

# Different moral values produce different judgments of intentional action

David Tannenbaum and Peter H. Ditto

University of California, Irvine

David A. Pizarro

Cornell University

**Note:** This is a draft paper, please do not distribute without permission of the author

## Abstract

Recent work in social psychology and experimental philosophy has suggested that moral considerations (praise vs. blame) can influence judgments about the intentional status of an act, contradicting both lay and legal assumptions about the relationship between theory of mind and morality. A corollary of this account suggests that different assessments of intentional action should emerge whenever people hold different moral values. Five studies validated this implication. Participants were more likely to report an action as intentionally caused if it led to a negative side-effect that had strong moral significance to the participant. This pattern was found when looking at differences in participants' protected values (Experiments 1 & 2), political orientation (Experiments 3 & 4), and gender (Experiment 5). These findings support the claim that moral values can strongly influence lay concepts of intentional action, providing an explanation of how people may arrive at different judgments of intentionality while nevertheless agreeing on the actors state of mind.

Comic-book heroes (as well as villains) are often endowed with the superpower of telepathy—the ability to “listen in” on the private thoughts and beliefs of others. Telepathy is a compelling superpower because the ability to accurately perceive the thoughts and intentions of any complex agent allows for one to explain, understand, and preempt that agents behavior. The better we can mind-read, it seems, the more likely we are to be successful in navigating through a complex and unpredictable world<sup>1</sup> (Dennett, 1987).

If superheroes have the power to read minds, what is there for the rest of us? Ordinary humans of course also make inferences about the mental states of others, and almost always do so with limited information. A long and rich tradition in social psychology (Heider, 1958; Jones & Davis, 1965; Kelley, 1972; Malle, Knobe, O’Laughlin, Pearce, & Nelson,

---

<sup>1</sup>This is not to say that complete knowledge of others’ mental states is always beneficial. In bargaining situations—where elements of both conflict and mutualism can be found—a lack of knowledge about the others motives can actually increase ones bargaining power, as long as the opponent is knowledgeable of this deficiency (Schelling, 1963).

2000; Shaver, 1985; Weiner, 1995), as well as an equally rich literature in developmental psychology and philosophy of mind (Baron-Cohen, 1995; Brattman, 1987; Dennett, 1987; Meltzoff & Gopnik, 1993), has established many of the operating principles underlying this naïve psychology or “theory of mind.”

The current research is specifically interested in how people make judgments about whether an action was performed *intentionally*. Intentionality, as used here, is a basic conceptual framework that is used to assess and evaluate the meaning of a person’s actions. When we judge another’s behaviors (or our own) as intentional, we are determining that the act was performed in a particular manner. More specifically, a recent model of the lay concept of intentionality (Malle & Knobe, 1997) has posited that intentional behavior is composed of five, hierarchically-arranged components. According to this model, an intentional act is composed of an intention to perform the act (component 1), which is further comprised of a desire to perform the action (component 2), and a belief that one can perform the action (component 3), along with having the ability to perform the action (component 4) and a conscious awareness that one is performing the action (component 5). All five components are thought necessary to judge an act as intentional; if one of them is missing, then the behavior is usually judged as unintentionally performed.

Importantly, empirical investigations of this model demonstrate that people treat the concept of intentionality separate from that of having an intention (Malle & Knobe, 1997). The concept of intention differs from that of intentionality in several important ways. Intention is a particular mental state that is ascribed to an *agent*, whereas intentionality is an evaluative concept that is ascribed to an agent’s *actions*. This explains why intention is only one of several mental states that comprises our concepts of intentional action. Moreover, intention and intentionality serve different functions. Intentions are future-oriented mental states—they are the expression (or ascription) of future plans that serve to coordinate behavior both interpersonally and intrapersonally (for a full treatment of this topic, see Brattman, 1987). By contrast, intentionality allows one to assess and evaluate the worth of an action that has already been performed (Alicke, in press; Malle & Bennett, unpublished manuscript).

The fact that judgments of intentional action also play heavily into moral evaluations should not come as a surprise. Intentionality is an evaluative concept, so it lends itself well to assessments of actions that draw forth moral praise or blame. This relationship is reflected at the very core of the legal code—for an act to be considered legally culpable, one’s bodily movements need to be determined as casually responsible for the act (*Actus Reas*) and also accompanied by a set of mental states that suggest wrong-doing (*Mens Rea*). The *Mens Rea* requirement is vital, as legal scholar Stephen Morse has suggested, “because full understanding of action itself and an action’s moral significance depends upon the meaning of the action. Mental states signal both that what the agent has done is wrong and how wrong it is” (Morse, 2003, p. 23). This common view implies that it is first necessary to establish intentionality (a theory of mind judgment) in order to establish blame (a moral judgment).

### *The Side-Effect Effect*

A recent set of findings by Knobe (2003a, 2003b, 2004) has called into question the notion of an unidirectional link from theory of mind to moral judgments. Participants were

asked to evaluate an agent who knowingly brings about an undesired side effect. When asked if the agent intentionally brought about the side-effect, participants' responses were largely determined by whether the side-effect led to a positive or negative outcome. If the agent's actions led to a negative side-effect, a large majority of participants claimed that the agent intentionally brought about those side-effects. If the agent's actions led to a positive side-effect, a majority reported that the agent *did not* intentionally bring about those side-effects. The paradigmatic demonstration of this effect used the following scenario (Knobe, 2003a):

The vice president of a company went to the chairman of the board and said, "Were thinking of starting a new program. It will help us increase profits, and it will also harm[help] the environment."

The chairman of the board answered, "I dont care at all about harming[helping] the environment. I just want to make as much profit as I can. Lets start the new program."

They started the new program. Sure enough, the environment was harmed[helped].

Most participants reported that the chairman intentionally harmed the environment, but relatively few participants reported that the chairman intentionally helped the environment. This general tendency for the valence of a known side-effect to influence judgments of intentionality has been referred to as the *side-effect effect* (Leslie, Knobe, & Cohen, 2006).

The notion that the chairman should be judged as more blameworthy when the environment is harmed and less praiseworthy when the environment is helped is consistent with a large body of psychological research. People intuitively follow the principle that one is strongly obligated to prevent causing morally bad outcomes, but the obligation to bring about morally good outcomes is much weaker (Grueneich, 1982). The asymmetry in placing greater evaluative weight for negative acts relative to equally positive acts reflects the general negativity bias that pervades virtually all domains of human behavior<sup>2</sup> (Baumeister, Bratslavsky, Finkhenauer, & Vohs, 2001; Rozin & Royzman, 2001). Kahneman and Tversky's (1979) famous dictum that "losses loom larger than gains" aptly reflects this sentiment.

The vexing question, however, is why a similar asymmetry would arise with regard to judgments of intentional action? If the lay concept of intentionality is strictly concerned with a person's mental state (e.g. "Was she aware that her actions would lead to X? Did she have an intention to X?") then why would such judgments be influenced by the outcome of the act?

There is little consensus in how to answer this question. Several commentators have assumed that the moral status of the act (in this case, harming or helping the environment)

---

<sup>2</sup>Interestingly, (Morewedge, 2007) has also documented a negativity bias for perceptions of intentional agency, whereby people are more likely to attribute agentic qualities to misfortunate outcomes. For example, participants were more likely to report their computers and automobiles as having "their own beliefs and desires" the more these objects malfunctioned. However, it is not clear that this negative agency bias would explain the basic side-effect effect, as it is virtually certain that all participants—both those who reported that the chairman acted intentionally and those who did not—would nevertheless report that the chairman was an intentional agent.

does indeed influence intentionality judgments, but they differ in whether moral considerations *should* be influencing intentionality judgments. Some have argued that the intentionality asymmetry is due to a biasing effect where people are inappropriately over-extending the concept of intentionality (Alicke, in press; Nadelhoffer, 2004). Although these accounts differ on the exact mechanism driving the effect, they generally posit that the asymmetry is due to the biasing effect of the negative outcome, in which people inappropriately attribute intentionality.

An alternative explanation has been championed by Joshua Knobe, who claims that “issues of praise and blame” are central to lay concepts of intentional action (Knobe, 2003b). On this account, intentional attributions are fundamentally tied to issues of praise and blame, so moral considerations should be influencing participant’s intentionality judgments. It follows then that because the way we think about praise and blame is not symmetrical (as discussed earlier), that our intentionality judgments should not follow a symmetrical pattern either. Importantly, Knobe’s account departs from all other models of intentionality in that it requires us to eschew the idea that intentionality can be thought of as a purely mentalistic evaluation of someone’s behavior. In short, when it comes to intentional attribution, moral considerations matter in an important and appropriate way.

This prescriptive side to the debate—whether intentional attribution should or should not be influenced by the moral outcome of an act—is to some extent a glass half-empty/half-full argument, and it is not entirely clear that empirical studies can conclusively resolve this debate (see Machery, in press). At a more descriptive level, however, researchers have questioned whether moral considerations are truly driving the asymmetry in intentionality judgments (Turner, 2004; Machery, in press). These alternative, non-moral accounts have instead suggested something akin to the notion that judgments of intentional action may track any behavior that violates some kind of norm, moral or otherwise. Norms, as used in this paper, can be thought of as any plausible reason that might constrain one’s actions. So according to these accounts, intentional attribution is likely to arise whenever there are reasons for not performing a given action. For simplicity, I will refer to these arguments as *norm-violation accounts*.

One such account, put forth by Machery (in press) is the *trade-off hypothesis*. This hypothesis argues that if an agent is perceived as making a trade-off (i.e., a decision that requires one to incur a cost if one is to reap a benefit) then participants will report that the cost was intentionally incurred in order to reap the benefit. Applying this explanation to the original Knobe example, participants report that the chairman intentionally harmed the environment because the cost of harming the environment is traded-off for the gain of making a profit. Conversely, the chairman is not seen as intentionally helping the environment because no cost is being incurred (i.e. a tradeoff has not taken place).

As a non-moral account of the side-effect effect, the trade-off hypothesis argues that this should occur for both moral and non-moral costs alike<sup>3</sup>. To test this second claim,

---

<sup>3</sup>A similar explanation has been offered by (Turner, 2004). Turner proposes that people will attribute intentionality whenever they believe an agent performed an action that also entails any reasons that “*the agent takes* to be a reason not to do it.” (Turner, 2004). To couch this in terms familiar to the trade-off hypothesis, the participant must judge that the agent would believe the side-effect to be a cost in order for them to have intentionally brought about that side-effect. This stipulation can be contrasted with that where only the participant need believe that the action in some way incurred a cost, but does not need to believe that the agent feels similarly. Machery, on the other hand, explicitly states that he is notcommittal

Machery (in press) devised a scenario that either required or did not require an agent to make a non-moral tradeoff. In one version, the agent desires to get a smoothie but before doing so is told that the action will incur an added cost—she will have to pay a dollar more than previously expected (i.e. a trade-off decision). In the other version, the agent desires a smoothie but is also told that doing so will get one a free commemorative cup (a non-trade off decision). In both cases, the agent doesn't care about this new information and continues to purchase the smoothie anyway. Machery found a large asymmetry in participants' intentionality judgments, where most participants reported that the agent intentionally paid an extra dollar but did not intentionally obtain the free cup. Based on these findings, Machery concluded that “the Knobe effect does nothing to show that the folk concept of intentional action has been shaped by its role ... in ascribing blame and praise” (p. 34).

### *The Current Studies*

The trade-off hypothesis (as well as other norm-violation accounts) provides for a parsimonious explanation of the differences in intentionality judgments across these scenarios, but does it provide for a complete explanation? Although it may be that a similar intentionality asymmetry can occur for situations with morally-neutral consequences, it does not disprove that moral judgments may also be influencing intentionality judgments in a similar fashion. Norm-violation accounts are not logically incompatible with moral-based accounts of the side-effect effect. However, none of the past studies are in a position to determine if intentionality judgments are also influencing moral considerations, because they have typically conflated these two accounts (Turner, 2004; Machery, in press).

There is, however, an alternative approach that can be taken to resolve this question, and can be done by appealing to a rather mundane observation: moral values can vary substantially from person to person. Since people hold different (and sometimes conflicting) moral attitudes—attitudes such as abortion, preemptive war, or the death penalty—how would this influence intentional attribution of an agent who performs an action that is believed to be immoral by some, but not by others? According to a strict interpretation of the norm-violation accounts, judgments of intentional action should be unaffected by a participant's set of moral values. If intentionality judgments are merely tracking acts that involve making a trade-off (or violating some norm while performing an action), as such accounts argue, then the participant's moral values are simply beside the point. But if moral considerations are influencing and shaping judgments of intentional action, as Knobe and others have argued, then one would expect these judgments to vary as a function of the participant's own moral values. To this end, five studies were conducted to examine the role that differing moral values may play in explaining the variance in judgments of intentionality.

### Experiment 1: The Side-Effect Effect Revisited

Nearly everyone agrees that harming the environment is bad, *ceteris paribus*. But does harming the environment qualify, in some sense, as morally wrong? Certainly for some, but perhaps not for others. The approach taken for Experiment 1 was to examine how the

---

on this point (Machery, in press).

sharp qualitative distinctions in value bestowed by individuals on a specific resource, such as environmental welfare, would influence their judgments of another's individuals actions.

As discussed earlier, one non-moral account of the side-effect effect is that people will ascribe intentionality to an agent so long as she is perceived as having made a trade-off (Machery, in press). However, psychologists have noted that for some people there are values that are treated as absolute and inviolable—that making tradeoffs on these values is taboo (J. Baron & Spranca, 1997; Tetlock, Kristel, Elson, Green, & Lerner, 2000). These “protected” values provide the backbone to the deontological rules that constrain our actions: the principle of “do not harm innocent others” is taken seriously because of the deep moral value that we place on human life<sup>4</sup>. Those with a protected value for the environment, then, should be more likely than those without a protected value to view tradeoffs on the environment (even those made indirectly) as morally blameworthy. If intentional attributions are fundamentally shaped by concerns of “issues of praise and blame,” then this difference should also emerge in intentionality judgments vis-à-vis protected values. If intentionality attributions are not shaped by such concerns, but rather by the act of making a trade-off, then we should not expect to see differences in moral absolutists’ and non-absolutists’ intentionality judgments.

For Study 1 the following hypotheses were made:

1. Moral absolutists for the environment (i.e., those with a protected value) would show a greater asymmetry in intentionality judgments compared to non-absolutists: the difference between attributions of intentionality for the *Harm* and *Help* conditions should be greater for those who believe that making a trade-off on the environment is taboo, compared to those who do not.

2. Moral absolutist’ and non-absolutist’ judgments will diverge most in the *Harm* condition. Because these protected values are primarily concerned with the prevention of impermissible actions (i.e., harms), no predictions were made concerning for how moral absolutists’ and non-absolutists’ judgments would differ in the *Help* condition.

3. Moral judgments would track intentionality judgments. If differences in intentionality judgments are found between moral absolutists and non-absolutists, then moral judgments should also follow suit.

## *Method*

### *Participants*

Three hundred twenty-nine students (189 females and 5 participants not specifying gender) at the University of California, Irvine were recruited from undergraduate courses in psychology and economics.

### *Materials and Procedure*

The study was advertised as an online study interested in “everyday moral reasoning.” Students were offered course credit in return for their participation. A generic email was sent out to students that briefly explained the details of the study, and provided a link students

---

<sup>4</sup>Of course, other values also matter when it comes to following deontological rules. For example, many people also value following moral rules that they believe to be just, simply for the sake of following such rules.

could access if they chose to participate. It was explained that participant responses would remain anonymous, and that they could choose to skip any questions they felt uncomfortable answering (though few did so). After answering all questions, participants were asked to report their student ID (in order to assign course credit) and then were finally shown a debriefing page that explained the purpose of the study.

After reading the initial instructions, participants were randomly assigned to read one of two versions of the Knobe scenario (2003a) described earlier. Briefly, the scenario asks participants to judge a corporate executive who undertakes a business venture that will increase profits for his company but also harms or helps the environment (*Harm* and *Help* conditions, respectively).

The primary dependent variable was judgment of intentionality on behalf of the chairman (e.g., “Did the chairman intentionally harm/help the environment?”). Participants were presented with a dichotomous yes/no option to this question, and then asked to explain their judgment. Participants were also asked to judge the moral status of the decision on a 7-point scale (1: Very Immoral, 7: Very Moral).

We were also interested in examining the specific mental attributions that participants would ascribe to the chairman. The five mental states thought to constitute folk concepts of intentional action—awareness, skill, belief, desire, and intention—were assessed<sup>5</sup> (Malle & Knobe, 1997). Accordingly, all participants were provided with the following items on 7-point scales: (a) Awareness: “Was the chairman aware that his decision would harm/help the environment?” (b) Skill: “Did the chairman have the ability to harm/help the environment?” (c) Belief: “Did the chairman believe that going through with the program would harm/help the environment?” (d) Desire: “Did the chairman have a desire to harm/help the environment?” (e) Intention: “Was it the chairman’s intention to harm/help the environment?”

*Assessing protected values for the environment.* Similar to past research (Tanner & Medin, 2004; Bartels & Medin, 2007), protected values towards the environment were assessed by asking the following question:

How do you feel about protecting the environment?

- (a) People should only undertake this action if it leads to some benefits that are great enough.
- (b) People should do this no matter how small the benefits.
- (c) Not undertaking the action is acceptable if it saves people enough money.

Only participants who answered (b) were considered to have a protected value for the environment (henceforth simply denoted as PVs). Participants who reported a willingness to make tradeoffs on the environment by answering (a) or (c) were considered to not have a protected value for the environment (non-PVs).

Although protected values are thought to be qualitatively different from strong attitudes in driving behavior (Skitka, Bauman, & Sargis, 2005), by definition protected values

<sup>5</sup>Skill is not a distinct mental component per se, but rather a judgment of the persons ability to perform an action effectively. For the purposes of the current study, however, this distinction is not an important one, and will be glossed over for purposes of analytical convenience.

also require strong attitudes towards the target domain. Because we hypothesized that protected values would explain judgments over and above strong attitudes, we attempted to control for attitude strength by including only participants who exhibited pro-attitudes towards the environment.

After responding to all of the dependent measures, participants were presented with competing statements on a number of political issues, and asked to favor one statement over the other. Within this battery was a single face-valid item that was used to determine pro-attitudes towards the environment (e.g., “Stricter environmental laws and regulations cost too many jobs and hurt the economy” vs. “Stricter environmental laws and regulations are worth the cost”). Only participants who favored the latter statement were considered for analysis that used protected value status as a predictor variable<sup>6</sup>.

## *Results and Discussion*

### *General findings*

The original side-effect effect was replicated: When the chairmans actions had the negative side-effect of harming the environment, 75% of subjects judged the chairman as intentionally harming the environment. By contrast, when the chairman’s actions had the positive side-effect of helping the environment, only 11% of subjects judged the chairman as intentionally helping the environment ( $\chi^2(1) = 135.97, p < .0001$ ). Subjects also judged the chairmans decision as more immoral when making a profit also harmed the environment ( $M = 1.99$ ) than when making a profit also helped the environment ( $M = 3.37$ ) ( $t(298) = 9.52, p < .0001$ ).

### *The role of protected values on judgments*

Would the drastic difference in judgments between the *Harm* and *Help* conditions be qualified by a willingness or unwillingness to make tradeoffs on environmental resources? To determine this, PVs were compared to non-PV subjects. Of the total 329 subjects, 258 met the initial inclusion criteria of having a pro-environment preference. This group was comprised of 203 PVs and 55 non-PVs. Similar to past research on protected values (J. Baron & Spranca, 1997), female subjects were more likely to have a protected value (76%) than were male subjects (54%) ( $\chi^2(1) = 15.33, p < .0001$ ).

A logistic regression was performed to test for the independent effects of protected value status (PVs vs. non-PVs) and valence of side effect (*Harm* vs. *Help* conditions) on judgments of intentionality. A reliable main effect was found for side-effect condition but not for protected value status ( $B = -9.26, p < .001$  &  $B = .23, ns$ , respectively). The main effect for condition reflects the basic tendency for subjects to ascribe intentionality more often in the *Harm* condition than in the *Help* condition (i.e., the original side-effect effect).

A second logistic regression was then performed that included the PV status  $\times$  condition interaction term to the model (Cohen, Cohen, West, & Aiken, 2003). As predicted, a

---

<sup>6</sup>It would seem odd that one would even need to exclude participants that had both a PV for the environment and who also showed a lack of support for the environment, since these two beliefs are incompatible. But this was indeed the case for a small number of subjects ( $N = 31$ ). Doing so points to an added benefit of our exclusion criteria—there is likely to be some degree of social desirability bias inherent to the PV item, and the exclusion criteria is a partial step towards sifting out “insincere” PVs from more genuine responses.

Table 1: Percentages (frequencies) of ascription’s of intentionality to the chairman in Experiments 1 &amp; 2

	Harm condition	Help condition	$\chi^2$	p-value
Experiment 1 (all participants)	75% (127/170)	11% (43/170)	135.97	< .001
No PV for the environment	63% (17/27)	21% (6/28)	9.75	0.002
PV for the environment	80% (89/111)	7% (6/91)	108.69	< .001
Experiment 2 (all participants)	52% (77/147)	18% (29/159)	39.33	< .001
No PV for the economy	44% (8/18)	28% (5/18)	1.08	0.300
PV for the economy	75% (18/24)	16% (3/19)	14.88	< .001

reliable interaction effect emerged—PVs, compared to non-PVs, were more likely to judge the chairman as intentionally harming the environment, and less likely to judge the chairman as intentionally helping the environment ( $B = -2.85$ ,  $p < .01$ ). Approximately 63% of non-PVs believed the chairman intentionally harmed the environment, and this percentage increased to about 80% for PVs ( $\chi^2(1) = 3.61$ ,  $p = .057$ ). Differences were also found between PVs and non-PVs in the *Help* condition; about 21% of non-PVs believed the chairman intentionally helped the environment, whereas this dropped to about 7% for PVs ( $\chi^2(1) = 5.20$ ,  $p < .05$ ).

PVs and non-PVs also differed in their moral assessments of the chairmans actions, following a pattern consistent with subjects’ intentionality judgments. A  $2 \times 2$  (protected value status  $\times$  valence of side-effect condition) analysis of variance (ANOVA) was performed on moral judgments of the chairmans actions. Main effects for condition and PV status were both found to be reliable ( $F(1, 254) = 84.5$ ,  $p < .0001$ , &  $F(1, 254) = 7.87$ ,  $p < .01$ , respectively). The condition main effect reflects the general tendency for subjects to judge the chairman’s actions as more immoral in the *Harm* condition compared to the *Help* condition. The main effect for PV status reflects the general tendency for PVs to assign lower morality ratings than non-PVs, consistent with the prediction that PVs and non-PVs moral judgments would track their intentionality judgments. Just as PVs were more likely than non-PVs to ascribe intentionality in the *Harm* condition (intentionality here implies blame), PVs were also more likely than non-PVs to judge the act as more immoral. Because PVs were also less likely than non-PVs to ascribe intentionality in the *Help* condition (intentionality here implies praise), PVs were also less likely than non-PVs to judge the act as moral. No reliable PV status  $\times$  condition interaction effect was found ( $F(1, 254) = 1.32$ ,  $p = .25$ ).

#### *Judgments of the mental states supporting intentional attribution*

The results thus far suggest strong differences between PVs and non-PVs in their intentional attributions and moral assessments of the chairman’s behavior. How might the difference in moral considerations between PVs and non-PVs drive this effect? One possible explanation is that because of the deep sense of value placed on the environment by those with a protected value, witnessing a harm to the environment will cause these participants to

go into "blame validation" mode (Alicke, 1992). On this account, participants will engage in biased information processing that is skewed towards affirming blame, with PVs construing the chairman's "real intentions" in a manner more likely to imply culpability.

A second possible explanation may be that moral considerations play a central role in intentionality judgments, and simply factor into how we judge an act to be intentional or unintentional (Knobe, 2004). On this account, there shouldn't be noticeable differences between PVs and non-PVs in their attributions of specific mental states that support intentional attribution, because moral considerations influence intentionality judgments independently of the mental facts ascribed to the agent.

To this end, PVs and non-PVs were compared for their attributions of the chairman's awareness, belief, skill, desire, and intention towards harming/helping the environment were also examined. Separate  $2 \times 2$  (protected value status  $\times$  valence of side effect condition) ANOVAs were performed for each of the five items. A main effect for condition did exert a strong influence on attributions of awareness, belief, desire, and intention (see Table 2, all  $p$ 's  $< .0001$ ), but not for attributions of skill ( $F(1, 256) = 0, ns$ ). The main effect for condition reflected the general tendency of subjects to ascribe to the chairman greater degrees of awareness, belief, desire, and intention to harm the environment compared to analogous ascriptions of the chairman's mental states when his actions helped the environment. Although the valence of the side-effect (i.e. condition) exerted a strong influence on subjects' depiction of the chairman, PVs' and non-PVs' depictions were quite similar in many respects. PV status did not have a reliable influence on any of these mental state attributions, nor did the interaction terms, with the exception of the desire item. The desire item showed a reliable PV status  $\times$  condition interaction effect ( $F(1, 257) = 4.89, p < .05$ ). PVs rated the chairman as having a greater desire to harm the environment ( $M = 3.68$ ) than non-PVs ( $M = 2.7$ ) ( $t(52) = -3.6, p < .001$ ). By contrast, PVs and non-PVs subjects did not reliably differ in their attribution of the chairman's desire to help the environment ( $M$ 's: 1.63 vs. 1.57,  $t(41) = -.21, ns$ ).

To determine if the difference in PVs' and non-PVs' intentionality ascriptions to the chairman were due to their attributions of the chairman's desire to harm/help the environment, a mediation analysis was performed using attributions of desire as the mediator variable (R. M. Baron & Kenny, 1986). Because the independent variable being used to perform the mediation was an interaction term (PV status  $\times$  side-effect condition), all steps of the mediation also regressed the main effects of PV status and side-effect condition. As established earlier, there was a reliable association between the interaction term and judgments of intentionality ( $B = -2.85, p < .01$ ). Also established earlier was the reliable association between the interaction term and attributions of desire to harm the environment ( $F(1, 254) = -2.21, p < .05$ ). Importantly, when attributions of desire to harm were controlled for, the effect of the interaction term on judgments of intentionality was reduced ( $B = -2.39, p = .017$ ). A Sobel test demonstrated this partial mediation to be reliable ( $Z = -2.09, p < .05$ ).

To summarize, subjects with a protected value for the environment showed a larger asymmetry in their intentionality judgments than those without a protected value (hypothesis 1). More specifically, PVs were more likely than non-PVs to report that the chairman intentionally harmed the environment (hypothesis 2). Although no prediction was made for how judgments might differ in the *Help* condition, PVs were also less likely to report that

Table 2: Means (std. deviations) for Experiment 1

	Harm condition		Help condition	
	Non-PVs	PVs	Non-PVs	PVs
(1) How moral was the chairman's decision?	2.07 (.96)	1.78 (1.02)	3.93 (1.56)	3.22 (1.28)
(2) Was the chairman aware his decision would harm/help the environment?	6.56 (1.28)	6.78 (.81)	5.96 (1.26)	5.67 (1.71)
(3) Did the chairman have the ability to harm/help the environment?	6.19 (.96)	6.44 (.90)	6.25 (1.35)	6.37 (1.15)
(4) Did the chairman believe his actions would harm/help the environment?	5.69 (1.78)	6.35 (1.06)	4.43 (1.91)	4.21 (1.91)
(5) Did the chairman have a desire to harm/help the environment?	2.70 (1.17)	3.68 (1.60)	1.57 (1.26)	1.63 (1.12)
(6) Did the chairman have an intention to harm/help the environment?	3.22 (1.58)	3.95 (1.89)	1.75 (1.40)	1.69 (1.10)

**Notes:** Participants responded to all questions using 7-point scales, with the endpoints 1 (*Not at all*) and 7 (*Completely*). One exception was the morality item, which used the endpoints 1 (*Completely Immoral*) and 7 (*Completely Moral*). PVs are those with a protected value for the environment, Non-PVs are those without a protected value for the environment.

the chairman intentionally helped the environment. Differences were also found between PVs and non-PVs in their moral assessments of the chairman's actions, following a similar pattern to their intentionality judgments (hypothesis 3). PVs were more likely to judge the chairman as blameworthy when his actions also harmed the environment, and less likely to judge the chairman as praiseworthy when his actions also helped the environment.

The differences in intentionality judgments between PVs and non-PVs cannot be explained by the trade-off hypothesis. PVs and non-PVs were equally likely to say that the chairman was aware of the consequences his actions would have on environment, and believed his actions would in fact harm the environment (see Table 2). These results indirectly suggest that the chairman was perceived as having made a tradeoff by both PVs and non-PVs. PVs and non-PVs, however, differed in how morally wrong they saw the actions of the chairman, which followed a pattern similar to that of their intentionality judgments. Taken as a whole, these results support the original claim by Knobe (2003a) that moral considerations are influencing intentional attribution.

Moreover, most of the specific mental states imputed to the chairman—his awareness, skill, beliefs, and intentions in harming or helping the environment—did not differ reliably

for those with different values for the environment. Although attributions of desire (the one item that did show a reliable difference between PVs and non-PVs) did partially mediate judgments of intentionality, subjects values towards the environment continued to exert a substantial influence on such judgments even when attributions of desire were statistically controlled for. These last results suggest that people who may otherwise agree about an agents state of mind can nonetheless disagree about whether that agent performed an action intentionally or unintentionally, based on differences in moral values for the environment.

## Experiment 2: Side-effects and the Economy

Experiment 1 suggests that judgments of intentional attribution can be influenced by outcomes that involve a resource under protected value (i.e., a moral judgment). However, it is unclear to what extent these results are a function of the nature of specific protected values (i.e., the environment); protected values more generally; or the specific stimuli presented to participants. Asking participants—a majority of whom reported a pro-attitude towards the environment—to judge an enterprising chairman who callously disregards the environment’s well-being may be of little use when generalizing the question of how moral values come to shape intentionality judgments. Experiment 2 parallels Experiment 1 by examining another (albeit more rarefied) value that is often perceived as in direct conflict with environmental welfare—protecting jobs and the health of the economy.

### *Method*

#### *Participants*

Three hundred seven students (169 females and 2 participants not specifying gender) at the University of California, Irvine were recruited from undergraduate courses in psychology and economics.

#### *Materials and Procedure*

Like in Experiment 1, participants were recruited to complete an online study in return for course credit. Participants were randomly assigned to one of two versions of a hypothetical scenario virtually identical to that used in Experiment 1, but this time about an environmentalist who aims to preserve the environment but with the added consequence of either harming or helping the economy (see Appendix). In the scenario, the chairman of an environmental protection organization is told that he can go through with a decision that will preserve the environment, but that it will also lead to an increase or decrease in unemployment (*Harm* vs. *Help* conditions, respectively).

Participants were presented with the same set of dependent variables as in Experiment 1: They were asked to judge the intentionality of the chairman in harming or helping the economy, to explain why or why not the action was intentional, and to assess the moral status of the chairmans decision. Participants were also asked to make assessments on the same set of awareness, skill, belief, desire, and intention judgments as in Experiment 1. All items were now, of course, assessing harm/help to the economy (rather than harm/help to the environment as in Experiment 1).

*Assessing protected values for the economy.* This time we were interested in the role that protected values for the economy, rather than the environment, would play in intentionality judgments. The same method as in Experiment 1 was used to assess protected value, except the item now assessed value for the economy. Those who indicated a willingness to make trade-offs on economic growth were identified as not having a protected value for the economy (non-PVs), while those who indicated an unwillingness to make such tradeoffs were identified as having a protected value (PVs).

Like in Experiment 1, an effort was made to control for PVs' and non-PVs' attitudes towards the economy. The same criterion item was used as before, but this time including only those who reported pro-attitudes towards the economy. That is, only those who agreed with the statement, "Stricter environmental laws and regulations cost too many jobs and hurt the economy" were considered for any of the analysis that used protected value status as a predictor variable.

Study 2 had the same hypotheses as Study 1:

1. PVs would show a greater asymmetry in intentionality judgments than non-PVs.
2. Differences in PV and non-PV judgments will be especially pronounced in the Harm condition.
3. Moral judgments would track intentionality judgments.

## *Results and Discussion*

### *General findings*

A basic asymmetry in intentionality judgments did emerge, though not as dramatic as that found in Experiment 1. About half (52%) of subjects judged the chairman as intentionally harming the economy in the negative side effect condition. By contrast, only 18% judged the chairman as intentionally helping the economy in the positive side effect condition ( $\chi^2(1) = 39.33, p < .0001$ ). Although this 34% difference in intentionality attributions is impressive, it is nonetheless considerably smaller than the 64% difference found in Experiment 1. Might this discrepancy be partially explained by the subjects' value (or lack thereof) towards harming to the economy?

Moral judgments of the chairman's actions suggest that this may be the case. Subjects did judge the chairman's decision to preserve the environment while harming the economy as more immoral ( $M = 3.51$ ) than when preserving the environment also helped the economy ( $M = 4.6$ ) ( $t(303) = -6.86, p < .0001$ ). However, it is also worth noting that the mean scores for moral judgments in both conditions of Experiment 2 were higher than moral judgments in either condition of Experiment 1. That is, the negative side effect condition in Experiment 2 (chairman preserves environment but harms economy) was still viewed as more moral than the positive side effect condition in Experiment 1 (chairman makes profit but also helps environment). Although comparing results across two separate studies should be interpreted with caution, these results suggests that the subjects viewed the actions of the environmentalist chairman in Experiment 2 as less immoral than the actions of the enterprising chairman in Experiment 1.

*The role of protected values on judgments*

A basic asymmetry on intentionality judgments was found, but as noted earlier, the difference was considerably smaller than that found in Experiment 1. Subjects' moral evaluations suggest this dampening of the asymmetry effect may be due to the relative value subjects place on the economy vis-a-vis the environment. If this is the case, then one would expect that for those with strong moral objections to making tradeoffs on the economy (i.e. those with a protected value) might produce a large asymmetry effect similar to that found in Experiment 1.

To determine this, PVs were compared to non-PV subjects. As expected, considerably fewer subjects had a preference for the economy over the environment—only 79 of the 307 subjects in Experiment 2 reported a pro-economy preference on the criterion item. The remaining sample was comprised of 43 PVs and 36 non-PVs. Like in Study 1, female subjects were more likely to have a protected value (68%) than were male subjects (40%) ( $\chi^2(1) = 5.58, p < .05$ ).

A logistic regression was first performed using PV status (PVs vs. non-PVs) and valence of side-effect (*Harm* vs. *Help* conditions) as predictor variables on judgments of intentionality. A reliable main effect was found for side-effect condition but not PV status ( $B = -3.44, p = .001$  &  $B = 1.01, ns$ , respectively). The main effect for condition reflects the general tendency of subjects to ascribe intentionality more often when the chairmans actions harmed the economy compared to when it helped the economy.

A second logistic regression that also included the PV status  $\times$  condition interaction term was then performed. A marginally reliable interaction effect emerged—PVs showed a large asymmetry in their intentionality judgments compared to non-PVs ( $B = -1.93, p = .054$ ). Consistent with predictions, 44% of non-PVs believed the chairman intentionally harmed the economy, but this percentage jumped to 75% for PVs ( $\chi^2(1) = 4.07, p < .05$ ). About 28% of non-PVs judged the chairman as intentionally helping the economy, while this percentage dropped to 16% for PVs. This difference, however, did not reach statistical significance ( $\chi^2(1) = .78, ns$ ).

Moral assessments of the chairmans actions appeared to track subjects' intentionality judgments, albeit this pattern was not as pronounced as in Experiment 1. A  $2 \times 2$  (PV status  $\times$  valence of side effect condition) between subjects ANOVA was performed on moral judgments. A reliable main effect for condition was found, which reflected the general tendency to judge the chairman's actions as more immoral in the *Harm* condition compared to the *Help* condition ( $F(1, 78) = 7.45, p < .01$ ). No reliable effect was found for PV status ( $F(1, 78) = 1.69, p = .20$ ). Although a PV main effect was not found, planned post-hoc comparisons support the prediction that moral assessments would track intentionality judgments. Just as PVs and non-PVs reliably differed in their intentionality judgments for the *Harm* condition, so too did they for their moral assessments: PVs judged the chairmans actions as more immoral than did non-PVs when the economy was harmed (M's: 2.95 vs. 3.61,  $t(40) = 1.75, p < .10$ ). And just as PVs and non-PVs did not reliably differ in their intentionality judgments for the *Help* condition, neither did they for their moral assessments: PVs and non-PVs did not reliably differ in judging the chairman's actions as moral or immoral when the economy was helped ( $t(40) = .14, ns$ ). Lastly, no significant PV  $\times$  condition interaction was found ( $F(1, 78) = 1.32, p = .28$ ).

Table 3: Means (std. deviations) for Experiment 2

	Harm condition		Help condition	
	Non-PVs	PVs	Non-PVs	PVs
(1) How moral was the chairman's decision?	3.61 (1.33)	2.96 (1.08)	4.06 (1.26)	4 (1.15)
(2) Was the chairman aware his decision would harm/help the economy?	6 (1.03)	6.46 (1.35)	5.06 (1.55)	5.74 (1.28)
(3) Did the chairman have the ability to harm/help the economy?	5.89 (1.37)	5.88 (1.60)	5.67 (1.54)	5.47 (1.39)
(4) Did the chairman believe his actions would harm/help the economy?	5.17 (1.72)	5.88 (1.62)	4.44 (1.98)	4.74 (1.82)
(5) Did the chairman have a desire to harm/help the economy?	2.67 (1.71)	3.33 (1.61)	3 (2.03)	3.47 (2.04)
(6) Did the chairman have an intention to harm/help the economy?	2.72 (1.74)	2.92 (1.38)	3.28 (2.14)	3.16 (2.12)

**Notes:** Participants responded to all questions using 7-point scales, with the endpoints 1 (*Not at all*) and 7 (*Completely*). One exception was the morality item, which used the endpoints 1 (*Completely Immoral*) and 7 (*Completely Moral*). PVs are those with a protected value for the economy, Non-PVs are those without a protected value for the economy.

#### *Judgments of mental states supporting intentional attribution*

Finally, I also examined if PV and non-PVs subjects differed in their attributions of the chairman's awareness, belief, skill, desire, and intention towards harming/helping the economy. Did PVs construe the specific mental states of the chairman differently from non-PVs? Like in Study 1, separate ANOVA's were performed for each specific mental state attribution item. Very few effects emerged across these items. There was a reliable main effect of condition for the awareness and belief items ( $F(1, 78) = 7.76, p < .01$  &  $F(1, 78) = 5.33, p < .05$ , respectively). This main effect reflected the general tendency for subjects to rate the chairman as having a greater degree of awareness and belief that his actions would harm the economy compared to analogous ascriptions of the chairman when his actions helped the economy. For all items, protected value status produced no reliable differences in subjects' attributions, nor did the PV status  $\times$  condition interaction term (see table 2).

To summarize, it appears that having a protected value does influence our judgments when considering whether a known side-effect was intentional or unintentional. Consistent with the results of Experiment 1, those who had a protected value for the economy showed a large asymmetry in their intentionality judgments across the two scenarios; those without

a protected value for the economy did not (hypothesis 1). More specifically, PVs were more likely than non-PVs to judge the chairman as intentionally harming the economy (hypothesis 2), but they did not differ in their intentionality judgments when the economy was helped. PV and non-PV subjects also differed in their moral judgments of the chairman’s actions. Although moral judgments followed a similar pattern to PVs’ and non-PVs’ intentionality judgments (hypothesis 3), this difference was not as pronounced. And lastly, PV and non-PV subjects did not seem to differ in the specific mental ascriptions that support intentional attribution. That is, PVs and non-PVs appeared to depict the “real intentions” of the chairman in much the same manner.

These results suggest that moral considerations may play an important role in folk concepts of intentional agency. Echoing arguments by Knobe (2003a), people’s concept of intentional action may be both a folk psychological assessment and a moral assessment of another person’s behavior. And because moral assessments are at least partly based on the moral values we hold, these results suggest that people who may otherwise agree about an agent’s state of mind can nonetheless disagree about whether that agent performed an action intentionally or unintentionally, simply because of differences in their intrinsic value systems.

### Experiment 3: The Collateral Damage Side-Effect

Experiments 1 and 2 provided initial support for the hypothesis that strongly held, specific values play an important role in our judgments of intentional action. Experiment 3 attempted to see if a similar pattern would emerge when assessing a broader, more diffuse set of values. For this task, subjects’ political orientation was assessed. Political orientation is often thought to serve as a proxy for different sets of moral values (Lakoff, 2002; Haidt & Graham, in press) and past research has also shown that political liberals and conservatives diverge on moral judgments (Skitka & Tetlock, 1993; Tetlock et al., 2000).

Might there be a set of political issues that involve foreseen but undesired side-effects, and that liberals and conservatives would also disagree on for moral reasons? The issue of civilian collateral damage appears to meet these criteria. In brief, the issue of collateral damage is concerned with the question of whether it can be morally justified to carry out military attacks that will likely entail the death, albeit unintended, of innocent civilians. Proponents of such military action often justify their position on the grounds that: (a) these actions are necessary to effectively combat the enemy, which may potentially save many more future lives; and (b) the death of innocent civilians is never intended, but are rather the unfortunate consequences of such actions.

Translated into the realpolitik of current US military efforts in the Middle East, which has generally received support from conservatives and opposition from liberals, the issue of civilian collateral damage provides a more realistic and relevant context in which to test peoples intentional attributions of known side-effects.

### *Method*

#### *Participants*

Two hundred forty-seven students (168 females and 7 participants not specifying gender) at the University of California, Irvine participated in the study. Students were recruited

from an introductory psychology course to participate in an online study in exchange for course credit. Fourteen participants were dropped from any further analyses because the time spent completing the online survey was either too short ( $< 5$  minutes) or too long ( $> 60$  minutes).

### *Materials and Procedure*

As in the previous studies, subjects were recruited to participate in an online study on everyday moral reasoning. Participants were given one of two scenarios. In both scenarios, military leaders in Iraq initiated an action that had the foreseeable but undesired side effect of killing innocent civilians. Half of the participants received a version in which American troops attacked Iraqi insurgents, and the other half received a version in which Iraqi insurgents attacked American troops. The scenario described American [Iraqi insurgent] leaders deciding to carry out an attack to stop key leaders of the Iraqi insurgency [American military] in order to prevent future deaths of American troops [Iraqi insurgents]. It was stated in each scenario that while the decision-makers were aware of the possibility of innocent deaths, they reasoned that sometimes it is necessary to sacrifice innocent people for the sake of a greater good (in this case the saving of many future lives). It also specifically stated that the decision makers did not intend the death of any innocent civilians, they merely foresaw it as an unwanted consequence of their military actions.

Participants then responded to a set of dependent variable items virtually identical to those used in Experiments 1 and 2. Participants judged if the harm caused to civilians was intentional (e.g., “Did the military leaders intentionally harm innocent civilians?”), to explain why or why not, and how moral or immoral the action was. Participants were also asked to make assessments on the same set of awareness, skill, belief, desire, and intention items as had been used in the previous studies, this time in relation to the military leaders’ actions.

After responding to the scenario, participants were asked to rate their political orientation on a seven point scale (1: Very Liberal, 7: Very Conservative) with the added option of responding “Haven’t given it much thought” or “Completely unsure.”

The following hypothesis were made for Study 3:

1. Politically liberal and conservative participants would differ in their intentionality judgments depending on the civilian population that was harmed as a result of military action. More specifically, because liberals are generally opposed to the current use of military force in the Middle East, they will be more likely than conservative participants to judge the harm caused to Iraqi civilians as intentional, but less likely than to judge harm caused to American civilians as intentional.

2. Moral judgments would track intentionality judgments. If liberals are more likely than conservatives to judge an act as intentionally harming innocent civilians then they should also judge the act as more immoral than conservatives, and vice versa.

## *Results and Discussion*

### *General findings*

Overall, Iraqi insurgent action that had the byproduct of harming American civilians was judged as more intentional than American military action that had the byproduct of

harming Iraqi civilians. Approximately 55% of subjects reported that Iraqi insurgents had intentionally harmed innocent American civilians, but only about 33% of subjects reported that American leaders had intentionally harmed Iraqi civilians ( $\chi^2(1) = 10.99, p = .001$ ). Moral assessments of the military decision appeared to track intentionality judgments: the military actions of the Iraqi leaders were judged as less moral ( $M = 3.45$ ) than that of the American military ( $M = 3.83$ ) ( $t(231) = -2.07, p < .05$ ).

### *The role of political orientation on judgments*

Did intentionality judgments vary as a function of the subjects' political orientation? Because only politically motivated subjects were desired for the current study, 41 subjects who reported "Completely unsure" or "Havent given it much thought" to the political orientation item were excluded from the analysis.

A logistic regression was performed using political orientation (very liberal–very conservative) and civilian casualty population (American civilians vs. Iraqi civilians conditions) as predictor variables on judgments of intentional harm<sup>7</sup>. A reliable main effect was found for civilian casualty population but not for political orientation ( $B = -2.6, p < .01$  &  $B = -.95, ns$ , respectively). The condition main effect reflected the general tendency for subjects to judge the harm caused to American civilians as more intentional than the harm caused to Iraqi civilians.

A second logistic regression that included the political orientation  $\times$  condition interaction term in the model was then performed. As predicted, a reliable interaction effect emerged ( $B = -2.11, p < .05$ ). The predicted regression lines for each condition are shown in Figure 1 as a function of political orientation. As can be seen in Figure 1, the more liberal the participant was, the more likely she was to judge the American leaders as intentionally harming innocent Iraqi civilians. By contrast, there was an opposite trend in the American civilians condition; the more conservative the participant was, the more likely she was to judge Iraqi leaders as intentionally harming innocent American civilians.

To assess moral judgments, a linear regression was then performed using the same predictor variables. A marginally reliable main effect was found for civilian casualty population condition, and a reliable main effect was found for political orientation ( $F(1, 189) = 1.9, p = .06$  &  $F(1, 189) = 3.05, p < .01$ ). The marginally reliable condition main effect reflected the tendency of subjects to judge the harm caused to American civilians as more immoral than the harm caused to Iraqi civilians. The political orientation main effect reflected the general tendency for conservative subjects to judge military action as less immoral compared to liberal subjects. When the political orientation  $\times$  civilian casualty condition interaction was entered into the model, however, a reliable interaction effect emerged ( $F(1, 188) = 2.89, p < .01$ ). Moral assessments followed a similar pattern to liberals' and conservatives' intentionality judgments—liberals were more likely than conservatives to judge harm to Iraqi civilians as immoral, and conservatives were more likely than liberals to judge harm to American civilians as immoral.

<sup>7</sup>Political orientation was centered on the mean response (3.65), and the interaction term was then created using the re-centered political orientation variable. This was done to remove non-essential multicollinearity from the model (Aiken & West, 1991). A similar procedure was done for all other logistic and linear regressions discussed in this paper.

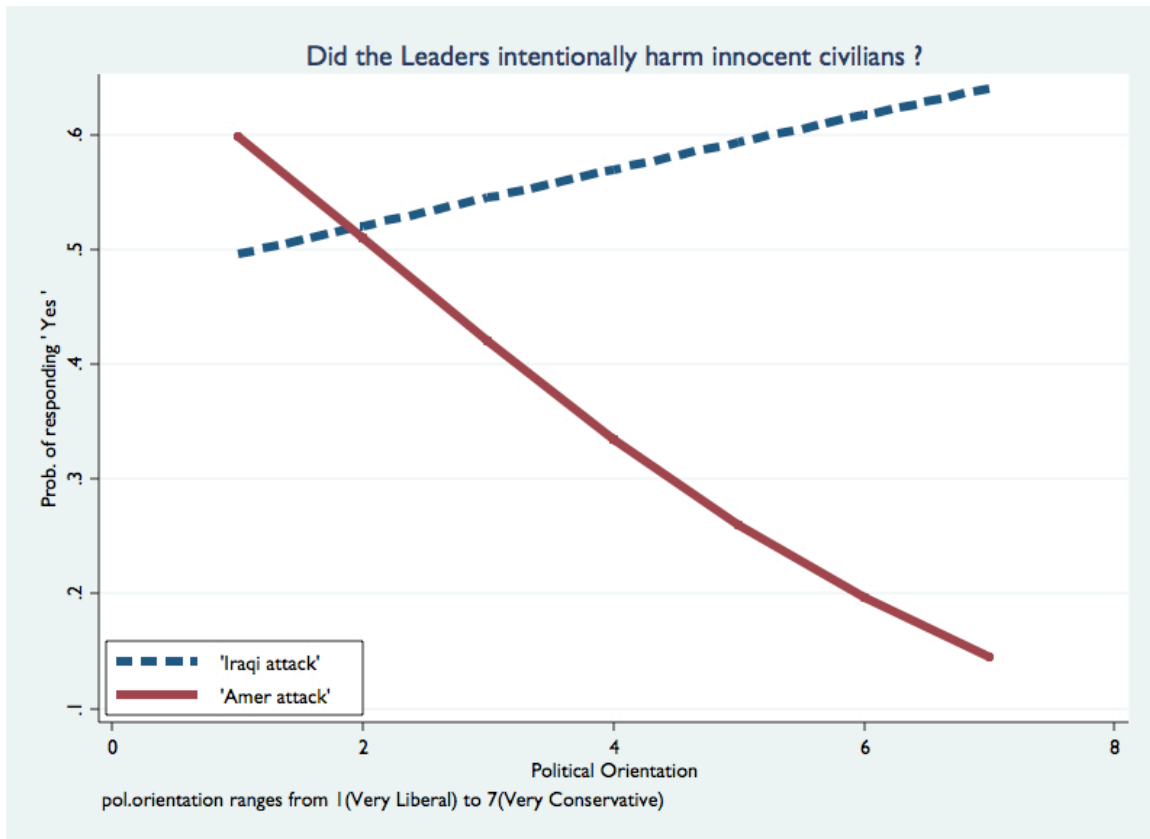


Figure 1. The predicted regression lines for each condition of Experiment 3, as a function of political orientation. The y-axis represents the predicted probability of responding "yes" to the question "Did the military leaders intentionally harm innocent civilians?". On the x-axis is political orientation, ranging from 1 (Very Liberal) to 7 (Very Conservative). The solid line represents the *Americans attack* condition, and the dotted line represents *Iraqis attack* condition.

#### *Judgments of the mental states supporting intentional attribution*

Attributions of the specific mental states that support intentionality judgments were also assessed as a function of political orientation and condition. For each of these items (i.e., awareness, belief, skill, desire, and intention) separate linear regressions were first performed using political orientation and civilian casualty condition as predictor variables. A main effect for political orientation was only found for the skill item ( $F(1, 187) = -2.76$ ,  $p < .01$ ). The political orientation main effect reflected the general tendency for more liberal subjects to ascribe a greater ability to harm innocent civilians compared to more conservative subjects. Liberals and conservatives did not, however, differ in their general tendency to ascribe awareness, belief, desire, or intention to harm innocent civilians. A main effect was found for the civilian casualty condition only for the desire and intention items ( $F(1, 189) = -4.16$ ,  $p < .001$  &  $F(1, 188) = -3.60$ ,  $p < .001$ , respectively). The condition main effect reflected the general tendency for subjects to ascribe a greater desire

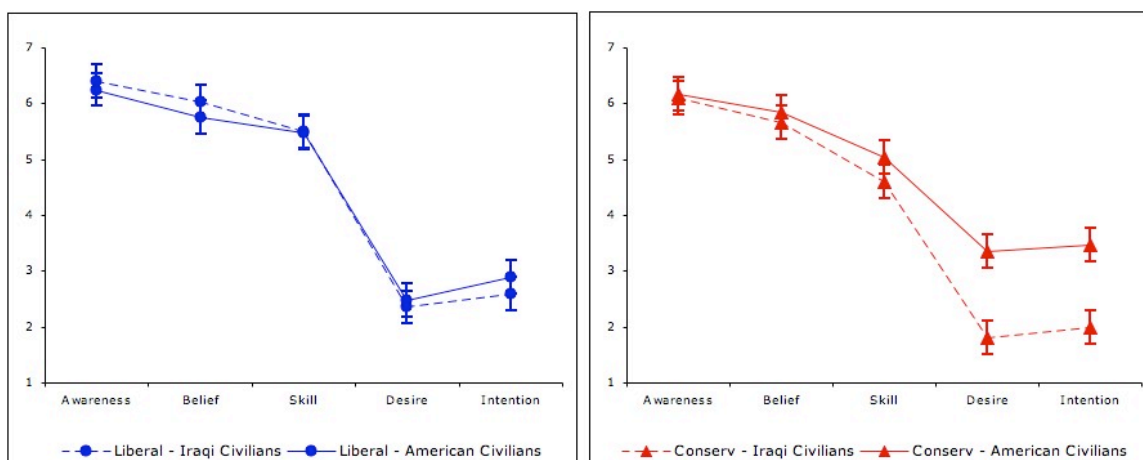
and intention to harm innocent American civilians (on behalf of Iraqi insurgent leaders) compared to innocent Iraqi civilians (on behalf of American military forces).

For each of these five items, separate linear regressions were then performed that also included the political orientation  $\times$  condition interaction term in the model (Aiken & West, 1991). No reliable interaction effect was found for attributions of awareness, skill, or belief. A reliable interaction effect emerged, however, for both attributions of desire and intention to harm innocent civilians ( $F(3, 187) = -2.09, p < .05$  &  $F(3, 188) = -2.60, p = .01$ , respectively). The interaction effect for both the desire and intention items followed a similar pattern to liberals' and conservatives' intentionality judgments: liberals were more likely than conservatives to ascribe American military leaders with a greater desire and intention to harm innocent Iraqi civilians; in contrast, conservatives were more likely than liberals to ascribe the Iraqi insurgent leaders with a greater desire and intention to harm American civilians.

These last results suggest that liberals' and conservatives' had different depictions of the "true intentions" of the Iraqi and American leaders' actions. To test this, the desire and intention items were collapsed to create a general index of *motive to harm* (Cronbach's  $\alpha = .89$ ), and then a mediation analysis was performed on intentionality judgments using this index as the mediating variable (Baron & Kenny, 1986). Because the independent variable being used to perform the mediation was an interaction term (political orientation  $\times$  civilian casualty condition), all steps of the mediation also regressed the main effects of political orientation and condition. First, a logistic regression reliably established a link between the interaction term and judgments of intentionality ( $B = -2.11, p < .05$ ). Second, a linear regression established a reliable link between the interaction term and motive to harm ( $F(1, 187) = -2.43, p = .016$ ). Lastly, when judgments of motive to harm were controlled for, the effect of the interaction term on judgments of intentionality was dramatically reduced ( $B = -1.05, p = .30$ ). A Sobel test confirmed this mediation to be reliable ( $Z = -2.01, p < .05$ ).

To summarize, the results of Experiment 3 show a similar side-effect asymmetry to that found in Experiments 1 and 2, where intentionality judgments are being influenced by moral considerations. Liberals and conservatives differed in the degree to which they found the harming of civilian casualties to be intentional and immoral, depending upon who was being harmed. Liberals judged the collateral damage of Iraqi civilians as both more intentional and immoral compared to conservative subjects (hypothesis 1 and 2). On the other hand, conservatives judged the collateral damage of American civilians as more intentional and immoral than did liberals.

Unlike Experiments 1 and 2, the asymmetry in intentionality judgments between liberals and conservatives was fully mediated by the desires and intentions that were imputed to the agent of harm. As shown in Figure 2, conservatives inferred more malevolent intentions than did liberals to Iraqi insurgent leaders (whose actions harmed American civilians) and more benevolent intentions to American military leaders (whose actions harmed Iraqi civilians). When these inferences were statistically controlled for, the difference between liberals' and conservatives' intentionality judgments largely disappeared. It is uncertain, therefore, whether one can determine if moral considerations influenced subjects' judgments of intentionality, or if liberals and conservatives simply came into the experiment with pre-existing assumptions about American and Iraqi military forces that biased their judgments.



*Figure 2.* Participants attributions in Experiment 3 of the Military leaders' mental states that support intentionality judgments. These scores represent the predicted responses of liberals (blue markers) and conservatives (red markers) 1 std. deviation above and below the mean response on political orientation. Scenario conditions are denoted by either dotted (American civilians) or straight (Iraqi civilians) lines. The y-axis represents degree to which participants imputed the given mental state to the military leaders, ranging from 1 (Not at all) to 7 (Completely). Notably, conservatives' attributions across scenarios are considerably more variable than liberals, especially for attributions of desire and intention to harm innocent civilians.

Although the latter result would be an interesting finding in its own right, it is unclear that such a finding provides support for the notion that moral considerations are playing an important role in lay concepts of intentional attribution. Experiment 4 was designed to clarify this point.

#### Experiment 4: Civilian Casualties Revisited

In Experiment 3, the two scenarios differed in their side effects, but the agents of harm (American forces or Iraqi insurgents) also differed across scenarios. Although this was done to add a degree of realism to the scenarios, it also introduced a confound that makes interpretation of the results problematic. As previously discussed, conservatives may have simply inferred more malevolent intentions on behalf of the insurgent leaders. Experiment 4 attempted to control for this confound by holding constant the agent of harm, while varying the outcome of the side effect.

#### *Method*

##### *Participants*

238 students (170 females) at the University of California, Irvine participated in the study for course credit.

### *Materials and Procedure*

The study was part of a larger questionnaire given to a class of psychology students. As in Experiment 3, participants were given one of two scenarios. In both scenarios military leaders initiated an action that had the foreseeable but unintended side effect of killing innocent civilians. Unlike Experiment 3, however, both scenarios had American military forces carrying out the attacks. This time American forces were trying to attack a key Al Qaeda operative who was (depending upon the condition) hiding in a small rural mountain region of either Afghanistan or Idaho. Both scenarios stated that an air strike was necessary for tactical reasons, but that the attack would also likely incur Afghan [American] collateral casualties. In both cases, it was also stated that American forces did not intend the death of any innocent civilians, but that the attack was necessary in order to prevent a larger number of future American civilian casualties.

After reading the scenario, participants completed the same three basic items that were asked in Experiment 3—did the American forces intentionally harm innocent civilians, to explain if the harm caused to innocent civilians was intentional or unintentional, and to rate the moral status of the American forces actions. Because of time restrictions, participants were not asked to make assessments on the specific mental states related to intentionality (i.e., awareness, skill, belief, desire, intention).

At the end of the questionnaire participants were also asked to rate their political orientation on a 7-point scale, with the added options of “Completely unsure” or “Haven’t given it much thought.”

Study 4 made the similar hypotheses to those of Study 3:

1. Politically liberal and conservative participants would differ in their intentionality judgments depending on the civilian population that was harmed as a result of military action. Liberals will be more likely than conservative participants to judge the harm caused to Afghan civilians as intentional, but less likely than to judge harm caused to American civilians as intentional.

2. Moral judgments would track intentionality judgments. If liberals are more likely than conservatives to judge an act as intentionally harming innocent civilians then they should also judge the act as more immoral than conservatives, and vice versa.

### *Results and Discussion*

#### *General findings*

Overall, no reliable differences were found for subjects’ intentionality judgments across the two scenarios ( $\chi^2(1) = .15, p = .70$ ). For both conditions, about two-thirds of the subjects responded that military leaders did *not* intentionally harm innocent civilians (69% and 66% for American civilians and Afghan civilians scenarios, respectively). Similarly, moral assessments of the military’s action did not differ between scenarios (M’s: 3.95 vs. 3.96,  $t(233) = .03, ns$ ).

#### *The role of political orientation on judgments*

Although no reliable differences in either intentionality judgments or morality assessments emerged between conditions, would these judgments vary according to the partici-

pants political orientation? To determine this, I excluded 35 subjects (as in Experiment 3) who reported no real preference on political orientation.

A logistic regression was performed using political orientation (very liberal – very conservative) and civilian casualty population (*American civilians* vs. *Afghan civilians* conditions) as predictor variables on intentionality judgments. As before, no reliable main effect was found for either political orientation or civilian casualty condition ( $B = .86$ ,  $ns$  &  $B = -.16$ ,  $ns$ ).

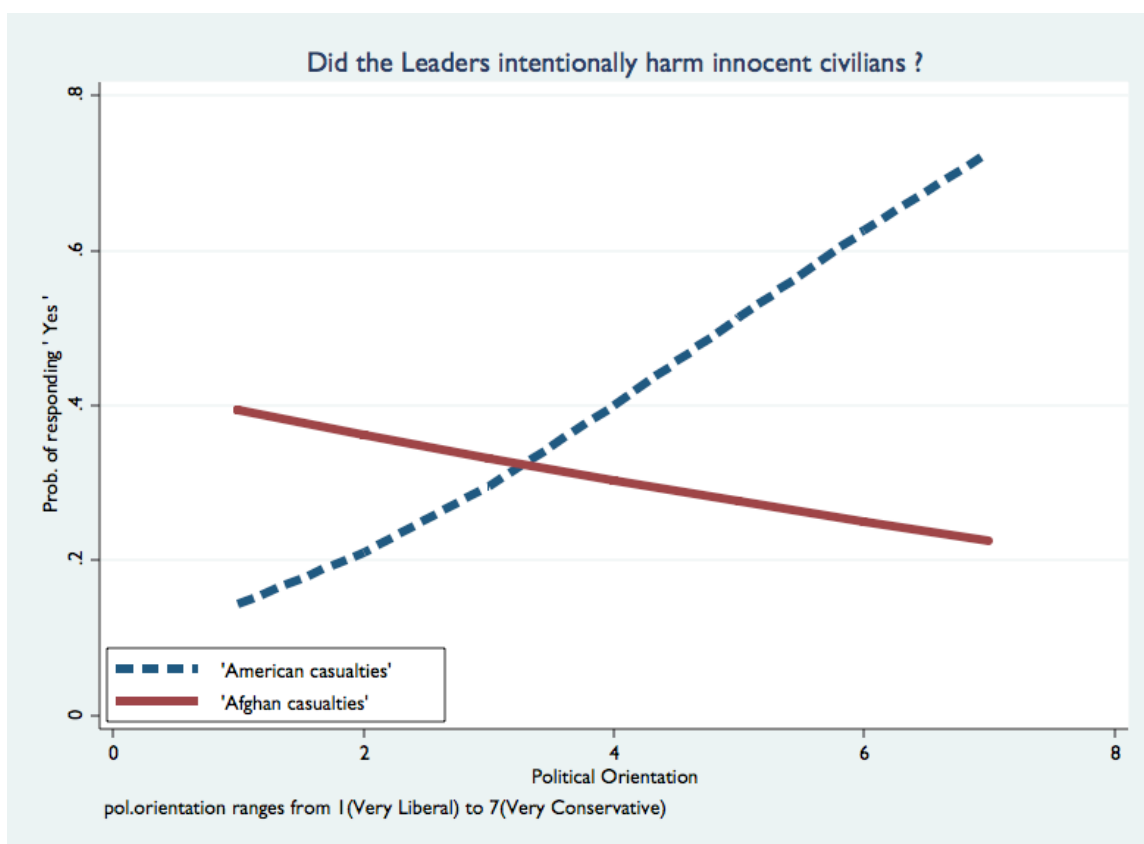
However, when a secondary logistic regression was performed that included the civilian casualty condition  $\times$  political orientation interaction term, a reliable interaction effect emerged ( $B = -2.58$ ,  $p = .01$ ). The interaction followed a pattern similar to that found in Experiment 3. As shown in Figure 3, the more liberal the participant was, the more likely she was to judge the American leaders as intentionally harming Afghan civilians. By contrast, there was an opposite trend in the American civilians condition; the more conservative the participant was, the more likely she was to judge American leaders as intentionally harming innocent American civilians.

A similar analysis was also carried out for morality assessments. A linear regression found a reliable main effect for political orientation but not for civilian casualty population ( $F(1, 199) = 3.89$ ,  $p < .001$  &  $F(1, 199) = -.29$ ,  $ns$ , respectively). The political orientation main effect reflected the general tendency for more conservative subjects to judge American military actions as less immoral than more liberal subjects. A linear regression was then performed that also included the political orientation  $\times$  civilian casualty population interaction term. Contrary to predictions, moral assessments in Experiment 4 did not appear to track intentionality judgments—no reliable interaction effect emerged for moral assessments of military action ( $F(3, 201) = .87$ ,  $p = .38$ ).

This last result seems somewhat puzzling. Liberals' and conservatives' moral assessments and intentionality judgments were consistent when Afghan civilians were harmed, but inconsistent when American civilians were harmed. As was expected, conservatives were more likely than liberals to judge American forces' actions as moral when Afghan civilians were harmed ( $r = .33$ ,  $p < .001$ ), consistent with their intentionality judgments. But conservatives' were also (unexpectedly) more likely than liberals to judge American forces' actions as moral when American civilians were being harmed ( $r = .18$ ,  $p < .10$ ), inconsistent with their intentionality judgments (conservatives viewed the harm to American civilians as more intentional than liberals). Can these results be reconciled with the findings in Experiment 3?

A recent set of studies by Pizarro and his colleagues might provide the answer (Knobe, in press; Pizarro, Knobe, & Bloom, unpublished manuscript). Pizarro et al. had participants judge a film director whose music video had the known side-effect of encouraging either homosexual or heterosexual couples to French-kiss in public. Although participants' explicit moral judgments did not differ between the two outcomes, participants were more likely to judge the director as intentionally encouraging gay kissing compared to encouraging straight kissing. Pizarro et al. suggested that these findings may be due to a divergence between implicit and explicit moral judgments (disapproval/approval of gay kissing in public), with intentionality judgments tracking only implicit moral judgments.

The findings reported here may be consistent with Pizarro et al.'s account. In the American casualties condition, conservatives may have implicitly disapproved of the harm



*Figure 3.* The predicted regression lines for each condition of study 4, as a function of political orientation. The y-axis represents the predicted probability of responding “yes” to the question “Did the American military leaders intentionally harm innocent civilians?”. On the x-axis is political orientation, ranging from 1 (Very Liberal) to 7 (Very Conservative). The solid line represents the *Afghan civilians* condition, and the dotted line represents the *American civilians* condition.

caused to American civilians relative to liberals, which would explain why conservatives judged the act to be more intentional than did liberals. However, there may have been reasons why conservatives' explicit responses on moral judgments would not follow a similar pattern. For example, conservatives may have felt unease at the prospect of reporting moral disapproval of American military efforts. In the Afghan civilians condition, on the other hand, liberals' and conservatives' implicit judgments would presumably be more congruent with their explicit judgments.

Why? It appears that liberals' explicit moral judgments followed a pattern similar to their intentionality judgments and thus were more negative than conservatives' moral judgments. After all, liberals judged the harm caused to Afghan civilians as more intentional than did conservatives, and liberals probably do not have reservations about reporting their disapproval of American military use of force. However, the above interpretation should be treated as a tentative hypothesis that undoubtedly requires further empirical validation.

### Experiment 5: Is Harm Reduction also an issue about side-effects?

Experiments 1–4 support the general hypothesis that differences in participants' values influence their judgments of intentional action. Experiment 5 examines the implication of this hypothesis in a heretofore unexplored context in which the side-effect effect might arise—*dilemmas of harm reduction*. Dilemmas of harm reduction are concerned with the question of whether one should provide assistance to reduce harms caused by certain behaviors, even if one disagrees with those behaviors. Examples of harm reduction initiatives might include needle exchange programs for heroin addicts, legalization of prostitution, methadone clinics for substance abusers, or distributing condoms to high school students.

While such initiatives clearly aim to reduce existing harms, critics argue that such programs also have the side effect of implicitly condoning (and even promoting) the undesired behavior. Unlike Studies 1–4, where the direct side-effects of an agent's actions were evaluated, the negative side-effect of harm reduction is indirect in that the causal link is one step removed—the act helps to facilitate *another's* bad behavior. However, because debates on harm reduction initiatives are often value-laden in nature, and seem to center on the distinction between intended ends versus foreseen byproducts, such disputes may nevertheless be fertile grounds for testing how proponents and opponents would judge the intentional character of policy decisions concerned with harm reduction.

#### *Method*

##### *Participants*

179 students (144 females) at the University of California, Irvine participated in the study for course credit.

##### *Procedure and Materials*

The study was part of a larger questionnaire given to a class of psychology students to complete. Participants were given one of two scenarios dealing with the dilemma of harm reduction: should one reduce an existing harm even if it tacitly condones an act one disagrees with? Half of the participants read the *High School* version, which discussed a decision by California policy makers to distribute condoms in public schools. It was stated that premarital sex was widespread among students at this age, and the proposed measure may help to reduce the spread of sexually transmitted diseases among teenagers. The other half of the participants received the *Military* version, which discussed a decision by legislators to provide condoms to military personnel overseas. It was stated that sexual aggression, often in the form of rape, was widespread among soldiers serving military duty in foreign countries, and the proposed measure may help to reduce the spread of sexually transmitted disease to the victims. In both versions, it was explicitly stated that the policy makers did not condone the act of premarital sex/rape, but that these behaviors would probably continue to persist “no matter what” and that the decision was instead to simply reduce existing harms.

Participants were then asked whether the decision-makers were intentionally promoting premarital sex/rape, to agree or disagree with the decision, and to determine the moral status of the decision. All items were administered on 7-point likert scales (1: Completely

Unintentional/Disagree/Immoral, 7: Completely Intentional/Agree/ Moral). Lastly, participants were also asked to report some basic demographic information at the end of the questionnaire, including gender, race/ethnicity, and political orientation.

Based upon the findings of Studies 1–4 that moral considerations do influence judgments of intentionality, I hypothesized that:

1. Participants would find the policy decision to distribute condoms to military soldiers as more immoral than to distribute condoms to high school students.
2. Participants would be more likely to judge policy makers as intentionally promoting rape in the *Military* condition than to judge policy makers as intentionally promoting premarital sex in the *High School* condition.

### *Results and Discussion*

As expected, the tendency to judge military decision-makers as intentionally promoting rape was greater ( $M = 4.1$ ) than were judgments of policy makers as intentionally promoting premarital sex ( $M = 2.5$ ) ( $t(170) = -6.7, p < .001$ ). Participants were also more likely to disagree with the policy decision to distribute condoms to military soldiers ( $M = 3.6$ ) than to distribute condoms to high school students ( $M = 5.4$ ) ( $t(171) = 6.7, p < .001$ ). Lastly, participants were more likely to judge distributing condoms to soldiers as immoral ( $M = 3.3$ ) compared to high school students ( $M = 4.8$ ) ( $t(177) = 6.3, p < .001$ ).

Interestingly, the above results are qualified by an unexpected gender  $\times$  condition interaction effect, with females reacting more positively and negatively to the high school and military conditions, respectively. A  $2 \times 2$  (gender  $\times$  condition) between subjects ANOVA was performed for each of the three dependent variables. For judgments of intentionally promoting premarital sex/rape, a main effect was found for condition ( $F(1, 178) = 14.45, p < .001$ ) but not for gender ( $F(1, 178) = .02, ns$ ). However, these judgments were qualified by a reliable gender  $\times$  condition interaction ( $F(1,179) = 6.4, p < .05$ ). As depicted in Figure 4, men did not differ significantly in their intentionality judgments across the two scenarios ( $t(32) = -.6, ns$ ), but women judged the decision makers as much more intentionally promoting rape ( $M = 4.25$ ) than intentionally promoting premarital sex ( $M = 2.36$ ) ( $t(132) = -7.6, p < .0001$ ).

The agree/disagree and moral judgment items followed the same pattern as intentionality judgments. For the agree/disagree item, there was a main effect for condition ( $F(1, 178) = 18.52, p < .0001$ ) but not for gender ( $F(1, 178) = .64, ns$ ). These results were qualified by a marginally reliable condition  $\times$  gender interaction ( $F(1, 178) = 2.78, p < .10$ ). Post-hoc comparisons reveal that men were more likely to agree with the decision to distribute condoms to high school students ( $M = 5.13$ ) than to military soldiers, although this difference was only marginally reliable ( $M = 4.25$ ) ( $t(32) = 1.72, p < .10$ ). Women showed a similar but more reliable pattern; they were more likely to agree with the decision to distribute condoms to high school students ( $M = 5.42$ ) than to military soldiers ( $M = 3.42$ ) ( $t(134) = 6.6, p < .0001$ ).

For moral judgments, a main effect was found for condition ( $F(1, 178) = 13.29, p < .001$ ), but not for gender ( $F(1, 178) = 1.66, ns$ ). These results were qualified by a reliable condition  $\times$  gender interaction ( $F(1, 178) = 5.54, p = .02$ ). Men's moral judgments of the decision to distribute condoms did not differ reliably across conditions ( $M$ 's: 4.53 vs. 4.15,  $t(33) = .7, ns$ ), but women judged the decision to distribute condoms to high school

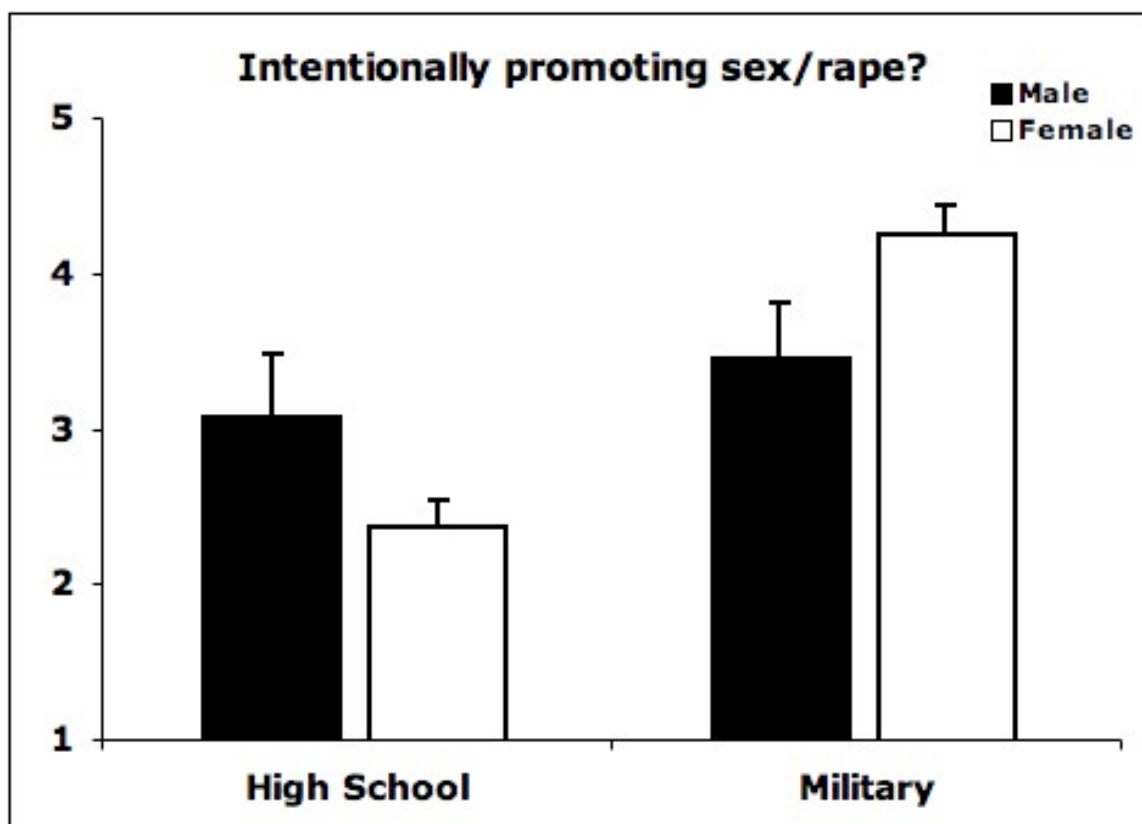


Figure 4. Average responses in Study 5 to the question “Did the decision-makers intentionally promote rape/ premarital sex?” as a function of condition and gender. Error bars represent standard error of the mean.

students as more moral ( $M = 4.84$ ) than distributing condoms to military soldiers ( $M = 3.07$ ) ( $t(142) = 6.88, p < .0001$ ).

Moreover, gender was not reliably associated with political orientation ( $r = .10, p = .17$ ), age ( $r = .04, p = .60$ ), or ethnicity ( $\chi^2(5) = 6.4, p = .27$ ). To investigate whether a gender  $\times$  condition interaction uniquely accounted for the variance of morality and intentionality judgments, the procedures described by Aiken and West (1991) for testing moderated regression were followed. The separate effects of age, ethnicity, and political orientation by condition interaction terms were entered in a standardized regression to predict both judgments of intentionality and morality (separate regressions were performed for each of these items). None of the interaction terms were found to be significant for intentionality judgments, so the regression analyses were repeated excluding these terms from the model (Aiken & West, 1991). However, a significant political orientation  $\times$  condition interaction was found for judgments of morality, so this interaction term was retained for the repeated regression analyses of the morality item.

A secondary regression analysis was then performed that regressed age, ethnicity, political orientation, gender, and condition before entering the gender  $\times$  condition interaction

Table 4: Standardized and unstandardized regression weights of predictors for judgments of morality and intentionality to distribute condoms to High School students/Military soldiers

	Moral/Immoral				Intentional/Unintentional			
	$\beta$	B	<i>SE</i>	<i>t</i>	$\beta$	B	<i>SE</i>	<i>t</i>
Age <sup>a</sup>	-.04	-.02	0.04	-.57	0.05	0.03	0.04	0.74
Gender	0.06	0.30	0.45	0.67	-.15	-.68	0.46	-1.49
Political Orientation <sup>a</sup>	-.10	-.13	0.13	-1.02	0.06	0.08	0.09	0.85
Ethnicity <sup>b</sup>								
African American	0.03	0.51	1.14	0.45	0.08	1.27	1.17	1.09
Latino/Hispanic	-.10	-.44	0.36	-1.21	0.03	0.14	0.37	0.39
Asian/Asian American	-.07	-.25	0.29	-.86	0.09	0.33	0.30	1.10
Middle Eastern	0.05	0.40	0.57	0.70	0.05	0.38	0.55	0.55
Other	-.09	-.79	0.61	-1.30	0.02	0.15	0.62	0.24
Condition	-.13	-.47	0.54	-.86	0.09	0.30	0.55	0.55
Political Orientation $\times$ Condition intx.	-.04	-.07	0.17	-.42	—	—	—	—
Gender $\times$ Condition intx.	-.36	-.13	0.61	-2.15*	.43	1.57	0.62	2.53**
Observations			179				179	
Adj <i>R</i> <sup>2</sup>			0.25				0.25	

**Notes:** Both DV items were assessed using 7-point scale with the endpoints 1 (*Completely Immoral/Unintentional*) and 7 (*Completely Moral/Intentional*).

<sup>a</sup>continuous variables were first re-centered on the mean score to reduce multicollinearity

<sup>b</sup>ethnicity was dummy-coded using Caucasian/White as the reference group

\*  $p < .05$

\*\*  $p < .01$

term. For both the intentionality and morality items, the gender  $\times$  condition interaction was virtually unaffected when controlling for these other variables in the model (see table 5). Given the limited amount of demographic information that was available from the sample, this moderated regression provided for as stringent a test as was possible. Nonetheless, gender continued to strongly moderate judgments of intentionality and morality across the two conditions.

In summary, the results of Experiment 5 are consistent with those found in Experiments 1–4: Dilemmas of harm reduction that involved morally bad outcomes (i.e., supplying condoms to soldiers who rape foreign civilians) were also viewed as more intentionally promoting the harmful act itself (i.e., rape) than when judging a similar action that involved a less immoral outcome (e.g., supplying condoms to high school students who engage in premarital sex). Surprisingly, these results appear to be moderated by the gender of the participant. Female participants' judgments showed a large asymmetry across the two scenarios. In contrast, male participants' judgments did not show reliable differences across the two conditions for either judgments.

These results are suggestive of the troubling implication that the male participants

may not have judged rape to be as morally wrong as did female participants. Although speculative, there may be real reasons why this would unfortunately be the case. Women are considerably more likely to be the victims of sexual aggression than are men, and because these outcomes are less consequential for men than for women, men may place less value on such outcomes as a result. Personal relevance, defined here to include outcomes which are more likely affect oneself in some direct manner (i.e., more consequential), may be the key mediating variable that explains the difference between men and women's judgments in Experiment 5. If this is correct, then an interesting future study could be to test these scenarios on male populations for whom rape is more personally relevant outcome, such as fathers who have teenage daughters, to see if such gender differences still arise.

Another explanation is that the differential gender judgments may be due to a more general willingness by male participants' to consider trade-offs when deliberating on the issue of harm reduction. Similar to past findings (J. Baron & Spranca, 1997), Experiments 1 and 2 in the current paper found that female participants were more likely than male participants to report an unwillingness to make tradeoffs (i.e., have a protected value) for both the environment and the economy. The results in Experiment 5 could simply reflect a more general tendency for males to see tradeoffs as more acceptable than females, regardless of the content of the trade-offs. This explanation, however, fails to account for the absence of main effects of gender for all of the dependent variables in Experiment 5. If males were generally less trade-off averse than females, then one would expect this tendency to manifest itself with greater endorsement of harm reduction policies across conditions. Indeed, women were more likely than men to endorse the decision of California policy makers to distribute condoms to high school students.

### General Discussion

The principal finding of the current research is that judgments of intentional action vary systematically as a function of the observer's moral values. When assessing an agent whose action produced a known side-effect, participants were more likely to say that the agent intentionally caused the side-effect if it violated a moral value. Participants were more likely to report that a chairman's decision led to an intentional harming of the environment (Experiment 1) or the economy (Experiment 2) if the decision jeopardized a protected value. Political partisans differed in their intentionality judgments of a military strike that knowingly harmed innocent civilians, depending on who was carrying out the attack or who was being harmed (Experiments 3 and 4). In general, political liberals were more likely than conservatives to view an act as intentional when foreign civilians were harmed, but the opposite trend was found when American civilians were harmed. Lastly, a policy decision that aimed to reduce overall harm but that also facilitated a morally abhorrent act was judged as more intentionally promoting that act, especially by women (Experiment 5).

Taken together, these findings suggest that moral considerations influence intentionality judgments over a range of actions that involve known but undesired side effects. The differences in intentionality judgments cannot be explained by appealing to strictly non-moral norm-violation explanations of the side-effect effect. The trade-off hypothesis (Machery, in press), for example, argues that an agent should be seen as intentionally bringing about the costs involved in making a trade-off (i.e., incurring a cost that is commensurate

with obtaining a benefit). This explanation fails to account for the variance across participants in all five studies. In Experiments 1 and 2, the trade-off hypothesis serves as a possible explanation for the general asymmetry found across *Harm* and *Help* conditions (i.e. the original side-effect effect), but cannot explain why moral absolutists—who viewed such trade-offs as taboo—were also more likely than non-absolutists to claim that the action was intentional. In Studies 3–5, the trade-off hypothesis would predict that intentionality judgments should not differ substantially, since all conditions required an agent to make a trade-off. This was not the case—participants’ intentionality judgments varied according to the nature of the tradeoff (i.e., who was harmed, who was the agent of harm) as well as the participant’s political orientation (Studies 3 and 4), or according to the participant’s gender (Study 5). Other norm-violation accounts (Turner, 2004) fail to explain the current findings for the same reasons that the trade-off hypothesis does. Moral considerations are clearly playing a role in judgments of intentionality.

Moral considerations may influence intentional attribution through multiple channels. One channel, suggested by the results of Experiment 1 and Experiment 3, is that morally blameworthy acts may actually distort participants’ depictions of an agent’s mental states in a manner that would support intentional attribution. In Experiment 1, those that placed a morally absolute value on the environment were more likely than non-absolutists to conclude that the chairman actually had a greater desire to harm the environment, even though it was made clear in the scenario that the chairman simply didn’t care, one way or the other, about the welfare of the environment. These imputations on behalf of the environmental absolutists partially explained how they arrived at their judgments of intentional action. In Experiment 3, conservatives came to very different conclusions about the “real intentions” of a military strike, depending on who was conducting the strike. Conservatives were more likely to conclude that Iraqi leaders really did have a “motive to harm” innocent American civilians, but gave American leaders the benefit of the doubt when making a similar assessment. Liberals, on the other hand, appeared to take the desires and intentions of both military strikes at face value (see Figure 2). Moreover, the difference in the motives that conservatives and liberals ascribed to the respective military strikes appeared to fully explain their intentionality judgments. That is, conservatives thought that Iraqi insurgent leaders had really bad intentions all along, so they would naturally assume that the act was indeed intentional.

This account makes intuitive sense and doesn’t require that one revise standard models of lay concepts of intentionality (Malle & Knobe, 1997). If we believe a person does indeed have bad intentions, then we may be justified in labeling their actions as intentional. However, this does not appear to be the only route in which moral considerations can influence intentional attribution. If perceptions of intentions and desires were the whole story, then how does one explain that even when these were statistically controlled for in Experiment 1 a large difference between absolutists’ and non-absolutists’ intentionality judgments was still observed? Moreover, Experiment 2 found a similarly large difference in moral absolutists’ and non-absolutists’ intentionality judgments for good and bad side-effects, but found no difference between their depictions of the agents underlying mental states. These results provide support for an alternative account in which intention need not be necessary in order to perform an act intentionally (Harman, 1976; Brattman, 1984; Knobe, 2004). On these accounts simply foreseeing a bad act is enough for the act to be intentional. More-

over, participants' verbal reports reflect this latter account when they are asked to explain why they believed an act was intentional or unintentional (Tannenbaum, Ditto, & Pizarro, unpublished data).

### *Generalizability and Future Directions*

All five experiments in this paper sampled from university student populations, so the familiar concerns of external validity apply here as well (Sears, 1986). However, there is reason to presume that these findings would generalize to non-student populations. The general side-effect effect asymmetry has been documented with New York pedestrians (Knobe, 2003a), 4-year old children (Leslie et al., 2006), non-Western populations (Knobe & Burra, 2006), and even in clinical populations with brain lesions to emotion-relevant areas of the brain (Young, Cushman, Adolphs, Tranel, & Hauser, 2006). The general side-effect asymmetry in intentionality judgments is so robust that one commentator has remarked, "... it is more challenging to eliminate the effect than it is to demonstrate it" (Alicke, in press).

What about the generalizability of moral values, which I have posited (Knobe, 2004) is partly responsible for the variation in participants' intentionality attributions? Moral values will certainly vary from population to population (a central assumption to this paper), and I would predict that intentionality judgments would follow accordingly: American Midwesterners might not show asymmetric intentionality judgments when a cow is harmed or helped, but a Hindu probably would. Moral values have been demonstrated in past research to be powerful predictors of "real world", high-stakes moral judgments (Ginges, Atran, Medin, & Shikaki, 2007; Skitka & Bauman, unpublished manuscript). Thus, the claim that moral values will vary substantially across cultures is not a threat to the current findings. The current study simply claims that all moral values, culturally variable or otherwise, should influence intentionality judgments in a manner consistent with the general side-effect effect.

Another plausible critique of the current research is that the intentionality asymmetries may be an experimental artifact of the hypothetical, "game-like" nature of many of the stimuli. While there is a healthy ongoing debate of the pros and cons of using fanciful hypothetical dilemmas to assess moral intuitions (Mikhail, 2005, 2007; Sunstein, 2005), it would be unfair to liken the stimuli discussed in the current studies to that of a unrealistic and often bizarre scenarios, such as the trolley problem and its ilk (Thompson, 1986). Importantly, participants were asked to entertain plausible tradeoff decisions that if implemented would have real consequences; corporate executives do make decisions that have consequences for the environment, and military leaders certainly face tough decisions when determining when to risk civilian lives.

Although Studies 3–5 were partially motivated by an attempt to test intentionality judgments in more realistic and involving contexts, future research is certainly needed on this matter. One promising line of research is in the area of juror decision-making. Discourse has already begun to examine the implications of the intentionality asymmetry on basic legal assumptions of the relationship between *Actus Reus* and *Mens Rea* (Nadelhoffer, 2004). In a similar vein, it would be interesting to see how differences in intentionality judgments (as a result of moral considerations) constrain and direct future information processing, especially as mitigating and aggravating evidence becomes available. For jurors who must assess the merits of a given criminal case, concerted attention to specific information becomes

necessary in order to maintain a coherent narrative of all the information (Pennington & Hastie, 1992; Simon, Chadwick, & Read, 2004). So an initial belief that one acted intentionally or unintentionally, then, could possibly influence which pieces of information a juror was predisposed to seek out. There is already evidence suggesting that intentional attributions of a blameworthy act influence how we reconstruct the event from memory and even distort non-mentalistic details of the act, such as how much money was stolen (Pizarro, Laney, Morris, & Loftus, 2006). More research on this topic would be a natural direction for future studies.

Another interesting implication of the current research is that intentional attribution may be used as indirect measure of moral evaluations. This may be useful when moral evaluations are being elicited about politically sensitive judgments (Pizarro et al., unpublished manuscript). Using an example mentioned earlier, Pizarro et al. found that politically liberal college students were no more likely to judge an act that encouraging gay couples to kiss in public as morally wrong than they were an act that encouraged straight couples to kiss in public. However, actions that encouraged gay kissing was seen as more intentional than were actions that encouraged straight kissing, and this difference was moderated by the participants' disgust sensitivity. Although by no means conclusive, the notion that intentionality judgments may serve as a proxy for implicit moral judgments is an interesting one and is deserving of future research.

### *Conclusion*

The findings reported here support the idea that intentional attribution may be both a mental evaluation and a moral evaluation of an agent's behavior. And since moral values vary from person to person, this also means that intentional attribution will vary systematically from person to person. An implication of this is that people who otherwise agree about an actors state of mind—about what an agent intended and what she merely foresaw—may nonetheless disagree about whether that agent performed the action intentionally, because of different of moral values.

### References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. London: Sage Publications, Inc.
- Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368-378.
- Alicke, M. D. (in press). Blaming badly. *Journal of Culture and Cognition*.
- Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*(1), 1-16.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182.
- Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.
- Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*(1), 24-28.

- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323-370.
- Brattman, M. E. (1984). Two faces of intention. *Philosophical Review*, *93*(375-405).
- Brattman, M. E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. London: Lawrence Erlbaum Associates.
- Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Ginges, J., Atran, S., Medin, D. L., & Shikaki, K. (2007). Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences*, *104*, 7357-7360.
- Grueneich, R. (1982). The development of children's integration rules for making moral judgments. *Child Development*, *53*, 887-894.
- Haidt, J., & Graham, J. (in press). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*.
- Harman, G. (1976). Practical reasoning. *Review of Metaphysics*, *79*, 431-463.
- 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, p. 371-388). Hillsdale, NJ: Erlbaum.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*(2), 263-291.
- Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones & ??? (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.
- Knobe, J. (2003a). Intentional action and side-effects in ordinary language. *Analysis*, *63*, 190-193.
- Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309-324.
- Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*, 181-187.
- Knobe, J. (in press). Reason explanations in folk psychology. *Midwest Studies in Philosophy*.
- Knobe, J., & Burra, A. (2006). Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition*, *6*, 113-132.
- Lakoff, G. (2002). *Moral politics: How liberals and conservatives think* (2nd ed.). Chicago: University of Chicago Press.
- Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect. *Psychological Science*, *17*(5), 421-427.
- Machery, E. (in press). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*.
- Malle, B. F., & Bennett, R. E. (unpublished manuscript). *People's praise and blame for intentions and actions: Implications of the folk concept of intentionality*. University of Oregon.
- Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101-121.

- 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(3), 309-326.
- Meltzoff, A. N., & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (p. 335-366). Oxford: Oxford University Press.
- Mikhail, J. (2005). Moral heuristics or moral competence? reflections on sunstein. *Behavioral and brain sciences*, 28, 557-558.
- Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *TRENDS in Cognitive Sciences*, 11(143-152).
- Morewedge, C. K. (2007). *A mind of its own: Negativity bias in the perception of intentional agency*. Dissertation, Harvard University.
- Morse, S. J. (2003). Inevitable mens rea. *Harvard Journal of Law and Public Policy*, 27(1), 51-64.
- Nadelhoffer, T. (2004). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 24, 259-269.
- Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, 62, 189-206.
- Pizarro, D. A., Knobe, J., & Bloom, P. (unpublished manuscript). *College students implicitly judge interracial sex and gay sex to be morally wrong*. Cornell University.
- Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory and Cognition*, 34, 550-555.
- Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 296-320.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology’s view of human nature. *Journal of Personality and Social Psychology*, 51(3), 515-530.
- Shaver, K. G. (1985). *The attribution of blame*. New York: Springer-Verlag.
- Simon, D., Chadwick, S. J., & Read, S. J. (2004). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology*, 86, 814-837.
- Skitka, L. J., & Bauman, C. W. (unpublished manuscript). *Moral conviction as a political motivator: Does it lead to a conservative electoral advantage?* University of Illinois, Chicago.
- Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength of something more? *Journal of Personality and Social Psychology*, 88(6), 895-917.
- Skitka, L. J., & Tetlock, P. (1993). Providing public assistance: Cognitive and motivational processes underlying liberal and conservative policy preferences. *Journal of Personality and Social Psychology*, 65(6), 1205-1223.
- Sunstein, C. R. (2005). Moral heuristics. *Behavioral and brain sciences*, 28, 531-573.
- Tannenbaum, D., Ditto, P. H., & Pizarro, D. A. (unpublished data). *How do people explain intentional attribution for good and bad outcomes?*

- Tanner, C., & Medin, D. L. (2004). Protected values: No omission bias and no framing effects. *Psychonomic Bulletin and Review*, *11*(1), 185-191.
- Tetlock, P., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853-870.
- Thompson, J. J. (1986). *Rights, restitution, and risk: Essays in moral theory*. Cambridge, MA: Harvard University Press.
- Turner, J. (2004). Folk intuitions, asymmetry, and intentional side-effects. *Journal of Theoretical and Philosophical Psychology*, *24*, 214-219.
- Weiner, B. (1995). *Judgments of responsibility*. New York: The Guilford Press.
- Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the effect of an action's moral status on its intentional status? neuropsychological evidence. *Journal of Culture and Cognition*, *6*, 291-304.

## Appendix Stimulus Materials

### *Experiment 1*

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

*Help Environment:* The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. It will help us increase profits, but it will also help the environment.” The chairman of the board answered, “I don’t care at all about helping the environment. I just want to make as much profit as I can. Let’s start the new program.” They started the new program. Sure enough, the environment was helped.

### *Experiment 2*

*Harm Economy:* The vice-president of an environmental-protection organization went to the chairman of the board and said, “We are thinking of starting a new program. It will help us preserve the environment, but it will also harm the economy: There will be a increase in unemployment as a result.” The chairman of the board answered, “I don’t care at all about harming the economy or the increase in unemployment. I just care about preserving the environment in any way I can. Let’s start the new program.” They started the new program. Sure enough, the economy was harmed and unemployment increased.

*Help Economy:* The vice-president of an environmental-protection organization went to the chairman of the board and said, “We are thinking of starting a new program. It will help us preserve the environment, but it will also help the economy: There will be a decrease in unemployment as a result.” The chairman of the board answered, “I don’t care at all about helping the economy or the decrease in unemployment. I just care about preserving the environment in any way I can. Lets start the new program.” They started the new program. Sure enough, the economy was helped and unemployment decreased.

### *Experiment 3*

*Americans civilian casualties:* Recently, an attack on Iraqi insurgence leaders was conducted by American forces. The attack was strategically directed at a few key rebel leaders that have been responsible for a number of recent attacks on American forces, and it was strongly believed that eliminating these key leaders would cause a significant reduction in the casualties of both American military forces and American civilians working in Iraq. It was known that in carrying out this attack there was a chance of Iraqi civilian casualties, although these results were not intended and American forces sought to minimize the death of civilians. The attack was successful—it eliminated all of the targets and is certain to ensure the safety of American soldiers and civilians. Unfortunately, a number of Iraqi civilians were killed and injured in the attack. American representatives say that while this was not intended, it was necessary to take the risk of civilian deaths in order to ensure

the future safety of a larger number of individuals. They also stated that sometimes it is necessary to allow the death of a small number of innocent people in order to promote a greater good.

*Iraqis civilian casualties:* Recently, an attack on American Forces was conducted by Iraqi insurgence leaders. The attack was strategically directed at a few key military outfits that have been responsible for a number of recent attacks against the Iraqi rebels, and it was strongly believed that eliminating these key outfits would cause a significant reduction in the casualties of both Iraqi military forces and of Iraqi civilians. It was known that in carrying out this attack there was a chance of American civilian casualties, although these results were not intended and American forces sought to minimize the death of civilians. The attack was successful—it eliminated all of the targets and is certain to ensure the safety of Iraqi civilians and civilians. Unfortunately, a number of American civilians were killed and injured in the attack. Iraqi rebels say that while this was not intended, it was necessary to take the risk of civilian deaths in order to ensure the future safety of a larger number of individuals. They also stated that sometimes it is necessary to allow the death of a small number of innocent people in order to promote a greater good.

#### *Experiment 4*

*American civilian casualties:* Recently, an attack on a key Al Queda leader, Umar Abd akhtar, was conducted by American Forces. Abd akhtar has ties to Osama Bin Laden and has played a role in past bombings. American intelligence officials had good reason to believe that Abd akhtar was planning yet another attack, and was currently hiding in a rural mountain region of Idaho. Because of tactical reasons, American forces had only a small window of time (24 hours) to attack. It was known that in carrying out this attack there was a chance of American civilian casualties, although these casualties were not intended and American forces sought to minimize the death of civilians. The attack was successful—an air strike killed Abd akhtar and is certain to ensure the safety of American soldiers and civilians. Unfortunately, a number of American civilians were killed and injured in the attack. American representatives say that while this was not intended, it was necessary to take the risk of civilian deaths in order to ensure the future safety of a larger number of individuals.

*Afghan civilian casualties:* Recently, an attack on a key Al Queda leader, Umar Abd akhtar, was conducted by American Forces. Abd akhtar has ties to Osama Bin Laden and has played a role in past bombings. American intelligence officials had good reason to believe that Abd akhtar was planning yet another attack, and was currently hiding in a rural mountain region of Afghanistan. Because of tactical reasons, American forces had only a small window of time (24 hours) to attack. It was known that in carrying out this attack there was a chance of Afghan civilian casualties, although these casualties were not intended and American forces sought to minimize the death of civilians. The attack was successful—an air strike killed Abd akhtar and is certain to ensure the safety of American soldiers and civilians. Unfortunately, a number of Afghan civilians were killed and injured in the attack. American representatives say that while this was not intended, it was necessary to take the risk of civilian deaths in order to ensure the future safety of a larger number of individuals.

*Experiment 5*

*High School:* Policy makers representing all of California's public school districts recently convened to discuss the prevalence of sexual behavior in schools. It is a well-established fact that many young middle-school and high-school students engage in sexual activity. Because the policy makers all agreed that teens would probably engage in intercourse no matter what, they decided to distribute condoms to the students so that, at the very least, these teens would not spread disease to their partners. The committee stated that although it may be wrong for children to have sex at such an early age, the least we can do is minimize the harm that results from their decision to have sex.

*Military:* Legislators in charge of making policy for the US Armed Forces recently convened to discuss sexual behavior among soldiers in foreign countries. It is a well-established fact that many soldiers engage in sexual aggression in the form of rape, generally targeting women from the countries in which they are currently serving duty. Because the legislators all agreed that soldiers would probably rape women no matter what, the policy makers decided to distribute condoms to the soldiers so that, at the very least, these soldiers would not spread disease to their victims. The legislators stated that although it may be wrong for soldiers to rape women, the least we can do is minimize the harm that results from their decision to rape.