Inquiry,
Vol. 52, No. 5, 449–466, October 2009

Routledge
Taylor & Francis Group

At the Heart of Morality Lies Folk Psychology

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

Brown University, USA

(Received 23 February 2009)

ABSTRACT *Moral judgments about an agent's behavior are enmeshed with inferences about the agent's mind. Folk psychology—the system that enables such inferences—therefore lies at the heart of moral judgment. We examine three related folk-psychological concepts that together shape people's judgments of blame: intentionality, choice, and free will. We discuss people's understanding and use of these concepts, address recent findings that challenge the autonomous role of these concepts in moral judgment, and conclude that choice is the fundamental concept of the three, defining the core of folk psychology in moral judgment.*

I. Introduction

When people make a moral judgment they evaluate an agent's behavior in light of a system of norms. Such evaluations of behavior are enmeshed with inferences about what was in the agent's mind before, while, and even after performing the behavior. If folk psychology[1] is the system of concepts and processes that enable a human social perceiver to make such inferences from behavior, then folk psychology lies at the heart of moral judgment.

We can see the folk-psychological roots of moral judgment in specific phenomena such as blaming, assigning responsibility, feeling resentment or sympathy. Explications of each of these psychological phenomena refer to the assumptions that social perceivers make about human capacities—about what makes a being an *agent* and how agents can act *intentionally*—and assumptions about how mental processes contribute to such actions. Blame and responsibility, to take one example, are assigned in consideration of a

Correspondence Address: Steve Guglielmo, Brown University, Department of Psychology, 89 Waterman St., Providence, RI 02912, USA. Email: steve_guglielmo@brown.edu

0020-174X Print/1502-3923 Online/09/050449–18 © 2009 Taylor & Francis
DOI: 10.1080/00201740903302600

person's principled capacity to reason about various paths of action and to actually choose one such path. Even when harm occurs unintentionally, if the person could have and should have chosen a harm-avoiding alternate path, blame and responsibility apply (Malle, Moses, and Baldwin, 2001). Resentment, to take another example, is a more complex emotion than mere anger, because it relies on the assumption that the target of one's sentiment chose an unjust act but could have chosen a more just one.

In this paper we explore two core concepts of folk psychology and their role in moral judgment: intentionality and choice. We will discuss people's understanding and use of these concepts, as well as how the concepts fit together to shape people's moral judgments. We first discuss a basic model of the role of intentionality in moral blame; then we address recent findings that challenge this model; and finally we explore the folk concept of free will and its relationship to blame and the larger folk-psychological framework.

II. Moral judgment

Recent psychological work on moral judgment has focused on the processes that make up such judgment—cognitive and affective, deliberate and automatic (e.g., Greene, Sommerville, Nystrom, Darley, and Cohen, 2001; Haidt, 2001; Hauser, Cushman, Young, Jin, and Mikhail, 2007; Pizarro, Uhlmann, and Bloom, 2003). An equally important question is what *concepts* are employed when moral judgments are made, especially the concepts that categorize features of agents and their behavior. Even though harm and destruction can by themselves arouse negative emotions (e.g., anger or horror), humans do not normally make moral judgments about earthquakes or hurricanes. What makes judgments moral is that they are directed at *agents* who are presumed, accused, or shown to have performed behaviors that caused or permitted harm to occur. But people do not stop at establishing the presence of an agent and a behavior; they look to the agent's mind to better understand what kind of behavior was performed and how it relates to the harm. That is, moral judgments are sensitive to the agent's desires and intentions, beliefs and knowledge, social obligations (and their recognition), and abilities to bring about and avert the outcomes at issues (Cushman, Young, and Hauser, 2006; Greene and Haidt, 2002; Hamilton, 1978; Shaver, 1985; Weiner, 1995). We suggest that these various factors can be integrated in a model of blame that is sketched in Figure 1 and discussed next.

III. Intentionality and blame

The step model of blame tries to capture the questions that social perceivers normally ask and the concepts they employ when assessing blame.[2] Steps 1 and 2 are uncontroversial: detecting a negative event and identifying a potential target for blame, typically grounded in judgments of causal

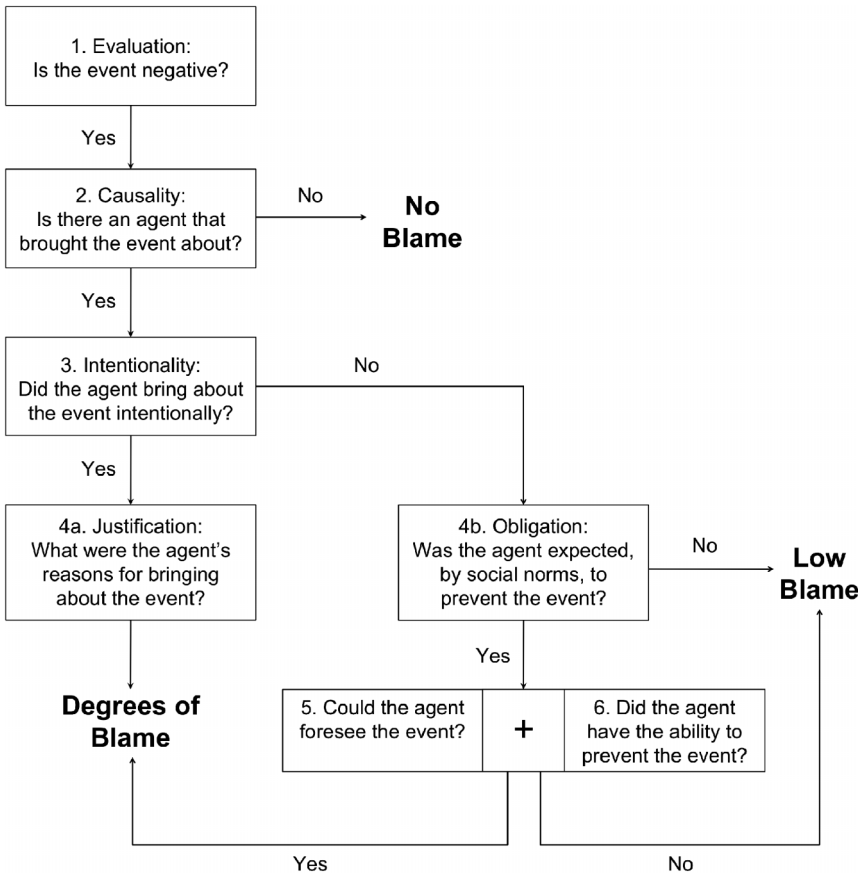

Figure 1. Step model of folk assessments of blame.

involvement (Sloman, Fernbach, and Ewing, 2008). The key role of folk psychology emerges in step 3 with the powerful contribution of the intentionality concept. This contribution goes beyond the well-documented fact that intentional actions are blamed more than (comparable) unintentional behaviors or events[3] (e.g., Lagnado and Channon, 2008; Ohtsubo, 2007; Shultz and Wells, 1985). To clearly see the role of intentionality in blame we need briefly consider the components that make up the folk concept of intentionality—conditions that all have to be met for a behavior to count as intentional (Figure 2). People require evidence for the agent's *desire* for an outcome, *beliefs* about the action in question leading to the outcome, the *intention* to perform the action, *awareness* of the act while performing it, and a sufficient degree of *skill* to reliably perform the action (Malle and Knobe, 1997).

The distinction between intentional and unintentional events provides the central step 3 of the blame model by bifurcating the additional information

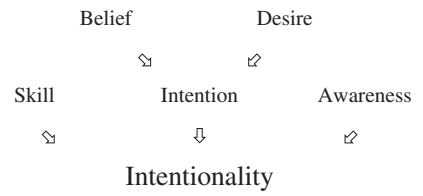

Figure 2. A model of the folk concept of intentionality.

the perceiver considers: the agent's particular reasons to act on the one hand (step 4a) and the combination of obligation and ability to prevent on the other hand (step 4b). The reasons in step 4a are of course the beliefs and desires that lead to an intention to act, indicating that once people determine that an agent acted intentionally they want to know *why* he or she acted this way. These reasons why will strongly increase or decrease blame. An agent who hurt someone intentionally may have had acceptable reasons (the dentist wanting to extract an unhealthy tooth) or unacceptable reasons (a schoolboy wanting to provoke a fight) and will be blamed less or more, respectively.[4]

But even when the agent did not bring about the negative event intentionally, intentionality judgments play a critical role. In such cases, the perceiver examines whether the agent *should have* tried to prevent the event (i.e., step 4b: obligation) and *could have* tried to prevent it (i.e., steps 5 and 6: necessary knowledge, skills, and opportunities). Failing to prevent a negative event when one should and could prevent it is an intentional rejection of one's obligation, which will trigger substantial blame. Here, people again base their blame judgments on the concepts of intention and intentional action, though this time they are applied to counterfactual considerations.

In this analysis of moral judgment, then, intentionality judgments critically guide and influence subsequent judgments of blame (Malle *et al.,* 2001; Solan, 2003). Recent work by Knobe (2003a, 2003b), however, challenged this model and proposed instead that early (and perhaps unconscious) assessments of blame can bias people's intentionality judgments. That is, already before step 3, the social perceiver has assigned blame to the agent and is *therefore* more likely to see the behavior as intentional. People do not, according to this hypothesis, assess intentionality to designate blame but instead assess blame to designate intentionality. Such a pattern not only has theoretical implications for models of moral judgment but would also threaten the practice of asking people to make *mens rea* judgments in legal proceedings (Nadelhoffer, 2006). We should be wary of jurors' assessments of a defendant's intent if the negative valence of the presumed criminal act biases them toward seeing such intent.

A second threat to the above model of blame comes from Alicke (2008), who suggests 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)—and such prescriptions include the stepwise progression in Figure 1. Alicke (2000) proposed that social perceivers' blame judgments can be influenced by "spontaneous evaluations" that, according to normative models, should not influence blame. That is, observation of certain agents (e.g., social outcasts, persons with a criminal history) or of certain events (e.g., those with severe consequences or that threaten beliefs in a just world) either directly increase feelings of blame; or they indirectly increase blame by influencing one of the other steps, such as being inclined to assume a stronger causal link (step 2), infer a malevolent motive (step 4a), or assume greater ability to prevent than is warranted (step 6). However, even though Alicke's findings show that spontaneous evaluations can influence important steps in the assessment of blame, they do not contradict the structure or order of steps we depict in our model.

Knobe's challenge, generally consistent with Alicke's (2000) analysis, makes a stronger claim. Alicke's model is one of blame judgments, and his studies never assess whether people consider a given behavior intentional. Knobe (2003a, 2003b) argued that the very concept of intentionality is imbued with moral meaning—that the use of this concept in judgment depends itself on the moral goodness or badness of a behavior (Knobe and Burra, 2006).[5] The empirical evidence for this claim comes from Knobe's own studies and replications thereof, so we must carefully examine the evidence so as to probe whether Knobe's challenge rattles the very core of the model of blame sketched earlier—that is, whether moral feelings can influence an intentionality judgment so early in the process of blaming that they override or circumvent the conceptual structure of intentionality as laid down in folk psychology. If so, the central role of folk psychology in moral judgment would be weakened because people's moral judgments actually serve as input to applications of folk-psychological concepts (Knobe, 2005), not as output of such applications.

IV. Knobe's challenge

In brief, Knobe (2003a, 2003b) suggested that people sometimes view morally significant behaviors as intentional while seeing structurally identical neutral behaviors as unintentional. In Knobe's and subsequent researchers' studies (Cushman and Mele, 2008; Knobe and Burra, 2006; Machery, 2008; Mallon, 2008; Nadelhoffer, 2004, 2005; Nichols and Ulatowski, 2007), the patterns of results were typically these: participants who considered neutral outcomes and learned that components such as the agent's intention or skill were missing concluded that the agent did not intentionally bring about that outcome; by contrast, participants who considered morally significant outcomes and learned that the agent's intention or skill were missing nonetheless indicated that the agent intentionally brought about that outcome.

These findings challenge both the dominant models of intentionality (because two components believed to be necessary may not be necessary) as well as our step model of blame (because blaming evaluations would appear to precede the intentionality assessment in step 3, rather than the other way around). We will focus on the challenge to intentionality models and then return to reassess our model of blame.

Two sets of findings have been marshaled to support this hypothesis of a moral bias in intentionality judgments. One set has been dubbed the "side-effect effect" (Leslie, Knobe, and Cohen, 2006) and concerns intentionality judgments for side effects—outcomes that are consequences of an agent's action but that the agent did not intend to bring about. In Knobe's (2003a) study, when a CEO adopted an environment-helping program to increase profits but "did not care at all about" helping the environment, very few people (23%) indicated, when the environment was indeed helped, that the CEO helped the environment intentionally. However, when the CEO adopted an environment-*harming* program to increase profits and "did not care at all about" harming the environment, most people (82%) indicated, when the environment was indeed harmed, that the CEO harmed the environment intentionally. Notably, people did not think that the CEO *intended* to harm the environment (Knobe, 2004; McCann, 2005), only that he harmed it *intentionally*. These surprising results pose a challenge to most models of intentionality (e.g., Adams, 1986; Malle and Knobe, 1997; Searle, 1983), which assume that if an agent does not intend to perform an action, the action cannot be intentional. According to Knobe's (2003a) side-effect findings, such a model does not hold true for immoral behavior.

A second set of blame bias findings is dubbed the "skill effect" and has similar implications. According to several models of intentionality (e.g., Malle and Knobe, 1997; Mele and Moser, 1994), any action performed unskillfully (i.e., relying on luck) cannot be intentional. The skill effect findings suggest that this is true for neutral behaviors but not negative ones. Knobe (2003b) showed that when an agent fired a lucky shot to hit a bull's-eye, very few people (28%) thought he hit the bull's-eye intentionally. However, when the agent fired a lucky shot that hit his aunt and killed her, most people (76%) thought he killed her intentionally.

In a recent series of studies, we have systematically examined both the side-effect effect (Guglielmo and Malle, 2009a) and the skill effect (Guglielmo and Malle, 2009b) to determine why these puzzling results were found, whether the standard models of intentionality should indeed be replaced, and what the implications are for models of blame.

V. Intentionality, blame, and the side-effect effect

One factor relevant to the difference in intentionality judgments between Knobe's (2003a) harm and help conditions is the agent's attitude toward the

outcome. Both the law and people themselves expect others, when possible, to foster positive outcomes and to reject or avoid negative outcomes (Pizarro, Uhlmann, & Salovey, 2003). By not caring about harming/helping the environment, both CEOs 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 sharply distances himself from it; this person has no pro-attitude towards the environment (Davidson, 1963). In contrast, 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 a relatively 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 have supported this claim. One study revealed that pro-attitude judgments (i.e., "how much the CEO wanted to harm/help the environment") were higher in the harm condition than in the help condition and that these judgments strongly predicted intentionality: the more the CEO wanted the harmful or helpful outcome, the more likely people saw the harming or helping as intentional.

A second study compared the original harm scenario with one in which the CEO gave a more 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." In this regret version, people were less likely to view the act of harming as intentional. Moreover, pro-attitude judgments were lower for the regretful CEO than for the original uncaring CEO, and these judgments were again strong predictors of intentionality.

Thus, in Knobe's study people viewed the harming as intentional (but the helping as unintentional) at least in part because the CEO's lack of care was interpreted differently in these conditions. Whereas "not caring" about a positive outcome indicates a true lack of desire for the outcome, not caring about a negative outcome indicates an endorsing—even a slight desire—for the outcome. These differences in pro-attitude—not the difference in valence or blameworthiness—accounted for people's different perceptions of intentionality.

However, Knobe's side-effect findings were surprising for a second reason, namely that people said the CEO intentionally harmed the environment despite lacking an intention to do so (Knobe, 2004; McCann, 2005). How could this be? Importantly, nearly all previous studies have given people only a Yes/No intentionality question, and people may be reluctant to assert that the harm was entirely *unintentional* on the CEO's part (Adams and Steadman, 2004). After all, many actions were intentional, such as adopting the program and defying the norm to prevent harm, and people may feel that answering "no" means claiming that he did nothing intentionally and further means excusing the CEO's behavior. Therefore, people are left with

saying "yes" to the intentionality question as the most reasonable response option. But do they really see the act of harming as intentional?

To answer this question we gave 30 participants five descriptions of the CEO's behavior and asked them to indicate which of the descriptions was the most accurate. Participants overwhelmingly endorsed the description: "The CEO intentionally adopted a profit-raising program that he knew would harm the environment", which 70% of people selected as the most accurate. By contrast, only 16% of people selected as most accurate or second-most accurate either of the descriptions that framed the CEO's behavior primarily as harming the environment (i.e., "The CEO intentionally harmed the environment" or "The CEO intentionally adopted an environment-harming program"). Importantly, blame judgments were very high regardless of which description people chose as the most accurate. In a subsequent study, participants could select as many statements as they considered to be correct descriptions of the scenario. Here, 88% indicated that it was correct to say "The CEO intentionally adopted a profit-raising program that he knew would harm the environment", whereas only 35% believed that it was also correct that "The CEO intentionally harmed the environment".

Thus, most people don't see the act of harming itself as intentional—they appear to so judge it only when given a forced-choice Yes/No question. In this case, the "intentional" response is warranted by the CEO's pro-attitude toward the outcome and blatant intentional disregard for his obligation to prevent the harm. Our findings therefore counter Knobe's challenge and vindicate the concept of intentionality. Unintended side-effects are not seen as intentional; however, when an agent fails to prevent known harm, *this* choice to violate the norm of prevention is intentional and incurs a great deal of blame.

These findings underscore the central role of choice in judgments of blame. When an agent chooses a path of action that leads to a negative side-effect, blame judgments are heightened, but people understand that the effect itself was not intentional. As shown in the step model of blame (Figure 1), obligation (step 4b), foresight (step 5) and preventive ability (step 6) can all influence blame even when the negative outcome is not intentional. The CEO had the obligation to prevent harm, foresaw that it would ensue from his actions, and had the ability to prevent it. Consideration of these steps, then, explains why participants view negative side-effects as highly blameworthy while recognizing them to be unintentional.

VI. Intentionality, blame, and the skill effect

What do Knobe's (2003b) skill-effect findings tell us about the concept of intentionality and the role of choice in blame judgments? The original findings were surprising because an unskilled agent's (immoral) action was judged intentional. However, his lack of skill did not become evident until

after he had pressed the trigger intentionally ("[Jake] presses the trigger. But Jake isn't very good at using his rifle. His hand slips . . ."). Jake chose to initiate a basic action (i.e., pressing the trigger) that may count as killing, likely making his subsequent slip irrelevant to participants. Thus, we reasoned that if the intentionality of this basic action was called into question, people would less often deem the act of killing intentional (Guglielmo and Malle, 2009b). Our results supported this hypothesis. Whereas nearly everyone said the killing was intentional when Jake slipped after pressing the trigger (93%), fewer said it was intentional when he slipped before pressing the trigger (71%), and even fewer said it was intentional when he slipped but there was no mention of the trigger being pressed (42%).

Therefore, it appears that people in Knobe's original study who viewed the killing as intentional did so because the shooter intentionally performed a critical basic action that counted as the broader act of killing. For that easy action, the agent had sufficient skill. When the skillful performance of even this basic action was in doubt, people less often viewed the act of killing as intentional.

One additional factor helps explain Knobe's original findings without resorting to issues of morality. Knobe's two conditions (the neutral hitting the bull's-eye and the immoral killing of the aunt) differed substantially in action specificity. One referred to a general action of killing whereas the other referred to a specific action of hitting the bull's-eye. Action specificity, we hypothesize, is an index of difficulty, because the more specific an action, the fewer variations are allowed for that action to be successfully completed. Indeed, we found that people viewed hitting the bull's-eye as more challenging than killing the aunt. Moreover, when we equated the specificity of the actions at issue, the intentionality disparity disappeared: 38% said Jake intentionally *hit* the bull's-eye and 27% said he intentionally *hit* his aunt's heart.

In short, for easy actions (such as killing a person in one way or another), little skill is needed; therefore, even a generally unskilled person has sufficient skill to perform that action intentionally. For difficult actions (such as hitting the bull's-eye), considerable skill is needed; therefore, a generally unskilled person does not have sufficient skill to perform the action intentionally. People's careful consideration of the interplay between skill and difficulty explains Knobe's and other researchers' findings and restores confidence in the standard model of intentionality, in which skill is one of the necessary conditions for judging an action intentional.

VII. Interim conclusion: the significance of choice

In addition to providing evidence for people's sensitive judgments of intentionality even in the face of morally significant behaviors, these studies highlight what lies at the core of the intentionality concept: the notion of

intention or choice. In the side-effects studies, the agent chose to violate a norm of prevention (which was highly blameworthy), but he did not specifically choose to harm the environment. In the skill studies, the agent chose to press the trigger and thereby chose to kill (even without much skill), because he initiated an immoral action plan that led to the desired outcome, even if the execution of that plan was somewhat altered.

The capacity to choose, to consider information and integrate it with one's desires, is perhaps the key ingredient in the folk distinction between intentional action and other events. Rocks, water, and atoms don't make choices; humans do. Heider (1958) argued that people have two very different models of causality: that of intentional action, and all the rest. What makes causality in the form of intentional action unique is that all causal forces are funneled through an intention before they are behaviorally manifested.

The significance of the concept of choice comes further into focus when we examine people's folk conception of free will; for there, too, choice is central. This may be obvious to some, but the literature has not presented matters that way. Because the issue of free will is so intimately tied to questions of responsibility and the latter is another way of asking what underlies assignments of blame, we will briefly review what is known about the folk concept of free will and then apply it to our final observations about choice and blame.

VIII. The folk conception of free will

In recent discussions of people's concept of free will, the concept is often treated as rather metaphysical. We use the term *metaphysical* to capture two features: one, it refers to the deepest level at which reality can be described (e.g., as unfolding processes and events; Whitehead, 1929); two, it refers to matters that go beyond justification by scientific evidence. Focusing on the latter, scholarly portrayals characterize people's concept of free will as a special form of causation. According to Wegner (2002), ordinary people believe that "our experiences of conscious will cause our actions" (p. 318). Similarly, Prinz (2003) characterizes the ordinary concept of free will as the "notion that the act follows the will, in the sense that physical action is caused by mental events that precede them" (p. 26). A stronger charge is that mental causation is fundamentally different from the accepted scientific concept of causation and that the will implies some type of "nonstandard" causality. Prinz (1997) argues that the folk concept demands the "replacement of usual causal determination through another, causally inexplicable form of determination" (p. 161). It follows that the folk concept entails a "renunciation of explanation and [a] cutting short of causal chains" (Prinz, 1997, p. 162). Most radical is the charge that the "jargon of free will in everyday language . . . requires us to accept local pockets of indeterminism in an otherwise deterministically conceived world view" (Maasen, Prinz, and Roth, 2003, p. 8)

and that ordinary people hold the assumption that "willfulness somehow springs forth from some special uncaused place" (Bayer, Ferguson, and Gollwitzer, 2003, p. 100).

All these characterizations of the folk concept of free will rely on scholars' own intuitions about this concept. However, what people really believe, assume, and theorize about is a matter of empirical fact, not intuition. Curiously little work has examined ordinary people's beliefs about the nature of free will, and we therefore know very little about its role in people's moral reasoning. Although experimental philosophers have recently turned their attention to folk conceptions of free will, they have focused primarily on whether people adhere to a compatibilist or incompatibilist view of the universe (Nahmias, Morris, Nadelhoffer and Turner, 2005; Nichols, 2004, 2006). We must take the prior step of deciphering the folk concept itself. The criteria that underlie folk concepts can be identified by asking people about the meaning of expressions or terms for which the concepts are relevant (Malle and Knobe, 1997). In our case, if *free will* has a systematic and shared meaning, we would expect to find a reliable and consensual pattern of responses.

We (Monroe and Malle, in press) explored this question by asking 180 participants to "*explain in a few lines what you think it means to have free will*." The dominant definition (mentioned by 65% of participants) referred to the ability to make a decision or a choice. Acting in accordance with one's desires was mentioned by 33% of participants. Being free from external or internal constraints was mentioned by 29%. We interpret these responses as pointing to two core ideas: making a choice that is in harmony with one's desires, and being free from overwhelming constraints.

The data showed no indication that people consider choice as an uncaused cause or a magical, indeterministic process. Rather, choice is the act of forming an intention in light of and because of relevant desires and beliefs. Other research has documented that people explain intentions and intentional actions by appealing both to an agent's temporally proximal reasons that guide the forming of an intention and to more distal factors that lie in the "causal history of reasons," such as personality, social forces, or unconscious impulses (Malle, 1999; Malle, Knobe, O'Laughlin, Pearce, and Nelson, 2000). Thus, choice is seen as part of complex causal networks just like other events in the world.

We should note that our data are mute regarding people's understanding of how choice is implemented. People's conception is a functional one, referring to choice as a reasoning process within a network of mental states rather than a miraculous moment of causal inception. This functional concept appears to be quite underspecified. In our study, people provided little information on exactly how beliefs and desires combine into intentions, exactly what goes on in the mind during an act of choice, or where in the physical world such choice can reside. These questions of (physical and metaphysical) implementation are not what people's folk concepts are designed

to answer. In the evolution of the social mind, free choice and free action have not been recognized by their neural or physical signature but by their place in the nexus of information, deliberation, mental effort, and controlled movement. Many models of implementation are compatible with this conception, which is one reason why the folk conceptions of choice and intentional action are largely unscathed by recent findings on unconscious and neural mechanisms of action control (Malle, 2006).

If people's concept of free will does not really speak to the fundamental nature of reality and causality, then the tension that philosophers have diagnosed between free will and moral responsibility reduces, for ordinary people, to the question of how choice is related to responsibility. And if conditions of responsibility largely overlap with conditions of blame, then our earlier model of blame clarifies the folk-psychological handling of the sticky topics of free will and responsibility. We will illustrate this point in a brief discussion of blame mitigation in the law and everyday life, which helps clarify both the meaning of "free" in the folk conception of free will and highlights once more the folk-psychological foundations of moral judgment.

IX. Free choice and mitigation of blame

The law provides for mitigation of blame in special cases when a wrongful act was committed but the culprit lacks the requisite mental states to choose the action—cases of insanity, diminished capacity, and also coercion and duress. The diminished capacity and insanity defenses assert that, at least at the time of the crime, the accused lacked access to the mental abilities a person would normally utilize to choose a course of action: that is, the actor (e.g., a person in an acute psychotic state) did not have the capacity to reason and *make a choice* about his or her actions. Duress or coercion defenses, by contrast, make no reference to a limited choice *mechanism* but rather to a limitation of *options* in the person's choice: the actor was unduly constrained by either external or internal forces that blocked or dictated certain paths of action (e.g., being forced at gunpoint to commit a criminal act).

These two lines of mitigation map well onto the conditions of free will that participants provided in Monroe and Malle (in press): that an actual choice must be made corresponds to the first, the choice mechanism defense; that the choice should be in accordance with one's desires and not constrained by other forces corresponds to the second, the limited option defense. Thus, *free choice* is the key requirement for full blame: a genuine *choice* that is *free* from option limitations. An insanity defense claims that an individual could not make a choice, whereas a coercion defense claims that even though a choice was made it was not free.

The existing literature on blame mitigation has usually focused on one of these limitations or confounded them under the label of *control* (e.g., Alicke, 1990, 2000; Weiner, 1995; Woolfolk, Doris and Darley, 2006). In a recent

project (Monroe and Malle, unpublished data), we set out to examine both paths of blame mitigation and explore their relative success.

In our initial study we asked 43 participants to evaluate the blameworthiness of a severe act of aggression (shattering someone's jaw with a rock) in light of several limiting conditions that were designed to occur at roughly four different stages of action production: distal personal limits (e.g., child abuse), limits to deliberation (e.g., schizophrenia), limits to intention formation (post-hypnotic command), and limits to action execution (motor disturbance). Whereas distal limits should concern more the freedom of a choice (by making certain action paths undesirable or unperformable), limits to deliberation and intention formation should concern the very mechanism of choice. Limits of action execution come after the choice and, if choice is pivotal to blame, should have little mitigating power.

The mean levels of blame reduction suggested three groups of limitations. Blame was mitigated most strongly when the mechanism of choice appeared to be wholly disabled (e.g., brain tumor, post-hypnotic command, schizophrenia). Blame was mitigated less when the choice mechanism was arguably intact but the output of the choice was constrained in some way (e.g., overwhelming emotions, mob influence, and abuse as a child). Blame was not mitigated when only the execution of the act itself was perturbed (e.g., premature or delayed action).

These findings must be replicated with a larger array of scenarios, including actual court cases. However, we may draw the following preliminary conclusions in light of our earlier model of blame (Figure 1).

Choice is the core of step 3, *Intentionality,* and a disabled choice mechanism (e.g., psychosis) may either stop the process at once or quickly negate step 4b, *Obligation*, because somebody with no capacity for choice is not obligated to prevent harm. To the extent that the capacity limitation is temporary or weak, the person may still have the obligation to prevent harm—as in the case of strong emotions, which people are expected to regulate at least to some degree.

When the choice mechanism is intact and the behavior in question is not accidental, mitigation must operate via the reasons the person had for acting. When Sophie Zawistowska in *Sophie's Choice* (Styron, 1979) was forced, by a Nazi doctor, to decide which one of her two children would die immediately and which one would continue to live, the only other option she had was that both children would die. Such severely limited options provide justifiable reasons for an otherwise terrible choice.[6] By contrast, the soldiers in Abu Ghraib who tortured prisoners chose their actions for reasons that most people do not find acceptable (e.g., orders form higher up), and no situationist defense (e.g., Fiske, Harris, and Cuddy, 2004; Zimbardo, 2007) appears to convince people otherwise.

Finally, the lack of any mitigation from failures in action execution mirrors our earlier discussion of the skill effect. Choosing to perform a highly

negative behavior (killing the aunt) warrants a great deal of blame, even if the outcome does not perfectly occur as planned (the bullet's wayward path). Because step 3 is complete (the agent pulled the trigger), only the justification for the decision (step 4a) determines the degree of blame—here, the killing is motivated by personal financial gain, a socially unacceptable reason. There is no separate condition that checks for what happens after the decision.

In our next studies we will expand the number of mitigating factors (*cf.* Alicke, 1990) and test the hypothesis that their effectiveness to reduce blame depends largely on the degree to which they (a) paralyze the choice mechanism or (b) limit the options to choose from. We will also attempt to track in more detail the cognitive processes that social perceivers engage in so we can determine whether the blame model makes correct predictions about the perceiver's subsequent steps of reasoning (i.e., consideration of reasons vs. obligations).

X. Lessons about intentionality, choice, and blame

Our aim in this paper was to illustrate the role of intentionality and choice in people's moral judgments, particularly in judgments of blame. First, we addressed a set of recent findings by Knobe (2003a, 2003b) that showed an apparent blame bias in people's intentionality judgments—namely, that people ignore certain intentionality components when considering blameworthy behaviors and are therefore more ready to judge these behaviors as intentional. Our analysis of these findings revealed that there is no reason to suspect this kind of blame bias. Failing to prevent a known negative side-effect represents a choice to violate a norm of prevention, and this choice is worthy of much blame. In fact, our findings underscore the importance of choice on judgments of blame. The CEO chose to adopt an environment-harming program and the shooter chose to initiate a harmful plan of action—for these choices, people unwaveringly apply a great deal of blame.

We next examined people's conception of free will and found that it is characterized by two primary features: the capacity to choose a course of action based on one's beliefs and desires, and freedom from external or internal constrains when making such a choice. These findings and previous research on action explanation suggest that people do not view free will as an uncaused cause but as an integration of mental states in a process of reasoning and forming an intention to act.

Finally, we explored how this understanding of free will guides judgments of blame. If free will consists of a capacity for choice and the absence of overwhelming constraints, then the breakdown of this capacity or the presence of such constraints should mitigate blame. We found this to be the case, with blame judgments decreasing most significantly when the capacity for choice breaks down.

We can then summarize how folk psychology guides moral judgment. Folk psychology provides the fundamental distinction between intentional and unintentional behavior and directs blame judgments along two different paths. For intentional action, what counts is the capacity to choose and the degree to which the agent's choice was free from strong constraints; for unintentional behaviors, what counts is the interplay of foreknowledge, obligation to prevent negative outcomes, and the actual ability to prevent those outcomes. This conceptual framework appears to stand solid even when people face highly negative actions and outcomes, reaffirming the sophistication and resiliency of folk psychological concepts and judgments.

Notes

1. We use the term *folk psychology* here to refer to the same system that has also been labeled *theory of mind*, *naïve psychology,* and *common-sense psychology*. For a discussion of the ingredients of this system, see Malle (2005, 2008).
2. A few caveats: first, there is no assumption that people necessarily answer the questions in an accurate or unbiased way. Second, in the current sketch the model streamlines some steps by leaving out, for example, the agent's commitment to the act (i.e., degree of planning and investment), which would expand step 4a, and the interplay of the agent's ability and situational opportunities or limitations in step 6. Finally, the role of coercion is discussed in section 9 of this article.
3. An exploration of the origin of this asymmetry must be reserved for another occasion.
4. In the literature, this step is often described as the blame-mitigating factor of "justification", with the later-discussed factor of coercion being one such justification.
5. More recently, Knobe (2006) has weakened his original claim—arguing that not only moral but also nonmoral evaluations can relax intentionality assessments (Knobe & Mendlow 2004); that there is indeed an intentionality concept both in folk psychology and in moral analysis but that evaluative considerations can guide the perceiver to attend to some components of intentionality more than to others; and that such guidance may in turn be useful for judgments of blame or praise (Knobe, 2006). The challenge remains that (a) intentionality can be influenced by evaluative or moral considerations and is therefore not an independent assessment of the observed behavior and that (b) the concept of intentionality is not a stable set of necessary conditions but a loose collection of components that may be used in different ways for different purposes.
6. Nonetheless, people may have difficulty withholding blame altogether because a genuine choice was made; and indeed, Sophie never ceases to blame herself for what she did.

References

Adams, F. (1986) "Intention and intentional action: The simple view", *Mind and Language*, 1, pp. 281–301.

Adams, F. & Steadman, A. (2004) "Intentional action in ordinary language: Core concept or pragmatic understanding?", *Analysis*, 64, pp. 173–81.

Alicke, M. D. (1990) "Incapacitating conditions and alterations of blame", *Journal of Social Behavior and Personality*, 5, pp. 651–64.

Alicke, M. D. (2000) "Culpable control and the psychology of blame", *Psychological Bulletin*, 126, pp. 556–74.

Alicke, M. D. (2008) "Blaming badly", *Journal of Cognition and Culture*, 8, pp. 179–86.

Bayer, U. C., Ferguson, M. J., & Gollwitzer, P. M. (2003) "Voluntary action from the perspective of social-personality psychology", in: S. Maasen, W. Prinz, and G. Roth (Eds.), *Voluntary Action; Brains, Minds, and Sociality*, pp. 86–107 (New York: Oxford University Press).

Cushman, F. & Mele, A. R. (2008) "Intentional action: Two-and-a-half folk concepts?", in: J. Knobe and S. Nichols (Eds.), *Experimental Philosophy* (New York: Oxford University Press).

Cushman, F., Young, L., & Hauser, M. (2006) "The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm", *Psychological Science*, 17, pp. 1082–89.

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

Fiske, S. T., Harris, L. T., & Cuddy, A. J. (2004) "Why ordinary people torture enemy prisoners", *Science*, 306, pp. 1482–83.

Greene, J. & Haidt, J. (2002) "How (and where) does moral judgment work?", *Trends in Cognitive Sciences*, 6, pp. 517–23.

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, pp. 2105–08.

Guglielmo, S. & Malle, B. F. (2009a) "Can unintended side effects be intentional? Solving a puzzle in people's judgments of intentionality and morality", manuscript submitted for publication, Brown University.

Guglielmo, S. & Malle, B. F. (2009b) "Enough skill to kill: Intentionality judgments and the moral valence of action", manuscript submitted for publication, University of Oregon.

Haidt, J. (2001) "The emotional dog and its rational tail: A social intuitionist approach to moral judgment", *Psychological Review*, 108, pp. 814–34.

Hamilton, V. L. (1978) "Who is responsible? Towards a social psychology of responsibility attribution", *Social Psychology*, 41, pp. 316–28.

Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007) "A dissociation between moral judgments and justifications", *Mind and Language*, 22, pp. 1–21.

Heider, F. (1958) *The Psychology of Interpersonal Relations* (New York: Wiley).

Knobe, J. (2003a) "Intentional action and side effects in ordinary language", *Analysis*, 63, pp. 190–93.

Knobe, J. (2003b) "Intentional action in folk psychology: An experimental investigation", *Philosophical Psychology*, 16, pp. 309–24.

Knobe, J. (2004) "Intention, intentional action and moral considerations", *Analysis*, 64, pp. 181–87.

Knobe, J. (2005) "Theory of mind and moral cognition: Exploring the connections", *Trends in Cognitive Sciences*, 9, pp. 357–59.

Knobe, J. (2006) "The concept of intentional action: A case study in the uses of folk psychology", *Philosophical Studies*, 130, pp. 203–31.

Knobe, J. & Burra, A. (2006) "The folk concepts of intention and intentional action: A cross-cultural study", *Journal of Cognition and Culture*, 6, pp. 113–32.

Knobe, J. & Mendlow, G. (2004) "The good, the bad and the blameworthy: Understanding the role of evaluative reasoning in folk psychology", *Journal of Theoretical and Philosophical Psychology*, 24, pp. 252–58.

Lagnado, D. A., & Channon, S. (2008) "Judgments of cause and blame: The effects of intentionality and foreseeability", *Cognition*, 108, pp. 754–70.

Leslie, A. M., Knobe, J., & Cohen, A. (2006) "Acting intentionally and the side-effect effect: Theory of mind and moral judgment", *Psychological Science*, 17, pp. 421–27.

Maasen S., Prinz, W., & Roth (2003) *Voluntary Action; Brains, Minds, and Sociality* (New York: Oxford University Press).

Machery, E. (2008) "The folk concept of intentional action: Philosophical and experimental issues", *Mind and Language*, 23, pp. 165–89.

Malle, B. F. (1999) "How people explain behavior: A new theoretical framework", *Personality and Social Psychology Review*, 3, pp. 23–48.
Malle, B. F. (2005) "Folk theory of mind: Conceptual foundations of human social cognition", in: R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The New Unconscious*, pp. 225–55 (New York: Oxford University Press).
Malle, B. F. (2006) "Of windmills and strawmen: Folk assumptions of mind and action", in: S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition*, pp. 207–31 (Cambridge, MA: MIT Press).
Malle, B. F. (2008) "The fundamental tools, and possibly universals, of social cognition", in: R. Sorrentino and S. Yamaguchi (Eds.), *Handbook of Motivation and Cognition Across Cultures*, pp. 267–96 (New York: Elsevier/Academic Press).
Malle, B. F. & Knobe, J. (1997) "The folk concept of intentionality", *Journal of Experimental Social Psychology*, 33, pp. 101–21.
Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E, & Nelson, S. E. (2000) "Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions", *Journal of Personality and Social Psychology*, 79, pp. 309–26.
Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001) "The significance of intentionality", in: B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*, pp. 1–24 (Cambridge, MA: MIT Press).
Mallon, R. (2008) "Knobe vs. Machery: Testing the trade-off hypothesis", *Mind and Language*, 23, pp. 247–55.
McCann, H. J. (2005) "Intentional action and intending: Recent empirical studies", *Philosophical Psychology*, 18, pp. 737–48.
Mele, A. R. & Moser, P. K. (1994) "Intentional action", *Noûs*, 28, pp. 39–68.
Monroe, A. E., & Malle, B. F. (in press) "From uncaused will to conscious choice: The need to study, not speculate about, people's folk concept of free will", *European Review of Philosophy.*
Nadelhoffer, T. (2004) "The Butler problem revisited", *Analysis*, 64, pp. 277–84.
Nadelhoffer, T. (2005) "Skill, luck, control, and intentional action", *Philosophical Psychology*, 18, pp. 341–52.
Nadelhoffer, T. (2006) "Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality", *Philosophical Explorations*, 9, pp. 203–20.
Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005) "Surveying freedom: Folk intuitions about free will and moral responsibility", *Philosophical Psychology*, 18, pp. 561–84.
Nichols, S. (2004) "The folk psychology of free will: Fits and starts", *Mind and Language*, 19, pp. 473–502.
Nichols, S. (2006) "Folk intuitions on free will", *Journal of Cognition and Culture*, 6, pp. 331–42.
Nichols, S. & Ulatowski, J. (2007) "Intuitions and individual differences: The Knobe effect revisited", *Mind and Language*, 22(4), pp. 346–65.
Ohtsubo, Y. (2007) "Perceiver intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect", *Japanese Psychological Research*, 49, pp. 100–10.
Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003) "Causal deviance and the attribution of moral responsibility", *Journal of Experimental Social Psychology*, 39, pp. 653–60.
Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003) "Asymmetry in judgments of moral blame and praise: The role of perceived metadesires", *Psychological Science*, 14, pp. 267–72.
Prinz, W. (1997) "Explaining voluntary action: The role of mental content", in: M. Carrier and P. K. Machamer (Eds.), *Mindscapes: Philosophy, Science, and the Mind*, pp. 153–75 (Pittsburgh, PA: University of Pittsburgh Press).
Prinz, W. (2003) "How do we know about our own actions?", in: S. Maasen, W. Prinz, and G. Roth (Eds.), *Voluntary Action; Brains, Minds, and Sociality*, pp. 21–33 (New York: Oxford University Press).

Searle, J. R. (1983) *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press).

Shaver, K. G. (1985) *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. (New York: Springer).

Shultz, T. R. & Wells, D. (1985) "Judging the intentionality of action-outcomes", *Developmental Psychology*, 21, pp. 83–89.

Sloman, S. A., Fernbach, P., & Ewing, S. (2008) "Causal models: The representational infrastructure for moral judgment", in: B. H.Ross (Series Ed.) & D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Psychology of Learning and Motivation* (*Vol. 50: Moral judgment and decision making*) (San Diego, CA: Academic Press).

Solan, L. M. (2003) "Cognitive foundations of the impulse to blame", *Brooklyn Law Review*, 68, pp. 1003–28.

Styron, W. (1979) *Sophie's Choice* (New York: Random House).

Wegner, D. M. (2002) *The Illusion of Conscious Will* (Cambridge, MA: MIT Press).

Weiner, B. (1995) *Judgments of Responsibility: A Foundation for a Theory of Social Conduct* (New York: Guilford).

Whitehead, A. N. (1929) *Process and Reality* (New York: Macmillan).

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

Zimbardo, P. (2007) *The Lucifer Effect: Understanding How Good People Turn Evil* (New York: Random House).