alongside space, time, and causality as one of the fundamental categories with which the mind makes sense of the world.

The origins and developmental refinements of the intentionality concept are increasingly well documented (Astington, 2001; Baird & Baldwin, 2001; Malle, Moses, & Baldwin, 2001; Carpenter, Akhtar, & Tomasello, 1998). Within their first year of life, infants distinguish goal-directed human actions from other events (Premack, 1990; Sommerville & Woodward, 2005; Wellman & Phillips, 2001; Gergely, Nádasdy, Csibra, & Bíró, 1995) and detect intentions and actions as the structure underlying humans' stream of movement (Baldwin, Baird, Saylor, & Clark, 2001). Two-year olds are able to decipher numerous object-directed and social intentions (Baldwin, 1993; Meltzoff, 1995; Tomasello, 2001; Bloom, 2005) and expand their vocabulary and grammar of action to the sophisticated interplay of beliefs, desires, and intentions (Astington, 2001; Baird & Moses, 2001; Bartsch & Wellman, 1989).

In adulthood, considerations of intention and intentionality bring order to social perception by allowing the perceiver to structure and analyze the complex stream of behavior. The intentionality concept is also a requirement for coordinated social interaction and communication and allows humans to explain their own and others' behavior in terms

of its underlying mental causes (Malle, 1999). Finally, intentionality plays a normative role in the social and institutional evaluation of behavior through its close ties with assessments of responsibility and blame.

Intentionality is thus a tool with manifold functions, ranging from the conceptual to the interpersonal to the societal, and it is a guiding frame for various cognitive processes, ranging from perception to reasoning, from explanation to evaluation. But exactly what is this concept of intentionality? What are its components, what information are concrete judgments based on?

For a long time, the fields of psychology and philosophy relied largely on theoretical models of intentionality, and these models have disagreed widely both on the specific components that make up intentionality and on the judgment processes that employ it (Brand, 1984; Bratman, 1987; Davidson, 1963; Fiske 1989; Heider 1958; Jones & Davis 1965; Mele, 1992; Searle, 1983; Shaver 1985). In recent years, an empirical approach has taken hold that reconstructs the folk concept of intentionality using qualitative and experimental methods, and research is accumulating that reveals the components of intentionality and their interplay in concrete judgment.

In a series of studies, Joshua Knobe and I examined the agreement people show in their judgments of intentionality and identified the conditions that determine an intentionality judgment (Malle & Knobe, 1997). In a first study, participants read descriptions of 20 behaviors (e.g., Anne is sweating; got admitted to Princeton; drove way above the speed limit) and judged them for their intentionality. Agreement among judges was very high. Any two people's intentionality ratings showed an average intercorrelation of $r(20) = .64$, and any one person had an average correlation of $r(20) = .80$ with the remaining group, resulting in an interrater reliability of $\alpha = .99$.[1] More important, whether participants were provided with a definition of intentionality by the experimenter or not had absolutely no effect on average agreement. It appears, then, that people share a folk concept of intentionality that they spontaneously use to judge behaviors.

[1] A high level of agreement was also found in a replication of this study with Chinese respondents (Ames, 2000b).

In a further study, we aimed to identify the components (or "necessary conditions") that make up this folk concept (Malle & Knobe, 1997, Study 2). In an open-ended approach we asked participants to provide explicit definitions in response to the question "When you say that somebody performed an action *intentionally*, what does this mean?" After initial inspection of the definitions, two coders classified them into various categories, of which four reached substantial frequencies accounting for 96% of the meaningful definitions. These four components were *desire*, *belief, intention*, and *awareness*. To qualify for the desire category, a definition had to mention "the desire for an outcome or the outcome itself as a goal, purpose, or aim." To qualify for the belief category, a definition had to mention "beliefs or thoughts about the consequences of the act or the act itself *before* it takes place." To qualify for the intention category, a definition had to mention "the intention to perform the act, intending, meaning, deciding, choosing, or planning to perform the act." To qualify for the awareness category, a definition had to mention "awareness of the act *while* the person is performing it." Several respondents also drew careful distinctions between these components. They differentiated, for example, intention from desire ("The person meant to act that way and was motivated to do so"), belief from intention ("Someone gave thought to the action beforehand and chose to do it"), and intention from awareness ("They decided to do something and then did it with full awareness of what they were doing").

The folk concept of intentionality, as reconstructed from explicit definitions, thus encompasses four components. We suspected, however, that there was a fifth component: that of skill, or the ability to execute the action in a controlled, replicable manner (Mele & Moser, 1994). In a pilot study, participants read a story in which a novice at darts surprisingly hits a triple 20 (a very difficult throw) on his first try. His partner dismisses the throw as a fluke, so the novice tries again, this time missing badly. There was little doubt in participants' minds that the novice *wanted* to hit the triple 20 in each try (77% said so). But only 16% said that he hit it *intentionally* in the first try. Most important, when the scenario was altered to include an aspect of skill – the novice hit the triple 20 twice in a row – a significantly greater number of participants (55%) was willing to infer that the novice hit it intentionally even at his first try.

The skill component may have been omitted from explicit definitions (that yielded only four components) because people focused on social behaviors, for which skill can be assumed – in contrast to, say, artistic or athletic behaviors, for which skill cannot be assumed. But a more systematic study was needed to document this potential fifth component of intentionality.

In this study, we contrasted the conditions of ascribing an *intention* (having appropriate desires and beliefs) to the conditions of ascribing *intentionality* (also having awareness and skill; Malle & Knobe, 1997, Study 3). Awareness and skill both concern the manner in which an action is performed, which should be irrelevant for judging the presence of a mere intention (after all, the intention need not be fulfilled to recognize it). Focusing on skill, we predicted that a skill component should be necessary for judging whether an agent performed an action *intentionally* but not for judgments of intention. (See Figure 1.)

Participants read a vignette in which a person named David was flipping a coin to land on tails, which settled a debate among David and his friends over whether they should go to a movie or do something else. Three components of intentionality were manipulated: *desire* (whether or not David wants to see the movie); *belief* (whether or not he knows that "tails" stands for going to the movie), and *skill* (David's ability to make the coin land on the side he wants). (The presence of awareness was held constant throughout.) Specifically, we compared four conditions:

[D] desire present (belief and skill absent)
[B] belief present (desire and skill absent)
[D+B] desire and belief present (skill absent)
[D+B+S] desire and belief and skill present

As measures of intention judgments and intentionality judgments, participants indicated, respectively, whether they thought that David *tried* to make the coin land on tails (intention) and whether they thought that he made the coin land on tails *intentionally*.

As predicted, both belief and desire were necessary for an ascription of *intention*: 81% of participants said David tried to make the coin land on tails in condition D+B, compared to 21% and 31% in conditions D and B, respectively. Furthermore, the presence of skill was necessary for

an ascription of *intentionality*: 76% of participants said David intentionally made the coin land on tails in condition D+B+S, compared to only 3% in condition D+B. This finding not only identifies skill as a fifth component of intentionality but also highlights that people clearly distinguish between judgments of intention (trying, attempting, or planning) and judgments of intentionality (performing an action intentionally).

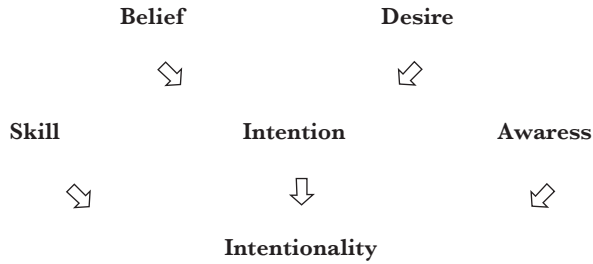

Figure 1. A model of the folk concept of intentionality (modified from Malle & Knobe, 1997, The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101-121. © Lawrence Erlbaum Associates.)

Joshua Knobe and I thus arrived at a five-component model of intentionality, displayed in Figure 1 (from Malle & Knobe, 1997). According to the folk conception, the central antecedent of an intentional action is an *intention* to perform that action, and for such an intention to be ascribed, a relevant *desire* for an outcome and one or more relevant *beliefs* about the action leading to the outcome are required. But in order for the action to be performed *intentionally*, *skill* and *awareness* have to be present as well. The awareness component specifies the agent's state of mind at the time of acting (knowing what he or she is doing), and the skill component refers to the agent's ability to perform the action as he or she intended. For example, to be convinced that my contractor intentionally affixed a bent top board on the cabinet, I need to believe all of the following: that he wanted to create, say, a modern appearance, that he believed a bent top board would assure such an appearance, that he has the skill to mount any top board he wants, and that, while working, he was aware that he was indeed installing a bent top board.

I now explore three connections between this folk concept of intentionality and morality: the influence of intentionality considerations on

moral evaluations (blame and praise); the role of intentionality judgments in legal decisions; and the influence of moral evaluations on intentionality judgments.

Asymmetries of Moral Evaluation in Light of Intentionality Judgments

In a first attempt to clarify the relationship between intentionality judgments and (moral) evaluation, Ruth Bennett and I investigated how people evaluate positive vs. negative behaviors as a function of the behavior's intentionality (Malle & Bennett, 2002). A straightforward cognitive model would predict that intentionality amplifies any evaluation – substantially increasing blame for negative behaviors and equally increasing praise for positive behaviors. A social-functional model would suggest that intentionality has more impact on the evaluation of negative behaviors than on the evaluation of positive behaviors. That is because perceivers incur great costs when confusing another person's intentional negative action as accidental (because they become vulnerable to a surprise attack), and both perceivers and agents incur costs when the perceiver falsely blames the agent's accidental negative behavior as if it were intentional, because this would amount to an act of injustice. By contrast, the costs of mistaking an intentional positive action as accidental are low (one can still consume the positive outcome), and there are no obvious costs to falsely praising an accidental positive action as if it were intentional (in fact, it has benefits for the agent). As a result, people may tend to distinguish carefully between intentional and accidental negative behaviors but can afford to be casual in distinguishing between intentional and accidental positive behaviors.

In one study we found such an asymmetry between people's use of intentionality for praising vs. blaming an action (Malle & Bennett, 2002, Study 1). Participants read a story about an "employee's phone call," whereby half of the sample read the positive version (indicated by a + sign), the other half read the negative version (indicated by a – sign).[2]

[2] After this study was conducted we discovered that Mele (1992, p. 151, n. 2) had discussed a similar thought experiment.

> Joan Ellen Edmonds, the featured agent, is a clerk at a local company. Her company hired a new clerk, Jonathan Baite, and Edmonds [+ appreciates him quite a bit] [– strongly dislikes him]. Edmonds heard that Baite [+ loves to get phone calls] [– absolutely hates to get phone calls at home], so she [+ joyfully] [– gleefully] decides to give him [+ an appreciation call the next morning when she gets to work] [– a "wake-up call" next morning when she gets to work at 6 a.m.]

A pretest showed that overall ratings of praise or blame for this first part of the positive or negative story, respectively, were equal in extremity.

The second part of the story manipulated whether the action in question was performed accidentally or intentionally. When Edmonds arrives at her office the subsequent morning, she also remembers that she wanted to call her mother. She dials that number, but, in the intentional condition, nobody answers. Then she calls Baite, who is [+ delighted] [– extremely annoyed]. In the accidental case, she dials her mother's number, but (due to a central switchboard error) she ends up reaching Baite, who is [+ delighted] [– extremely annoyed].

All participants answered (in counterbalanced order) the manipulation check question "Did Edmonds make this call to Baite intentionally?" and the question of how much praise or blame Edmonds deserves for her action of calling Baite (on 1-7 scales, labeled "a little" to "a lot"). The results showed that blame for the action performed intentionally (M = 5.7) was intensified by a factor of 3.8 over blame for the same action performed accidentally (M = 1.5). By contrast, praise for an action performed intentionally (M = 4.1) was intensified only by a factor of 1.6 over praise for the same action performed accidentally (M = 2.5), interaction $F(1, 144) = 19.3$, $p < .0001$. In terms of effect sizes, blame intensification was almost three times as large (d = 2.32) as praise intensification (d = 0.88). This asymmetry held even when we controlled for the perceived strength of intention. All subjects were asked: "How strong was Edmonds' initial intention to call Baite?" (0-10 scale, from "very weak" to "very strong"). Controlling for this rated strength of intentions did not alter the results.

This study suggests that when people evaluate a negative action, they somehow take intentionality more into account than when they evaluate a positive action. Such a tendency to pay special attention to inten-

tionality when dealing with negative actions may not be surprising to students of cultural (at least Western cultural) history. Over the centuries and millennia, the distinction between intentional and accidental negative actions has been continually reinforced. One of the earliest documentations of the intentional-accidental distinction refers to negative actions, namely, the Old Testament's distinction between intentional murder and accidental killing (Telushkin, 1994, ch. 58; see Kenny, 1973). The concept of intentionality entered the English law in the 12th century when religious concepts of free will and sin, influential in the Roman Law and Canon law, led to differentiations of punishment, such as more punishment for freely chosen crimes (Marshall, 1968). Religion and the law, then, have taught people how to differentiate blame depending on intentionality, but few cultural institutions teach people how to differentially praise depending on intentionality. One of the few historic institutions of praise, the "ars laudandi" in Renaissance Rome, had orators speak of God's and humans' laudable deeds without emphasizing their (assumed) intentionality (O'Malley, 1979). The same holds true for modern equivalents such as Nobel prizes, Olympic Games, and Oscars, where successes are assumed to be intentional and rarely discounted as "luck" (cf. the "hot hand" phenomenon, Gilovich, Vallone, & Tversky, 1985).

An archival study underscores the far closer connection of intentionality judgments with negative actions than with positive actions. I searched the LEXIS-NEXIS database of general newspaper articles for occurrences of the target words "intentional" and "intentionally." I set the time window for the search to yield about 125 excerpts for each word, which meant a very short window for the more frequent term *intentional* (November 21-30, 2004) and a longer window for *intentionally* (September 7, 2004 to March 4, 2005). After removing duplicate news stories and identical phrases within stories, 122 instances of *intentionally* and 97 of *intentional* remained. The results were striking: Overall, 94% of all instances concerned negative or socially undesirable events, 88% in the case of *intentional*, 99% in the case of *intentionally*. The negative events occurred in the domains of business (intentional deception in marketing, intentionally overvaluing stock), law (intentional homicide, intentionally setting a fire, destroying evidence, or spilling contaminated water), politics (legislators who intentionally mislead the public, voters who were

intentionally disenfranchised), and sports (e.g., intentionally walking a baseball player, intentional grounding or cut-blocking in football). The positive instances typically involved intentionally creating a positive appearance or message in art or business. So over and over again, when people encounter public declarations of intentionality, they find them associated with negative actions. The message is clear: people must prudently make these judgments because they are socially, culturally consequential.

If intentionality has a more substantial impact on evaluations of negative actions than on evaluations of positive actions, could the mere negativity of an action influence judgments of intentionality? Might the powerful association between negative actions and intentionality lead to biased assessments of intentionality when people face negative events? If so, then we have a serious problem in our legal system to the extent that it relies on judgments of intentionality: these judgments may be biased against the defendant, as people expect negative events to be more likely to be intentional. But before we reach this conclusion we need to examine more carefully how judgments of intentionality are made in legal contexts.

Intentionality in the Legal Domain

There is no doubt that the concept of intentionality plays a central role in legal decision making. Murder, for example, is typically defined as intentional killing, which requires both intentional "body movements" and the mental state of "intent to kill" (Morse, 1999, p. 148). More generally, criminal responsibility is often defined as the pairing of a harmful act and the "corresponding mental state or intent" (Felthous, 1999, p. 143). However, numerous disputes on the exact meaning of *intention* can be identified in the courts and in the literatures of philosophy and law (e.g., Duff, 1990; Kenny, 1973; Lacey 1993). The problem is that these disputes are based almost exclusively on the individual writer's intuition about what intentionality is and how it can be determined. Furthermore, these are not unfettered intuitions grounded in a representative folk concept of intentionality; rather, they are expert models often guided by theoretical and sometimes political considerations, not everyday practice.

Empirical studies on legal decision making about intentionality are surprisingly scarce. There is important work to be done, for example, on jury instructions regarding the concept of intent – instructions that often clash with jurors' own folk concept of intentionality (Malle & Nelson, 2003) – and on the specifics of people's intentionality judgments for legal events, including the concepts they rely on and the information they process. One study in this direction provides some intriguing initial results.

Angela Laurita (1998) explored ordinary people's intentionality judgments for cases that had been hotly debated in the legal or philosophical literature. (See the original paper for a full description of the cases.) The study mimicked jurors' joint decision processes without fully creating a mock-jury situation. Laurita presented four legal cases to groups of 2-4 subjects (total $N = 64$), had them discuss the merits of each case, and asked for a joint judgment about the intentionality of the defendant's behavior ("Should the defendant be convicted of intentional murder or unintentional manslaughter?"), followed by a justification of their judgment. The justifications were then content coded and the identified concepts tabulated against the intentionality judgments. The study yielded four noteworthy findings.

First, all components of intentionality that Malle and Knobe (1997) had identified were spontaneously discussed and had the predicted influence on people's verdicts. In particular, when any component (e.g., intention, skill, or awareness) was discussed as absent, intentionality was denied.

Second, there was a perfect predictive relationship between mentioning the presence of an intention and rendering an intentionality judgment (24 out of 24) and an equally perfect predictive relationship between mentioning the absence of an intention and denying an intentionality judgment (22 out of 22). This finding supports the "Simple View" of intentional action (Adams, 1986) according to which intention is a necessary requirement of intentionality.

Third, one case (*Fairview*) allowed for a focused comparison between actions performed as they had been intended and actions not performed as intended. (The case was modeled after Chisholm's story of the nephew who intends to kill his uncle by running him over and, while driving around before the planned killing, unknowingly runs over a person who turns out to be his uncle.) In the 8 instances in which people explicitly

mentioned that the action was performed as intended, all 8 verdicts were intentional murder; in the 6 cases in which people explicitly mentioned that the action deviated from the intention, all 6 verdicts were unintentional manslaughter.

Fourth, however, this same case (*Fairview*) elicited a much higher overall rate of intentionality judgments than the philosophical literature (and our original model of intentionality) would have predicted. The nephew was judged to have intentionally killed his uncle in 70% of the cases, suggesting that a good number of participants regarded the awareness component as unimportant: For at the moment the nephew ran over his uncle, he was not aware of fulfilling his intention of killing his uncle. Similarly, in the case *Paronne*, a general who killed civilians as collateral damage in the destruction of a factory was judged by 75% of the groups to have killed intentionally. That was surprising because the general insisted that he "only intended to destroy the factory, which served the war goals, but did not intend to kill the civilians." Moreover, "testimony from military strategists confirmed the assumption that no attack options were available that would have prevented civilian death." What happened here to the folk concept of intentionality that allegedly regards intention and awareness as necessary requirements of intentionality? Attempting to answer this question, Joshua Knobe launched a research program on the possible influence of moral considerations on intentionality judgments. This work I discuss next.

Asymmetries of Intentionality Judgments in Light of Moral Evaluation

Joshua Knobe (2003a, b) presented data that suggest people's judgments of a behavior's intentionality may be significantly influenced by moral considerations. In particular, Knobe (2003a) showed that when people judge the intentionality of an action that has moral consequences, they fail to consider an important component of intentionality (the agent's skill) and are quite likely to consider the immoral action intentional even if, by strict standards (and previous findings; Malle & Knobe, 1997) it may not be intentional. This finding raises a number of issues about the consistency of intentionality judgments and perhaps even the unity of

the folk concept of intentionality. Moreover, it raises the specter of a bias in people's thinking, namely to ignore important information when judging morally significant actions, which, if true, would have considerable impact on legal proceedings.

As a first step in my own explorations of this phenomenon, I conducted a replication of Knobe's (2003a) results within a different subject population (in this case, first- and second-year college students). Participants were presented with a one-page questionnaire as part of a larger survey packet. The sample consisted of 74% women, 77% Caucasians, with a median age of 18 years. The questionnaire asked participants to respond to Knobe's (2003a) rifle scenario in which an agent either hit a bull's eye with his rifle (morally neutral consequence of winning a contest) or shot a person with the rifle (immoral consequence of killing the person to acquire inheritance money). In addition, the agent was portrayed either as having skill ("expert marksman") or luck ("isn't very good at using his rifle"; "the shot goes wild, bouncing off a heavy post . . ." but nonetheless hits the target).

Of the entire sample of 228 participants, 155 read the neutral (contest) scenario, 73 read the moral (inheritance) scenario, and of each group, half saw the *skill* version and half saw the *luck* version, resulting in a 2 (moral vs. neutral) × 2 (skill vs. luck) design. The first research question was whether we could replicate Knobe's basic results; and indeed we did. When Jake hit the bull's-eye with skill, 90% of people considered it an intentional action; when he hit it with luck, only 30% considered it intentional, $d = 1.51$, $p < .001$. When Jake shot his aunt with skill, 97% of people considered it an intentional action; and when he shot her with luck, still 85% did, $d = 0.42$, $p < .06$. Thus, people adjusted their intentionality judgments in light of skill vs. luck far more for the neutral action than for the immoral action.

With this and other replications (e.g., McCann, forthcoming; Nadelhoffer, forthcoming) we have enough evidence to state confidently that people don't make intentionality judgments about immoral actions the same way they make intentionality judgments about neutral actions. But now we face an interesting puzzle. On the one hand, we learned that people seem more strongly to take intentionality into account when blaming negative actions than when praising positive actions (Malle & Bennett, 2002). On the other hand, we found that when people actually make

intentionality judgments for negative actions, they somehow take intentionality *less* into account – or seem to ignore certain components of intentionality. An account of the moral asymmetry findings will have to reconcile these two trends. The key question here is what the exact difference is – how we should interpret the findings of a "moral asymmetry" of intentionality judgments.

Interpretation of Moral Asymmetry Findings

When evaluating various interpretational models of the moral asymmetry results we have to assess how the model accounts for the available empirical data and how intelligible these asymmetries are in light of the model.

1. *A missing component of intentionality*. One attempt to account for the findings regarding moral influences of intentionality judgments is to argue that previous models (e.g., Malle & Knobe, 1997; Mele & Sverdlik, 1996) overlooked a component of intentionality: moral (or social) badness. The component doesn't play a role when people judge positive or neutral actions, but it does come into play when they judge potentially negative actions.

Badness couldn't be a necessary condition (because intentionality ascriptions are certainly made for neutral or positive actions); nor could it be a sufficient condition, because plenty of negative or immoral events are not seen as brought about intentionally. What badness might do is override considerations of other components and invite the social perceiver to render an intentionality ascription even if some standard components are missing. For example, belief+desire+skill+awareness+*badness* would be considered intentional because badness overrides the absence of intention; likewise, belief+desire+intention+awareness+*badness* would be considered intentional because badness overrides the absence of skill. Knobe's results reported in (2003a) and (2003b) are consistent with this model, showing that skill and perhaps intention can be overridden by badness considerations. In addition, Laurita's (1998) case *Fairview* (in which the nephew kills a pedestrian who turns out to be his uncle) suggests that awareness can be overridden by moral badness as long as a prior intention to bring about the immoral outcome can be assumed.

What are the limits of this override? Does moral badness supersede the absence of multiple components? Suppose the agent has no specific

intention to bring about an immoral outcome and both luckily (no skill) and unwittingly (no awareness) brings the outcome about. Would people still believe the agent acted intentionally? If so, it would imply that a desire or proattitude for an immoral outcome O and beliefs about how to achieve O can be sufficient for an intentionality judgment. But what if, in addition, either belief or desire are missing? It seems inconceivable that, just by having beliefs about how to achieve O but no proattitude toward O, the agent's (lucky and unwitting) act of bringing about O would ever be considered intentional. The interesting case is that of mere desire: An agent would like O to happen but has no idea about how to bring it about; then the agent brings O about unawares only to discover that O materialized. Could this be a case for folk intentionality judgments or merely one for the psychoanalyst?

The badness-as-component model doesn't by itself clarify why the intentionality concept has this override option. A plausible assumption is that social structures and functions invite people to set their detection criterion for negative intentional actions much more leniently than for positive intentional actions. As I argued earlier, the perceiver's costs of missing a negative intentional action are great enough to justify assuming intentionality when there is some (though limited) evidence that it might be present.

A serious problem for the badness-as-component model lies in the fact that Knobe (2003a) found intentionality ascriptions with skill absent even for an extremely *positive* outcome. This is the case of Klaus, the German soldier who luckily takes out a communication device and thereby saves innocent lives. Despite Klaus's obvious lack of skill, 92% of participants consider the lucky shot intentional (Knobe 2003a). Even though this study has, to my knowledge, not yet been replicated, it is a critical finding that any model of the morality-intentionality relation must account for, and the badness-as-component model does not. To amend the model we would have to postulate something like evaluative extremity as an override for intentionality judgments, and then the question is what function such a generalized override should serve. I will return to evaluative extremity in the discussion of model 3.

2a. *Two concepts of intentionality*. Somewhat radically, one might consider the possibility that *intentionality* simply has two different meanings – one is based on the cognitive analysis of behavior features and mental

states, the other is tracking blame (as opposed to praise). Depending on the content of a vignette and the experimenter's questions, either one or the other meaning is activated and then guides people's judgments.

We cannot rule out this interpretation, but we also don't have any evidence that favors it. There is so much overlap between the two supposed meanings that a true polysemy seems unlikely. Moreover, the evaluative aspects of intentionality develop in close connection to its cognitive aspects (e.g., Shultz, 1980), so we would need strong evidence to consider one concept to be truly separate from the other.

2b. *Two uses of intentionality*. A weaker, and perhaps more defensible thesis is that there is one intentionality concept that is somewhat flexible and can be *used* not only for cognitive tasks (e.g., behavior interpretation, explanation, and prediction) but also for evaluative tasks (e.g., praise, blame, punishment). Adams and Steadman (2004) have proposed such a thesis, suggesting that ordinary ascriptions of intentionality ("You did that on purpose!") can pragmatically imply blame, and people therefore associate judgments of intentionality with blame. Adams and Steadman (2004) further argue that people have a desire to blame the despicable acts described in Knobe's vignettes, and the intentionality judgment is one way of fulfilling that desire.

Knobe, too, sometimes favors a version of the use model, though he differs from Adams and Steadman in that he regards evaluative tasks as the primary application of intentionality judgments (Knobe, forthcoming). Knobe emphasizes that, in evaluative tasks, intentionality judgments are remarkably converging with blame judgments (but not with praise judgments) and so are made on the basis of different information features in the case of blame than in the case of praise. (For example, when blaming people don't look for information on intention or skill.) These moral considerations "are playing a helpful role in people's underlying competence itself. They make it possible for people to generate intentional action intuitions that prove helpful in the subsequent process of assessing praise and blame."

These models, though compelling, face shared as well as unique difficulties. Both models predict substantial correlations of blame and intentionality judgments when a negative action is evaluated. Knobe does not regularly report these correlations, but I have found low correlations across four studies (-.16, -.06, .37, and .40). This is problematic for the

claim that intentionality judgments are used to express blame or converge with the assessment of blame.

Adams and Steadman's (2004) model also leaves important questions unanswered: Why would study participants use intentionality judgments to express blame when they have the opportunity to indicate blame on a separate rating scale? And why do intentionality judgments normally imply blame but sometimes (as in the case of Klaus saving people) imply praise?

Knobe's model also does not specify why intentionality judgments often help with the assessment of blame and only sometimes with the assessment of praise. More important, Knobe's model leaves intentionality judgments without much function. If intentionality judgments were primarily a reflection of moral badness and, as we know, blame assessments are a reflection of moral badness (Knobe & Mendlow, forthcoming), why would we even need intentionality judgments? Their "helpful role" in blame assessments seems superfluous, no more than an idle cog (or affect) in the machinery. But of course we know that this is not true. From infancy on, intentionality judgments help in numerous nonevaluative tasks: parsing the action stream, responding to communication, learning new words, and explaining actions (Baird & Baldwin, 2001; Bartsch & Wellman, 1989; Bloom, 2005; Carpenter, Akhtar, & Tomasello, 1998; Malle, 1999).

3. *Attention and salience.* A third model tries to explain the experimental findings without necessarily assuming that the findings reflect wide-spread entanglement between intentionality and moral evaluation. The model focuses on people's attention in response to salient stimuli in the relevant experimental vignettes. Making both intentionality judgments and blame judgments side by side but keeping them cognitively separate might either tax people's attentional resources or, relatedly, may not be the salient task participants discern in the vignettes. Attentionally engulfed by the evaluative information in the stories, people might conclude that their task is to take that evaluative information into account rather than make a conceptual or "technical" judgment of intentionality. This is particularly true when the evaluative information is so extreme (e.g., killing, sacrificing or saving many people's lives) that *not* responding to it would be seen as moral indifference. People adopt a more lenient criterion for intentionality judgments of evaluatively extreme

actions not because of a conceptual dictate, but because of perceived demands to do so. This account does not deny that a moral asymmetry in judgment may exist, but it traces such an asymmetry to the force that experimental vignettes (and perhaps some real-world situations) can exert on people's attention, task interpretation, and judgments.

An advantage of the attention model is that it accommodates both moral badness findings (e.g., sacrificing lives, killing one's aunt) and the moral goodness finding (the soldier saving lives), because in each case extreme evaluative information is assumed to pull for attentional resources and to command a valence-influenced judgment. Moreover, the model is at least consistent with Malle and Bennett's (2002) phone call vignette that showed people to be *more* responsive to intentionality information (such as awareness) in the blaming of negative actions than in their praising of positive actions. At first glance, this finding seems to contradict people's apparent neglect of intentionality information in Knobe's vignettes. But the phone call story differs in two important respects from Knobe's moral vignettes: intentionality information was presented more saliently in the story than was valence information, and the valence itself was moderate rather than extreme (i.e., an unwelcome call at 6 a.m. vs. an appreciative call). With such a distribution of salience and no extreme evaluation engulfing attention, the greater social consequences of intentional negative actions may invite people to pay closer attention to intentionality information about those negative actions.

The attention model also makes testable new predictions. First, it predicts that, in the Knobe-style vignettes, participants' memory will not be as sharp about details of the intentionality information as about details of the badness perceived in the action or outcome. Second, if an experimental vignette successfully invited people to somehow decouple their blame and intentionality judgments, the morality effects should diminish or disappear (Malle & Nelson, 2003). For example, if participants can be convinced that they are judging blame in one role ("jury member") and intentionality in another ("expert psychologist"), the influence of morality on intentionality should be weakened, especially if the action is not evaluatively extreme.

4. *The scope of intentions. A* final explanatory proposal is compatible with the attention model but makes one additional assumption: that people's judgments of intentionality are sensitive to the scope of intentions.

An intention's scope refers to the range of actions and outcomes that would count as fulfilling the intention. The basic argument is simple: The vaguer the intention, the more lenient people's intentionality judgments will be. That is because the range of actions and outcomes that can fulfill a vague intention – and that can therefore be seen as "intentionally" brought about – is much broader.

An agent's intention is normally specified by verb choice. In the original scenario by Knobe (2003a), the verb chosen to describe the neutral intention was "to hit the bull's eye." However, the immoral intention was not specified by a verb; the vignette merely stated: "He knows that he will inherit a lot of money when his aunt dies." So he may have had all kinds of intentions: to kill her, to shoot her in the heart, to bring it about that she dies in any way possible, etc. This vagueness allows for many different outcomes to satisfy the agent's intention and thus for many different outcomes to be considered intentional, given that intention.

To examine more systematically the impact of verb choice and intention specificity I presented a variant of Knobe's (2003a) rifle vignette but compared three verb conditions for the immoral act in a between-subjects design: "hit his aunt's heart" ($N = 79$), "shoot his aunt" ($N = 38$), and "kill his aunt" ($N = 40$). (Malle, 2004).

The results in Table 1 suggest that when the intention is to *hit the aunt's heart* – the most specific and narrow intention – people differentiate luck from skill in their intentionality judgments, $F(1, 153) = 35.2$, $p <$

Table 1
Intentionality judgments for different verbs describing the stimulus action

	Percent Yes	
	Luck condition	Skill condition
Did Jake hit his aunt's heart intentionally? (% Yes)	49%	95%
Did Jake shoot his aunt intentionally? (% Yes)	84%	90%
Did Jake kill his aunt intentionally? (% Yes)	100%	100%

.001, $d = 1.18$. The vaguer verbs *shoot* and *kill* do not elicit luck-skill differentiation.

I would argue that this verb effect is primarily due to the difficulty of fulfilling a specific intention such as to hit a small body part such as someone's heart. To strengthen this interpretation I conducted an additional study in which the verb was held constant (*shoot*) but the object of the shooting intention varied: someone's arms, hips, or head. These body parts arguably decrease in size, therefore increase in difficulty of hitting with a rifle, and therefore increase in intention specificity. As expected, the body part manipulation had a significant effect on intentionality judgments such that more people regarded the act of shooting an arm as intentional (94%) than the act of shooting a hip (80%) than the act of shooting the head (69%), $F(1, 92) = 6.1$, $p < .05$. There was also an interesting (but nonsignificant) linear trend such that a differentiation between skillfully and luckily hitting the target varied with body part. When the shooter hit the arm, 94% of participants in the luck condition and 93% in the skill condition considered the shooting intentional. When he hit the hip, the numbers were 71% for luck and 89% for skill, and when he hit the head, they were 59% for luck and 80% for skill. Note that shooting someone's head is more harmful than shooting someone's arm or hip. So if it were mere moral badness that drove luck-skill differentiation in intentionality judgments, the head condition should show less differentiation ("it's really bad, so it's intentional, whether skilled or not"). In actuality, the head condition showed *more* differentiation (21%) than the arm condition (1%). This finding is consistent with the interpretation that vagueness of intention drives the asymmetry findings more than evaluative extremity.

Thus, at least part of people's blending of luck and skill conditions in Knobe's (and other) experiments may be due to the fact that an intention such as to *kill* is so vague as to allow all kinds of scenarios to fulfill that intention. When the intention becomes more specific (e.g., shooting a specific body part), people are more reluctant to ascribe intentionality under lucky circumstances.

Summary. I have considered four models that aim at explaining some intriguing findings on the relationship between morality and intentionality judgments. The models that propose a conceptual solution (1. and 2.) face a number of problems in accounting for the empirical data at hand. The salience model has quite a few strengths and offers new and

testable hypotheses. The scope of intention model, lastly, builds on the salience model and assigns a specific role to people's interpretation of what it takes to fulfill intentions – which is often vaguer and more flexible in the case of immoral intentions (at least as described in the standard vignettes).

Future Work

There is no doubt that questions about intentionality and morality are both inherently interesting and practically relevant for understanding social interactions, legal proceedings, and political events. There is also no doubt, however, that claims and theories about these phenomena are empirically on thin ice. Few studies have explored the complex responses of blame and praise as a function of intentionality; the leniency of intentionality judgments in the face of morally extreme events; or the impact of the folk concept of intentionality on legal decision making. The lacuna of empirical work on intentionality judgments in the legal domain is particularly deplorable, considering that intentionality has both a central and controversial status in the law.

Thus, we need more empirical research, but good empirical research; and we need more theory, but good theory. I close with a few recommendations about future research and theory building, hoping to elicit critique, adjustment, and expansion by other scholars and, perhaps in some cases, implementation in future work.

Improving methodology. One major problem with virtually all extant studies on morality and intentionality is that participants are forced to make judgments in the researchers' terms, whether on rating scales of forced-choice items. When asked, people prefer to say that a reprehensible behavior is intentional rather than unintentional. But when simply describing the behavior in question or retelling the entire story surrounding it, would they imply anything about the behavior's intentionality? Perhaps intentionality is not of great concern when people encounter a clearly reprehensible behavior; instead they are concerned with justice and proper moral sentiments. We don't know people's immediate thoughts when they read our experimental vignettes and we have only guesses about which salient aspects of those vignettes guide their judgments. Structured interviews could explore people's unfettered perceptions through

initially open-ended probes followed by increasingly specific questions and requests for clarification and justification.

A second problem with extant studies is that we present participants with impoverished stimulus materials that feature fictitious strangers in contrived scenarios. If intentionality and morality judgments are confounded or at least tightly related in ordinary life, we should conduct studies that capture these judgments in ordinary life – that is, vis-à-vis actual behaviors and outcomes, and within intimate relationships (Pearce, 2003).

Theory building. Initial studies on an interesting phenomenon typically document novel findings but cannot be expected to offer compelling theory. This is clearly the case for the findings on a possible moral asymmetry in intentionality judgments. But in light of a range of newly available experimental results we are in a position to focus more on theory development and to test specific predictions of competing theories. Many questions await answers, and a good theory should be able to provide them. If the "simple view" of action (according to which intentionality judgments necessarily imply intention ascriptions; Adams, 1986) is not correct, as Knobe (forthcoming) claims, how then do people make intentionality judgments? What are the exact affective and cognitive processes that lead people to consider or ignore relevant components of intentionality? Is the key ingredient for moral asymmetries in intentionality judgments the badness of the behavior or the desire to not let a "violator" get away with anything? If the violator was "punished" by an apparent act of fate, would people's judgments of intentionality be freed from the need to express moral affect and to effect justice?

Current research on the evaluative responses of blame and praise also need to be integrated into a comprehensive theoretical framework. Shaver (1985), Weiner (1995), and Alicke (2000) took major steps in this direction, but the methodological limitations of the empirical research base – involving deliberate responses to questionnaires depicting stranger behaviors – are limiting sound theory building (Pearce, 2003). A comprehensive model needs to integrate all of the following processes: immediate affective responses and delayed reasoning; responses to stranger and intimates; blame as a mental state and blame as a communicative act; blame for personally relevant outcomes and blame for "external" events; and blame as a function of various components of intentionality.

Interdisciplinary endeavor. The recent explorations into the relationship of morality and intentionality have been compellingly interdisciplinary –

with psychologists caring about conceptual problems and philosophers caring about empirical data. I would hope that other disciplines will join this exciting effort. We should invite sociologists to clarify the social-cultural role of intentionality beyond the cognitive level; anthropologists to speak to the evolutionary origins of intentionality and morality judgments; developmentalists to offer insight into the time course of developing intentionality judgments and their (perhaps increasing) susceptibility to affective influences; and neuroscientists to provide data on the relations between judgments of morality and intentionality at the neural substrate level. It will take a while to integrate the various theories and data, but only an interdisciplinary approach is the proper way to understanding this fundamentally human phenomenon in all its complexity.

REFERENCES

Adams, F.
1986 Intention and intentional action: The Simple View. *Mind and Language, 1*, 281-301.

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

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

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. 85-104). Cambridge, MA: MIT Press.

Baird, J. A., & Baldwin, D. A.
2001 Making sense of human behavior: Action parsing and intentional inference. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 193-206). Cambridge, MA: MIT Press.

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

Baldwin, D. A.
1993 Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*, 832-843.

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

Bartsch, K., & Wellman, H.
1989 Young children's attribution of action to beliefs and desires. *Child Development, 60*, 946-964.

Bloom, P.
2005 Word learning, intentions, and discourse. *Journal of the Learning Sciences, 14*, 311-314.

Brand, M.
1984 *Intending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.

Bratman, M. E.
1987 *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Carpenter, M., Akhtar, N., & Tomasello, M.
1998 Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21*, 315-330.

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

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

Felthous, A. R.
1999 Introduction to mental illness and criminal responsibility. *Behavioral Sciences and the Law, 17*, 143-146.

Fiske, S. T.
1989 Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought: Limits of awareness, intention, and control* (pp. 253-283). New York: Guilford.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S.
1995 Taking the intentional stance at 12 months of age. *Cognition, 56*, 165-193.

Gilovich, T., Vallone, R., & Tversky, A.
1985 The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

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

Jones, E. E., & Davis, K. E.
1965 From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219-266). New York: Academic Press.

Kenny, A.
1973 The history of intention in ethics. In A. Kenny, *The anatomy of the soul. Historical essays in the philosophy of mind* (pp. 129-146). Oxford: Basil Blackwell.

Knobe, J.
2003a Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology 16*, 309-324.
2003b Intentional action and side effects in ordinary language. *Analysis*, 63, 190-193.
forthcoming The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*.

KNOBE, J. & MENDLOW, G.
forthcoming The Good, the Bad, and the Blameworthy: Understanding the Role of Evaluative Considerations in Folk Psychology. *Journal of Theoretical and Philosophical Psychology*.

LACEY, N.
1993 A clear concept of intention: Elusive or illusory? *The Modern Law Review, 56*, 621-642.

LAURITA, A.
1998 *The concept of intentionality underlying people's judgments of criminal behavior*. Unpublished Honor's thesis, University of Oregon.

MALLE, B. F.
1999 How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*, 23-48.
2004 The moral dimension of intentionality judgments. *Technical Reports of the Institute of Cognitive and Decision Sciences, No. 04-2*, Eugene, Oregon. Available electronically at http://hebb.uoregon.edu/04-02tech.pdf.

MALLE, B. F., & BENNETT, R. E.
2002 People's praise and blame for intentions and actions: Implications of the folk concept of intentionality. *Technical Reports of the Institute of Cognitive and Decision Sciences, No. 02-2*, Eugene, Oregon. Available electronically at http:// hebb.uoregon.edu/02-02tech.pdf.

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.
forthcoming Intentional action and intending: Recent empirical studies. *Philosophical Psychology*.

MELE, A. R.
1992 *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.

MELE, A. R., & MOSER, P. K.
1994 Intentional action. *Nous, 28*, 39-68.

MELE, A. R., & SVERDLIK, S.
1996 Intention, Intentional Action, and Moral Responsibility. *Philosophical Studies*, 82, 265-87.

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

MORSE, S. J.
1999 Craziness and criminal responsibility. *Behavioral Sciences and the Law, 17*, 147-164.

Moses, L. J.
1993 Young children's understanding of belief constraints on intention. *Cognitive Development, 8*, 1-25.

Nadelhoffer, T.
forthcoming Skill, luck, control, and folk ascriptions of intentional action. *Philosophical Psychology*.

O'Malley, J. W.
1979 *Praise and blame in Renaissance Rome: Rhetoric, doctrine, and reform in the sacred orators of the papal court, c. 1450-1521*. Durham: Duke University Press.

Pearce, G. E.
2003 *The everyday psychology of blame*. Unpublished doctoral dissertation, University of Oregon, Eugene.

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

Searle, J. R.
1983 *Intentionality: An essay in the philosophy of mind*. Cambridge, England: Cambridge University Press.

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

Shultz, T. R.
1980 Development of the concept of intention. In W. A. Collins (Ed.), *The Minnesota Symposium on Child Psychology* (Vol. 13, pp. 131-164). Hillsdale, NJ: Erlbaum.

Sommerville, J. A., & Woodward, A. L.
2005 Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition, 95*, 1-30.

Telushkin, J., Rabbi
1994 *Jewish wisdom*. New York: William Morrow & Co.

Tomasello, M.
1999 Having intentions, understanding intentions, and understanding communicative intentions. In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), *Developing theories of intention: Social understanding and self control* (pp. 63-75). Mahwah, NJ: Erlbaum.
2001 Perceiving intentions and learning words in the second year of life. In
M. Tomasello and E. Bates (Eds.), *Language development: The essential reading* (pp. 111-128). Malden, MA: Blackwell.

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

Wellman, H. W., & 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: MIT Press.