Asymmetries in Judgments of Responsibility and Intentional Action

JENNIFER COLE WRIGHT AND JOHN BENGSON

Abstract: Recent experimental research on the 'Knobe effect' suggests, somewhat surprisingly, that there is a bi-directional relation between attributions of intentional action and evaluative considerations. We defend a novel account of this phenomenon that exploits two factors: (i) an intuitive asymmetry in judgments of responsibility (e.g. praise/blame) and (ii) the fact that intentionality commonly connects the evaluative status of actions to the responsibility of actors. We present the results of several new studies that provide empirical evidence in support of this account while disconfirming various currently prominent alternative accounts. We end by discussing some implications of this account for folk psychology.

Awareness that an action is intentional plays an important role in evaluations of an actor and her action. This is only natural: if *x* intentionally acts to bring about a bad outcome, we may form different judgments about *x* or *x*'s behavior than if that same outcome is simply an accident or the result of (non-willful) ignorance. In this way, whether or not a given action is intentional matters to us when we assess an action's or actor's evaluative status. This relation between judgments of intentionality[1] and judgments about the goodness/badness of actions or the responsibility (e.g. praiseworthiness/ blameworthiness) of actors seems to be a relatively straightforward part of folk psychology. What is surprising is that recent experimental research suggests that there is actually a *bi*-directional relation between attributions of intentional action and evaluative (or normative) considerations. A series of studies suggest that not only do attributions of intentional

Thanks to Jonathan Dancy, Shaun Nichols, Mark Phelan, George Sher, Ed Sherline, and especially Joshua Knobe for helpful comments and discussion. We are also grateful to two anonymous reviewers for *Mind & Language*, as well as audiences and commentators at the 2006 Mountain-Plains Philosophy Conference and 2007 Central APA. Finally, we wish to thank Christin Covello, Jerry Cullum, Bill Devlin, and Piper Grandjean for assistance with data collection/entry.

Address for correspondence: Jennifer Wright, Department of Psychology, College of Charleston, 65 Coming Street, Office #104, Charleston, SC 29424, USA; John Bengson, Department of Philosophy, University of Texas at Austin, 1 University Station, C3500 Austin, TX 78712, USA.
Email: wrightjj1@cofc.edu; jsteele@mail.utexas.edu

[1] Throughout, we use 'intentionality' to denote a particular property of *actions*, namely, the property of being done intentionally. It is somewhat of an open question whether intentionality requires the specific *mental state* (or event) of intending, as defenders of the so-called 'simple view' claim (Adams, 1986; McCann, 1986, 2005). Since our primary focus is the explanation of attributions of intentionality, not intention, we shall set this issue to the side here.

Mind & Language, Vol. 24 No. 1 February 2009, pp. 24–50.

action influence evaluative considerations, but evaluative considerations also influence attributions of intentional action.[2]

While we believe that these sorts of experimental results must be treated with care in a philosophical setting, it remains incumbent on philosophers concerned with the nuances of folk psychology to provide a descriptively correct understanding of this phenomenon, coined the 'Knobe effect' (Nichols and Ulatowski, 2007). In what follows, we defend a novel account of the Knobe effect which exploits two factors: (i) an intuitive asymmetry in judgments of responsibility (e.g. praise and blame) and (ii) the fact that intentionality commonly connects the evaluative status of actions to the responsibility of actors. Along the way, we present the results of several new studies that provide empirical evidence in support of this account while simultaneously disconfirming various currently prominent alternative accounts.

In §1, we present our account. In §2, we provide empirical evidence that strongly supports it. In §3, we critically discuss several currently prominent alternative accounts. Then, in §4, we consider the possibility that the Knobe effect arises in non-moral cases. We end, in §5, by discussing some implications of our account for folk psychology.

1. A Two-factor Account of the Knobe Effect

As an illustration of the Knobe effect, consider the following two scenarios (taken from Knobe, 2003a).

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

When given these two scenarios, participants' dominant (70-80%) response was to say that in HARM the chairman harmed the environment intentionally, whereas in HELP the chairman did *not* help the environment intentionally (Knobe, 2003a; see also Nichols and Ulatowski, 2007).

---

[2] See, e.g. Knobe, 2003a, 2003b, 2004, 2005; Knobe and Mendlow, 2004; Nadelhoffer, 2004a, 2004b, 2004c, 2006; McCann, 2005; Young *et al.*, 2006; Nichols and Knobe, 2007; Nichols and Ulatowski, 2007; Adams and Steadman, 2007; Phelan and Sarkissian, 2008; and Machery, 2008.

To many, this asymmetry seems odd. After all, the only ostensible difference between the two scenarios is that in HARM the environment was *harmed* as a result of the chairman's action and in HELP the environment was *helped*. If one judges that the chairman acted intentionally in HARM, then because the two scenarios are at first glance in all relevant ways similar, it seems that one should also judge that he acted intentionally in HELP (and likewise if one judges that he did not act intentionally). Yet, the fact remains that participants' intentionality attributions were very clearly asymmetrical.

This asymmetry requires explanation. A popular explanation posits the bi-directional relation between evaluative considerations and intentionality attributions described at the outset: not only do attributions of intentional action influence evaluative considerations, but evaluative considerations also influence attributions of intentional action.[3] While this response is well-motivated, it is clear that an appeal to a bi-directional relation alone cannot explain the asymmetry. For such an appeal does not by itself explain *why* or *how* such considerations, whatever they happen to be, lead to an asymmetry in intentionality attributions. What is it about evaluative considerations that enable them to influence intentionality attributions in an asymmetrical fashion?

In the remainder of this section, we detail an account of the Knobe effect designed to answer this question. We propose that there is an asymmetry in judgments of responsibility (e.g. praise and blame) that, because of the putative connection between responsibility and intentionality, generates the Knobe effect. The basic idea, as applied to HARM/HELP, is simple: in HARM, a judgment of responsibility (in this case, blame) results in an attribution of intentionality, whereas in HELP the absence of a judgment of responsibility (in this case, praise) results in the absence of an attribution of intentionality. On our account, then, the asymmetry in intentionality attributions in cases such as HARM and HELP can be explained by appeal to two factors: (i) an intuitive asymmetry in judgments of responsibility and (ii) the fact that intentionality commonly connects the evaluative status of actions to the responsibility of actors.[4]

Let us begin by considering the first component of our account, namely, that there is an asymmetry in judgments of responsibility.[5] Intuitively, agents are typically to some extent blameworthy, criticizable, or otherwise *negatively responsible* when they engage in an action that they know will bring about a bad outcome, or an outcome

---

[3] See, e.g. Knobe, 2003a, 2003b, 2004, 2005, 2006, 2007; Knobe and Mendlow, 2004; Nadelhoffer, 2004a, 2004b, 2004c, 2006; and Malle, 2006.

[4] As will become clear, our account differs in several important ways from alternative accounts. We critically discuss three prominent alternatives in §3.

[5] Throughout we focus on normative or *evaluative* (as opposed to causal) responsibility, which may involve both moral and non-moral evaluation. We take it that being blameworthy or praiseworthy is sufficient, though not necessary, for being evaluatively responsible. That is, being blameworthy or praiseworthy entails being responsible in the relevant sense. The same goes for being criticizable or laudable or deserving of credit. For an example of evaluative responsibility without moral responsibility, consider that an athlete may be praised or criticized for her performance. A judgment of responsibility of this sort is clearly not moral. This is not to say that it is an attribution of mere causal responsibility; on the contrary, it is clear that in such a case the athlete's performance is being *evaluated*.

which they have reason to not bring about, even if that outcome is simply a side-effect of an intended outcome. On the other hand, agents are not typically praiseworthy, laudable, or otherwise *positively responsible* merely for bringing about a good outcome, or an outcome which they have reason to bring about. This is so regardless of whether that outcome is an intended outcome or merely a side-effect, foreseen or not, of an intended outcome. For, typically, in order to be to some extent positively responsible for bringing about a good outcome, one must bring about that outcome *for the right reasons*—that is, *because* one has reason to bring it about (Wolf, 1990, p. 84). No corresponding requirement appears to hold for negative responsibility.[6]

This asymmetry appears to emerge in HARM and HELP. In HARM the chairman presumably knew that his action would have a bad outcome, thereby making blame (i.e. an attribution of negative responsibility) seem warranted. After all, he presumably knew that he had reason not to implement the new program—namely, that it would harm the environment—yet he still implemented the program anyway. On the other hand, in HELP the chairman brought about a good outcome—namely, helping the environment—but did not do so for the right reasons (viz., because it would help the environment). Since he implemented the program simply because he desired to make money, praise (i.e. an attribution of positive responsibility) seems unwarranted.

This brings us to the second component of our account, namely, the putative connection between responsibility and intentionality. Typically, those who intentionally act to bring about a bad outcome are negatively responsible and those who intentionally act to bring about a good outcome are positively responsible. In this way, the intentionality of actions commonly connects the evaluative status of actions to the responsibility of actors. Of course, intentionality is clearly not necessary: other factors, such as negligence (e.g. drunk driving) and willful ignorance, can also connect them. Nevertheless, intentionality commonly plays this connecting role. We can diagram the connection in question in the following manner: typically,

good/bad action+intentional action=positively/negatively responsible actor.

Normally, we infer the presence/absence of positive/negative responsibility from both the presence/absence of goodness/badness and the presence/absence of intentionality.

---

[6] Here and throughout, unless otherwise noted we restrict ourselves to cases in which x has a known reason to ϕ or not ϕ, ϕ-ing or not ϕ-ing is an action properly attributed to x, and x is normal (i.e. possesses the general capacities presupposed by agency). Given these restrictions, we believe that the observations in the text hold for both prima facie and all things considered judgments of positive/negative responsibility, as well as judgments of positive/negative responsibility that are made relative to some salient standard (even if the assessor does not herself accept this standard). We will ignore these complications in what follows. We will also ignore complications that may arise from the Doctrine of Double Effect. While our discussion at times runs together judgments of responsibility and judgments of praise/blame (laudability/criticizability, etc.), the Doctrine of Double Effect and various other considerations suggest that responsibility and praise/blame (laudability/criticizability, etc.) may come apart. See below for related discussion.

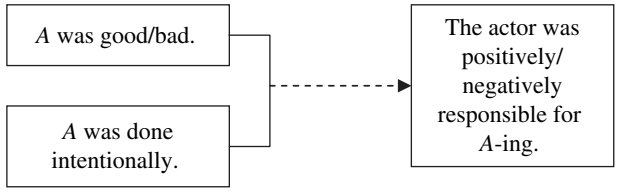

Figure 1 *Reasoning from goodness/badness and intentionality to positive/negative responsibility.*

However, the 'equation' above, which represents certain relations between goodness/badness, intentionality, and positive/negative responsibility, makes clear how other inferences are possible. Given that actors are typically held to be positively/negatively responsible for actions which are good/bad only if they act intentionally, the responsibility of actors serves as an indicator for intentionality. So, just as we can infer the value of *n* from the equation $m + n = o$ given specific values for *m* and *o*, it is possible to infer, albeit defeasibly, the presence/absence of intentionality given specific information regarding the goodness/badness of an action and the positive/negative responsibility of an actor. To be sure, we cannot simply 'read off' intentionality from such evaluative considerations. Rather, the idea is that given the relations represented in the 'equation' above, the goodness/badness of an action and the responsibility of an actor can be used to defeasibly infer the presence/absence of intentionality.[7]

Suppose, for instance, that one judges that an outcome of a given agent's action *A* is bad, and that the agent knows this but *A*-s nevertheless; accordingly, one judges that the agent is blameworthy for *A*-ing. One might then reason that since the agent is blameworthy (negatively responsible) for *A*-ing, then because typically an agent who is responsible for *A*-ing *A*-ed intentionally, it is probably the case that the agent *A*-ed intentionally. In this way, one can defeasibly infer the presence/absence of intentionality from the goodness/badness of an action and the responsibility of an actor.

In certain situations, such as those in which there is a relative paucity of direct information regarding whether an agent's action was intentional, it may be extremely useful to be able to reason thus. For instance, consider a graduate student who decides that if there is ever reason to think that one of her professors has humiliated her intentionally, she will leave the program. Suppose that in the middle of a class presentation for which the student has prepared an elaborate handout, one of her professors makes a loud noise, wads up the handout, and tosses it in the trash bin. The student is aware that the self-absorbed professor's primary goal in discarding her handout was not to humiliate *her*, but rather to express *his* disapproval; still, a side-effect of his action was that she was humiliated. Reflecting on the professor's actions later that evening, she considers that the professor is criticizable

[7] This inference is most likely inductive or abductive in nature. In effect, the analogy with a simple mathematical inference is merely an analogy. The '+' and '=' in the 'equation' in the text presumably should be interpreted as marking transitions in what Harman (1973) calls 'inference to the best total explanation' or some other non-deductive inference.

for bringing about this side-effect. After all, he 'should have known better': presumably, he knew that he had reason not to act as he did—namely, that it would humiliate her—yet he still acted anyway. Since the fact that he is criticizable for humiliating her is most likely not due to some sort of (say) negligence or willful ignorance, his criticizability for humiliating her is an indicator that he humiliated her intentionally; in other words, holding that he humiliated her intentionally may seem to her to provide the 'best total explanation' of his criticizability. The student thus concludes, rightly or wrongly, that she has reason to think that her professor humiliated her intentionally; consequently, she decides to leave the program.

As this example illustrates, if an agent who acts to bring about a (bad) outcome is blameworthy, it is possible to infer that the agent brought about that outcome intentionally. An analogous case could be constructed to illustrate that if an agent who acts to bring about a (good) outcome is praiseworthy, it is possible to infer that the agent brought about that outcome intentionally.

It is tempting to suppose that this inference to intentionality, like the fallacy of affirming the consequent, is not philosophically defensible. Although we wish to remain neutral regarding whether the inference in question is in fact justified, it is worth pausing for a moment to briefly register two reasons to think that it might sometimes be. First, it is not at all clear that this inference commits a fallacy such as affirming the consequent. But whether or not it does is, in a way, irrelevant, since formal fallacies often make for reasonable inductive or abductive inferences. Given that the inference to intentionality is presumably inductive or abductive in nature (see note 7), it might on certain occasions be justified even if it commits a formal fallacy. Second, many philosophers have argued that to the extent that one knowingly φ-s and is properly held responsible for φ-ing, one φ-s intentionally (Harman, 1976; Duff, 1982; Bratman, 1984). This position, or one like it, might very well vindicate the inference to intentionality in question. (Alternatively, one might view the majority judgments in the studies discussed below as supporting, rather than supported by, this position.)

Whether or not the inference to intentionality is justified, it appears to be present in HARM and HELP. Viewing the HARM chairman as blameworthy for a bad outcome puts one in a position to infer that he acted intentionally.

Viewing the HELP chairman as not praiseworthy not only does not put one in a position to infer that he acted intentionally, it actually creates a reason to not attribute intentional action to him. For if the chairman helped the environment (a reputedly *good* action) intentionally, then there would be good reason to hold

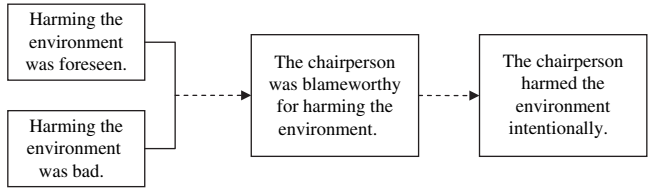

Figure 2 *Reasoning in HARM.*

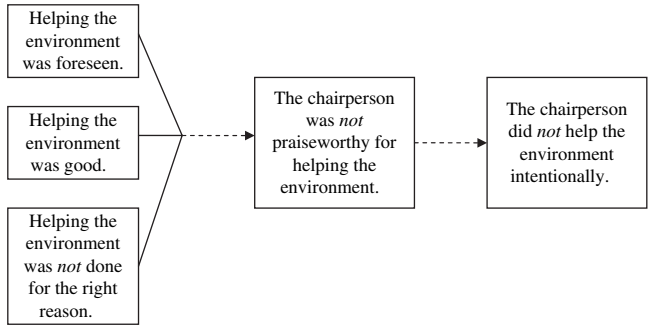

Figure 3 *Reasoning in HELP.*

him praiseworthy. Since one judges him to be not praiseworthy, one infers that he did not help the environment intentionally.

We believe that these considerations recommend an explanation of the Knobe effect in terms of the following two factors:

i. judgments of positive/negative responsibility are asymmetrical;
ii. the intentionality of actions commonly connects the evaluative status of actions to the responsibility of actors, the latter of which alone typically implies intentionality.

Factor (i) locates the source of the asymmetry in intentionality attributions; factor (ii) explains how this source connects up with intentionality. Together, these two factors provide a straightforward account of the Knobe effect. Participants in the relevant studies typically *blamed* in the (reputedly *bad*) HARM scenario, but did not *praise* in the (reputedly *good*) HELP scenario. Because one may attribute intentionality when responsibility and goodness/badness is present, this resulted in more frequent attributions of intentionality in HARM than in HELP.

2. Empirical Support: Two New Studies

We conducted two studies designed to test the empirical adequacy of this account. In the first study, 122 participants were given slightly revised HARM and HELP scenarios. To collect within-subjects data, all participants were given both scenarios, which were counterbalanced to eliminate the possibility of order effect. In addition to being asked (a) whether or not the chairperson (which was substituted for 'chairman' to eliminate the possibility of gender bias) acted intentionally, participants were asked (b) whether the harming/helping of the environment was good, bad, or neither and (c) whether the chairperson deserved any praise, blame, or neither for harming/helping the environment.

In HARM, the vast majority of participants (92.6%) judged that harming the environment was bad, that the chairperson deserved blame (88.4%), and that the

chairperson harmed the environment intentionally (64.8%). HELP elicited very different judgments: while the vast majority of participants (90.1%) judged that helping the environment was good, a small minority judged that that the chairperson deserved praise (14.9%) and that the chairperson helped the environment intentionally (4.1%).

Most of the intentionality attributions in HARM accompanied judgments of both a *bad* action and a *blameworthy* chairperson; likewise, most of the intentionality attributions in HELP accompanied judgments of both a *good* action and a *praiseworthy* chairperson. In both cases, participants were significantly more likely to judge that the chairperson acted intentionally when they stated *both* that the action was good/bad and that the chairperson was praiseworthy/blameworthy than when they only agreed to one or neither of these (HARM: 68% versus 44%, $\chi^2(121) = 3.7$, $p = .054$, $\varphi = .18$; HELP: 19% versus 2%, $\chi^2(120) = 9.8$, $p = .002$, $\varphi = .29$).

A close look at the data reveals which of these two judgments played the central role. In both cases, participants were significantly more likely to judge that the chairperson acted intentionally when they stated that the chairperson was praiseworthy/blameworthy than when they did not (HARM: 69% versus 29%, $\chi^2(122) = 8.9$, $p = .003$, $\varphi = .27$; HELP: 22% versus 1%, $\chi^2(121) = 17.5$, $p < .001$, $\varphi = .38$). On the other hand, in both cases, participants were no more likely to judge that the chairperson acted intentionally when they stated that the action was good/bad than when they did not (HARM: 63% versus 78%, $\chi^2(122) = .72$, $p = .40$, $\varphi = -.08$; HELP: 4% versus 8%, $\chi^2(121) = .59$, $p = .44$, $\varphi = -.07$). In short, participants' judgments of goodness/badness, when considered alone, were not significantly correlated with their intentionality attributions, whereas judgments of *both* praise and blame were.

Further evidence that judgments of responsibility played the central role comes from considering partial correlations between judgments of badness, blame, and intentionality. In HARM, when the variance explained by judgments of badness was controlled for, judgments of blame and intentionality became more strongly positively correlated (partial $r = .33$, $p < .001$) because error variance decreased. On the other hand, when the variance explained by judgments of blame was controlled for, judgments of badness and intentionality became *negatively* correlated (partial $r = -21$, $p = .022$).

Of course, this does not mean that judgments of badness played no role at all; after all, participants were significantly more likely to blame the chairperson when they considered the chairperson's action to be bad than when they did not (92% versus 44%, $\chi^2(121) = 18.4$, $p < .001$, $\varphi = .39$). It is just that judgments of badness became relevant to intentionality attributions only when coupled with judgments of blame. That is, as illustrated by Figure 2 above, judgments of badness were only *indirectly* related to attributions of intentionality insofar as they increased the likelihood of judgments of blame (and thus attributions of intentionality).

These results provide strong support for our account. Because of the asymmetry in judgments of positive/negative responsibility, participants were significantly more likely to hold the chairperson responsible in HARM than in HELP. This asymmetry, coupled with the fact that intentionality commonly connects the

goodness/badness of actions to the responsibility of actors, explains why participants were far more likely to make intentionality attributions in HARM than in HELP. Again, a judgment of responsibility was central. Judgments of goodness/badness alone did not lead to intentionality attributions: there was no need for participants to ascribe intentionality in order to link goodness/badness to praise/blame when the latter was judged to be absent. But when participants judged the chairperson to be praiseworthy/blameworthy for the relevant action, because responsibility typically implies intentionality, they ascribed intentionality. Given that responsibility was ascribed to the HARM chairperson far more frequently than to the HELP chairperson, intentionality, too, was ascribed to the HARM chairperson's action far more frequently than to the HELP chairperson's action. In this way, the asymmetry in intentionality attributions in HARM/HELP resulted from more frequent judgments of blame than praise.[8]

Of course, the statistical analyses we employed in this study provide results that are strictly speaking neutral regarding the directionality of the relation between judgments of responsibility and intentionality attributions. Nevertheless, these analyses establish that there is a strong relation between these judgments. Given the observations in §1 concerning, first, the asymmetry in judgments of responsibility and, second, the putative connection between responsibility and intentionality, we have good reason to believe that in HARM/HELP intentionality attributions were influenced by judgments of responsibility, and not the other way around. Because the conditions under which judgments of positive and negative responsibility are appropriate differ, participants were significantly more likely to hold the actor responsible in one case than in the other. Since responsibility typically implies intentionality, this led to more intentionality attributions in one case than in the other: hence, the asymmetry in intentionality attributions.

In search of empirical evidence for the hypothesis that intentionality attributions can influence judgments of responsibility, we conducted a second study that employed standard manipulation techniques to determine direction of influence.[9] Such techniques systematically vary the value of one (independent) variable in

---

[8] While the majority of participants (62%) made asymmetrical intentionality attributions, there was of course a minority (38%) who did not. Nichols and Ulatowski (2007) suggest that this sort of result indicates that there are multiple correct interpretations of 'intentional' (cf. Sosa, 2007). This 'interpretative diversity hypothesis' acknowledges that some individuals' responses demonstrate an asymmetry in intentionality attributions, and thus it does not challenge the need for an explanation of this asymmetry. So, despite our reservations about certain applications of this sort of hypothesis (see Bengson *et al.*, forthcoming), we are content to note that our account provides an explanation of the asymmetry in intentionality attributions when it appears, while remaining neutral on the question of whether or not there is a single correct interpretation of 'intentional'.

[9] To our knowledge, other researchers have not yet conducted studies designed to determine direction of influence. Thus, to the extent that existing accounts of the Knobe effect make claims about direction of influence (e.g. that judgments of badness, transgressions, or costs influence intentionality attributions), they remain empirically untested. We hope that the manipulations which follow go some distance towards filling this gap in current research on the Knobe effect.

order to determine whether it results in variation in the value of another (dependent) variable. If variation in the independent variable results in variation in the dependent variable, then there is strong evidence that the independent variable influences the dependent variable. In the present case, the independent variable was the presence or absence of a judgment of positive/negative responsibility and the dependent variable was the presence or absence of an intentionality attribution. The goal of the study was to systematically vary participants' judgments of positive/negative responsibility in order to determine whether such variation resulted in variation in participants' intentionality attributions. If variation in participants' judgments of positive/negative responsibility resulted in variation in participants' intentionality attributions, then this would provide strong evidence that judgments of positive/negative responsibility influence intentionality attributions. This, in turn, would provide additional support for our account.

Carrying out this sort of manipulation is difficult in the present case because it requires controlling for the actual judgment(s) that participants immediately form upon reading the HARM/HELP vignettes. Once participants have judged the chairperson to be positively or negatively responsible (or neither), it would be extremely difficult for them to override that initial judgment, thus making any request that they form the opposite judgment highly impractical. Our strategy for overcoming this obstacle was to give participants vignettes in which *other subjects* formed responsibility judgments about the original HARM/HELP cases.[10] For each vignette, participants were asked whether or not the character in the vignette would make an intentionality attribution, given that the character had made the responsibility judgment that he or she had. If judgments of positive/negative responsibility do in fact influence intention-ality attributions, then we would expect participants to respond to a variation in the characters' positive/negative responsibility judgments by reporting a corresponding variation in the characters' intentionality attributions. Thus, our hypothesis was that a systematic variation in the characters' positive/negative responsibility judgments would elicit from participants a corresponding variation in the characters' intentionality attributions.

While this hypothesis could be tested using a between-subjects design, a within-subjects confirmation of the hypothesis would be more powerful. Therefore, all participants ($N = 113$) were given four scenarios, which were counterbalanced to eliminate the possibility of order effect. The vignettes involving the original HARM case were set up as follows:

HARM - BLAME: Joe is given the following case:

The VP 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

[10] The use of such third-person manipulations is well established. In particular, such manipulations have a precedent in several areas of experimental psychology. For example, they have been used to assess theory of mind, perspective-taking, and moral cognition.

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.

After reading the case, Joe was asked to consider whether or not the chairman was blameworthy for harming the environment and also whether or not the chairman had harmed the environment intentionally.

Joe judged that the chairman WAS BLAMEWORTHY FOR HARMING THE ENVIRONMENT. Did Joe judge that the chairman harmed the environment intentionally?

[ ] PROBABLY, YES [ ] PROBABLY, NO

HARM - NO BLAME: Bob is given the following case:

The VP 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.

After reading the case, Bob was asked to consider whether or not the chairman was blameworthy for harming the environment and also whether or not the chairman had harmed the environment intentionally.

Bob judged that the chairman WAS NOT BLAMEWORTHY FOR HARMING THE ENVIRONMENT. Did Bob judge that the chairman harmed the environment intentionally?

[ ] PROBABLY, YES [ ] PROBABLY, NO

Participants were also given vignettes involving the original HELP case, which were modified in an analogous manner.

For the vignettes involving the original HARM case, when the character in the vignette judged that the chairperson was blameworthy for harming the environment, the vast majority of participants (90.3%) reported that the character also judged that the chairperson harmed the environment intentionally. On the other hand, when the character judged that the chairperson was *not* blameworthy, the vast majority of participants (79.6%) reported that the character also judged that the chairperson did *not* harm the environment intentionally. Most participants (71.7%) changed their report of the characters' intentionality attributions in the direction hypothesized: as predicted, when a judgment of blame was present/absent, an intentionality attribution was correspondingly present/absent. These results were significantly above chance ($p < .001$).

A similar pattern of judgments emerged for the vignettes involving the original HELP case. When the character in the vignette judged that the chairperson was praiseworthy for helping the environment, the majority of

participants (59.3%) reported that the character also judged that the chairperson helped the environment intentionally. And when the character judged that the chairperson was *not* praiseworthy, the vast majority of participants (90.3%) reported that the character also judged that the chairperson did *not* help the environment intentionally. Once again, most participants (54%) changed their report of the characters' intentionality attributions in the direction hypothesized: as predicted, when a judgment of praise was present/absent, an intentionality attribution was correspondingly present/absent. These results were significantly above chance ($p < .001$).[11]

These findings confirm our hypothesis that a systematic variation in the characters' positive/negative responsibility judgments would elicit from participants a corresponding variation in the characters' intentionality attributions. This provides strong evidence that judgments of positive/negative responsibility do in fact influence intentionality attributions.

In addition to the manipulation just described, participants were also given a set of control vignettes. These vignettes exactly mirrored the vignettes described above, except that participants were given the characters' intentionality attributions, rather than the characters' responsibility judgments. For each vignette, participants were asked whether or not the character in the vignette would make a positive/negative responsibility judgment, given that the character had made the intentionality attribution that he or she had. As noted at the outset, it is clear that intentionality attributions influence judgments of positive/negative responsibility. So, if the manipulation technique employed here is reliable, we should expect a systematic variation in the characters' intentionality attributions to elicit from participants a corresponding variation in the characters' positive/negative responsibility judgments. And this is what we found. For the vignettes involving the original HARM/HELP cases, the majority of participants (99.1% and 60.2%, respectively) reported that the character judged that the chairperson was blameworthy/praiseworthy when judged to have acted intentionally; a vast majority (71.7% and 87.6%, respectively) reported that the character judged that the chairperson was *not* blameworthy/praiseworthy when judged to have

[11] It is important not to be misled by the fact that 'only' 54% of participants changed their report of the characters' intentionality attributions for the vignettes involving the original HELP case. Since we would expect only a very small percentage of participants to change their report if judgments of responsibility did not influence intentionality attributions, 54% is actually quite substantial. That the percentage here is not as high as the percentage (71.7%) in the HARM case is presumably a byproduct of the quandary participants faced when presented with this case—a quandary not present in the HARM case. Given that it is apparent that the chairperson did not help the environment for the right reasons, it is quite likely that participants could not readily accept that the character made the responsibility judgment that he did. Despite this, the character's judgment that the chairperson was praiseworthy did have a highly significant effect on participants' reports of the character's intentionality attributions, providing strong support for our hypothesis that variation in positive/negative responsibility judgments leads to a corresponding variation in intentionality attributions.

not acted intentionally. The percentage of participants (71.7% and 51.3%, respectively) for whom a variation in intentionality attributions resulted in a variation in positive/negative responsibility judgments was significantly above chance ($ps < .001$).

Taken together, the results of our manipulation and the control vignettes provide empirical evidence for the existence of a bi-directional relation between ntentionality attributions and judgments of positive/negative responsibility. In short, this study indicates that not only do attributions of intentional action influence judgments of positive/negative responsibility, but judgments of positive/negative responsibility also influence attributions of intentional action—just as our account, articulated in §2 (and illustrated by Figures 2 and 3), predicts.[12]

3. Alternative Accounts of the Knobe Effect

We have thus far articulated a novel account of the Knobe effect and presented empirical research which supports it. In this section, we critically discuss several currently prominent alternative accounts. In particular, we will consider the views that judgments of badness explain the Knobe effect, that the Knobe effect is the product of judgments regarding costs, and that affective bias generates the Knobe effect.

3.1 The Badness Account

Knobe and Mendlow (2004) and Phelan and Sarkissian (2008) have reported preliminary research which suggests that a judgment of blameworthiness is not generally required to elicit intentionality attributions. In two pilot studies ($N < 25$), participants were given the following scenario, which we will call DECREASE:

> DECREASE: Susan is the president of a major computer corporation. One day, her assistant comes to her and says, 'We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts but decreasing sales in New Jersey.'

[12] In §5, we address worries to the effect that this conclusion is somehow objectionable. Although we lack the space to discuss the results of other studies which have elicited the Knobe effect, we believe that our account provides a straightforward explanation of participants' responses to these other cases as well. This includes, for example, the soldiers at Thompson Hill cases (see Knobe, 2003a), the rifle cases (see Knobe, 2003b, 2006), the die-rolling cases (see Nadelhoffer, 2004a), the New Jersey sales cases (see Knobe and Mendlow, 2004; Phelan and Sarkissian, 2008; and §3.1 below), the free cup and extra dollar cases (see Machery, 2008; and §3.2 below), and (given that a judgment of positive/negative responsibility may be made relative to some salient standard which the assessor does or does not ultimately accept) the racial identification law cases (Knobe, 2007), among others.

> Susan thinks, 'According to my calculations, the losses we sustain in New Jersey should be a little bit smaller than the gains we make in Massachusetts. I guess the best course of action would be to approve the program.'
> 'All right,' she says. 'Let's implement the program. So we'll be increasing sales in Massachusetts and decreasing sales in New Jersey.'

Participants in both studies did not blame Susan for decreasing sales in New Jersey, yet the majority in both stated that Susan brought about this outcome intentionally. Knobe and Mendlow and Phelan and Sarkissian conclude that (at least in these sorts of cases) something other than the blameworthiness of the actor is generating participants' intentionality attributions.

Knobe and Mendlow (2004) draw one further conclusion, namely, that the source of the asymmetry in participants' intentionality attributions is the perceived badness of decreasing sales in New Jersey. More generally, they claim that the Knobe effect can be explained in the following manner: the perceived badness of the foreseen outcome of actions, *not* the blameworthiness of actors, influences intentionality attributions. On this view, which we will call the *badness account*, judgments of badness, but not goodness (nor positive/negative responsibility), lead participants to attribute intentionality (see also Knobe, 2003a, 2003b, 2006; Pizarro *et al.*, 2008).

There are at least two reasons to be skeptical of the badness account. First, it does not appear to be supported by the results of DECREASE. In their study, Knobe and Mendlow did not ask participants whether or not decreasing sales in New Jersey was bad. Participants in Phelan and Sarkissian's study, on the other hand, were asked whether this action was bad; they judged that it was *not*. Thus, Phelan and Sarkissian conclude that (at least in these sorts of cases), *contra* the badness account, something other than the perceived badness of the action is generating participants' intentionality attributions.

A second reason to be skeptical of the badness account is that it is clearly disconfirmed by the findings reported in §2. Recall that in HARM, participants were no more likely to judge that the chairperson acted intentionally when they stated that the action was bad than when they did not; indeed, participants' judgments of badness considered alone were not significantly associated with their intentionality attributions. Moreover, when the variance explained by judgments of blame was controlled for, judgments of badness and intentionality became *negatively* correlated. These results clearly demonstrate the empirical inadequacy of the badness account.[13]

In addition, the badness account appears to be unable to explain the results of our manipulation study. We found that changes in judgments of responsibility resulted in changes in intentionality attributions. The badness account leaves us

[13] Knobe (2007) now retracts the badness view partly in response to an earlier version of the present paper.

without an explanation of why this should be so. Because the badness account maintains that judgments of badness influence intentionality attributions while judgments of responsibility do *not*, the badness account fails to explain these findings.[14]

Let us return, then, to the results of the New Jersey sales case, which might be interpreted as challenging our account (since DECREASE elicited intentionality attributions without judgments of blameworthiness). There are good reasons to think that they do not. For one, our account does not offer a generally necessary condition for intentionality attributions.[15] It is plain that factors other than responsibility, such as explicitly deliberating about whether to φ and then successfully φ-ing (*modulo* deviant causal chains), are typically sufficient for holding that φ was done intentionally. So, the fact that participants attributed intentionality to Susan's action while failing to judge her blameworthy for that action is entirely consistent with our account.

Indeed, it is not difficult to understand why DECREASE elicited intentionality attributions in the absence of judgments of blame. In that scenario, Susan goes through an explicit deliberative process, weighing the pros and cons of the potential outcome (including decreasing sales in New Jersey, which may or may not be considered a side-effect of her action), before deciding to implement the new program. Given this, it is unsurprising that participants judged that Susan's action was intentional despite the fact that she was not blameworthy: she explicitly engaged in practical reasoning, an action resulting from which would be, *ceteris paribus*, considered intentional (see, e.g. Anscombe, 1957; von Wright, 1983).

In any event, because DECREASE involves explicit deliberation, it is too dissimilar from HARM/HELP to function as a genuine test of the empirical adequacy of our (or, for that matter, any other) account of the Knobe effect. To test various accounts of the Knobe effect, it is necessary to consider scenarios which elicit the Knobe effect. Yet, it is doubtful that DECREASE, together with a variant in which, *à la* HELP, Susan *increased* sales in New Jersey, would have elicited the asymmetrical intentionality attributions characteristic of the Knobe effect, for the presence of explicit deliberation in both scenarios presumably would have led participants to attribute intentionality equally to the two agents' actions.

The lesson, we take it, is that scenarios much more similar to HARM/HELP are needed. With this in mind, we gave 122 participants the following variants on DECREASE:

---

[14] Note that the problem remains even if the badness account was modified so that it held that *non-conscious* judgments, whether of badness or transgression (Knobe, 2007), influence intentionality attributions while judgments of responsibility do *not*. This account would similarly fail to explain the results of our manipulation studies.

[15] In this and other ways, our account differs markedly from the view that Phelan and Sarkissian (2008) refer to as 'BLAM'.

DECREASE2: The VP of a company went to the chairperson of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also decrease sales in New Jersey.' The chairperson of the board answered, 'I don't care at all about decreasing sales in New Jersey. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, profits increased and sales in New Jersey decreased.

INCREASE: The VP of a company went to the chairperson of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also increase sales in New Jersey.' The chairperson of the board answered, 'I don't care at all about increasing sales in New Jersey. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, profits increased and sales in New Jersey increased.

In DECREASE, the main character (Susan) engaged in explicit deliberation. When explicit deliberation was removed, as in DECREASE2 and INCREASE, participants' responses became strikingly similar to those found in HARM/HELP.

First, most of the (many) intentionality attributions in DECREASE2 accompanied judgments of both a *bad* action and a *blameworthy* chairperson; likewise, most of the (few) intentionality attributions in INCREASE accompanied judgments of both a *good* action and a *praiseworthy* chairperson. In both cases, participants were significantly more likely to attribute intentional action to the chairperson when they stated both that the action was good/bad and that the chairperson was praiseworthy/blameworthy than when they only agreed to one or neither of these (DECREASE2: 68% versus 41%, $\chi^2(120) = 7.4$, $p = .007$, $\varphi = .25$; INCREASE: 47% versus 10%, $\chi^2(121) = 13.6$, $p < .001$, $\varphi = .34$). Second, participants were significantly more likely to judge that the chairperson acted intentionally when they stated that the chairperson was praiseworthy/blameworthy than when they did not (DECREASE2: 70% versus 34%, $\chi^2(120) = 14.9$, $p < .001$, $\varphi = .35$; INCREASE: 41% versus 11%, $\chi^2(121) = 10.8$, $p = .001$, $\varphi = .30$). However, they were no more likely to judge that the chairperson acted intentionally when they stated that the action was good/bad than when they did not (DECREASE2: 60% versus 43%, $\chi^2(120) = 3.5$, $p = .09$, $\varphi = .17$; INCREASE: 17% versus 11%, $\chi^2(121) = .70$, $p = .41$, $\varphi = .08$). Third, participants were significantly more likely to blame and marginally more likely to praise when they considered the chairperson's action to be good/bad than when they did not (DECREASE2: 83% versus 18%, $\chi^2(120) = 49.8$, $p < .001$, $\varphi = .64$; INCREASE: 18% versus 5%, $\chi^2(121) = 3.3$, $p = .069$, $\varphi = .17$). This indicates that judgments of goodness/badness became relevant to intentionality attributions only when coupled with judgments of praise/blame.

These findings further disconfirm the badness account. They also reinforce our contention that a judgment of both a good/bad action and a responsible actor typically leads to an intentionality attribution—and that, of the two, the

responsibility of the actor plays the central role. Consequently, rather than challenging our account, New Jersey sales cases support it.

3.2 The Trade-off Hypothesis

Recently, Machery (2008) has claimed that the perceived costs of actions, not a judgment of responsibility (e.g. praise/blame), influences intentionality attributions. On this view, which Machery calls the *trade-off hypothesis*, an actor is judged to act intentionally if she willingly incurs a cost in order to reap a benefit, since people 'believe that costs are intentionally incurred'. To illustrate, consider the HARM case. According to the trade-off hypothesis, the chairperson in HARM is judged to act intentionally because she willingly accepts harm to the environment in order to increase profits. Since harming the environment is a cost while increasing profits is a benefit, and for some reason participants believe that in general a cost is intentionally incurred, participants judge that the chairperson intentionally harmed the environment.[16]

An initial worry about the trade-off hypothesis is that it appears to be unable to explain the empirical findings—in particular, the results of our manipulation study—reported in §2. We found that changes in judgments of responsibility resulted in changes in intentionality attributions. The trade-off hypothesis, like the badness account, leaves us without an explanation of why this should be so. Because the trade-off hypothesis maintains that perceived costs influence intentionality attributions while judgments of responsibility (e.g. praise/blame) do *not*, the trade-off hypothesis fails to explain these findings. In effect, the studies presented in §2 provide support for our account over the trade-off hypothesis, all else being equal.

There is another reason to be skeptical of the trade-off hypothesis. Machery attempts to motivate the trade-off hypothesis by citing the results of a study involving the following vignettes:

> FREE-CUP: Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest-sized drink available. Before ordering, the cashier told him that if he bought a Mega-Sized Smoothie he would get it in a special commemorative cup. Joe replied, 'I don't care if I get a commemorative cup. I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie in a commemorative cup.
>
> EXTRA-DOLLAR: Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest-sized drink available. Before ordering, the cashier told her that the Mega-Sized Smoothies were now one dollar more

[16] To our knowledge, empirical support has yet to be provided for the claims that (i) participants judge the putative side-effect present in a given vignette to be (perceived as) a cost, (ii) participants judge the desired outcome present in a given vignette to be (perceived as) a benefit, or most importantly (iii) participants believe that in general 'costs are intentionally incurred', as Machery claims. In particular, we find (iii) in need of support, especially in light of its alleged role in generating the Knobe effect. In the absence of a link between cost and intentionality, a judgment of cost appears to be irrelevant.

> than they used to be. Joe replied, 'I don't care if I have to pay one dollar more. I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie and paid one dollar more for it.

In FREE-CUP, where receiving the free cup was an unintended consequence of the actor's purchase of a smoothie, the majority of participants judged that the actor did *not* receive a free cup intentionally. However, in EXTRA-DOLLAR, where paying an extra dollar was required in order to purchase the smoothie, the majority of participants judged that the actor *did* pay an extra dollar intentionally. In neither case did the majority judge that the actor was praiseworthy or blameworthy. Machery thus concludes that judgments of praise/blame cannot explain the Knobe effect. On the other hand, he claims that the trade-off hypothesis is well-positioned to explain the asymmetry in this particular case:

> In the extra-dollar case, the agent is confronted with a decision concerning whether to incur an extra cost (paying an extra-dollar) in order to reap a desired benefit (getting a smoothie). In the other case—the free-cup case—the agent is given a benefit (a free cup) in addition to the foreseen benefit that results from her decision (a smoothie).

This research is provocative. Nevertheless, we are not convinced that it supports the trade-off hypothesis as an explanation of the Knobe effect. While receiving a free cup is a side-effect of Joe's intended outcome (getting a smoothie), paying an extra dollar is not—or at least it is not clear that it is.[17] To the extent that the Knobe effect is properly viewed as a *side-effect* effect (see, e.g. Knobe, 2003a), this casts doubt on the claim that the asymmetry in intentionality attributions in these cases supports the trade-off hypothesis *as an explanation of the Knobe effect*.[18] For a similar reason, we are not convinced that this study shows that judgments of praise/blame cannot explain the Knobe effect. For to the extent that paying extra is not a side-effect whereas receiving a free cup is, there is no Knobe effect here to be explained.

What are the implications of the smoothie vignettes for our account, if it turns out that the vignettes *are* an instance of the Knobe effect? When answering this question, it is important to bear in mind that our account invokes, not merely

---

[17] Consider: while some people will find it correct to say that Joe acted with the intention of paying what he did (in EXTRA-DOLLAR), it is plain that he did not act with the intention of receiving a free cup (in FREE-CUP). Note also that some people will find it correct to say that Joe paid an extra dollar in order to get the smoothie. But it would be wrong to say that he received a free cup in order to get the smoothie. (See Knobe 2006, 2007 for discussion of the relevance of the phrase 'in order to' to intentionality attributions.)

[18] For further critical discussion of the trade-off hypothesis, see Mallon (2008), who presents compelling empirical evidence that this hypothesis cannot explain the Knobe effect. Cf. the remarks in note 16.

praise/blame, but responsibility more generally. While being praiseworthy or blameworthy is sufficient for being positively/negatively responsible, it is not necessary. As noted in §1, there are a variety of ways of being responsible: for instance, one might be responsible by being laudable, criticizable, deserving of credit, and so on (see, in particular, note 5). Our account therefore has the resources to explain Machery's results. Presumably, Joe should be held responsible for paying an extra dollar in a way that he should not be held responsible for receiving a free commemorative cup. Consequently, our account predicts that people will be more inclined to judge that Joe paid extra intentionally than to judge that he received the cup intentionally.

To test this hypothesis, we conducted a study in which 78 participants were given FREE-CUP and EXTRA-DOLLAR. To collect within-subjects data, all participants were given both vignettes, which were counterbalanced to eliminate the possibility of order effect. In addition to being asked whether or not the actor (which was changed to 'Suzy' in EXTRA-DOLLAR to reduce the possibility of gender bias) received the cup/paid extra intentionally, participants were asked whether the actor should be held responsible for receiving the cup/paying extra. The majority of participants attributed responsibility (79%) and intentionality (72%) in the extra dollar case; far fewer attributed responsibility (55%) and intentionality (22%) in the free cup case. While participants' responsibility judgments and intentionality attributions did not line up perfectly in the latter case, in *both* cases participants' judgments about responsibility and intentionality were nevertheless significantly positively correlated (free cup: $\varphi = .35$, $p = .002$; extra dollar: $\varphi = .21$, $p = .073$), just as our account would predict. Thus, to the extent that the smoothie vignettes are an instance of the Knobe effect, our account is well-positioned to explain the asymmetry in participants' judgments.

We saw above that the studies reported in §2 provide support for our account over the trade-off hypothesis, all else being equal. In addition, there is reason to doubt both that the smoothie vignettes motivate the trade-off hypothesis as an explanation of the Knobe effect and that they pose a problem for our account. Indeed, rather than challenging our account, as shown by the results of the study reported in the previous paragraph, the smoothie vignettes support it.

3.3 The Bias Account

In a recent discussion of the potential implications of the Knobe effect for the problem of jury impartiality, Nadelhoffer (2006) has proposed that the Knobe effect is due to an affect-driven bias. Nadelhoffer concedes the central contention of our account—that judgments of responsibility (in particular, blame), influence intentionality attributions—but adds that this, in turn, is explained by an affect-driven bias:

> … once morally loaded features are built into scenarios, these features often trump or override the standard application of the concept of intentional

> action—thereby distorting our judgments about intentionality … [A]ffective responses often undermine our ability to apply the concept of intentional action in an unbiased way (Nadelhoffer, 2006, pp. 213-214).

In explicating this view, which we will call the *bias account*, Nadelhoffer invokes Alicke's (2000, p. 557) psychological model of blame attribution, according to which 'cognitive shortcomings and motivational biases are endemic to blame.' According to Nadelhoffer, Alicke's model holds that a judgment that a given act is immoral can '*spontaneously* trigger [an agent] to go into the default mode of blame-attribution—a mode that causes them to be affected by negative and relatively unconscious reactions that prejudice [their judgment of the actor and his action]' (2006, p. 211). As a result, participants in HARM, for instance, are led to attribute intentionality as a result of their affect-driven attribution of blame.[19] Nadelhoffer concludes, 'even though moral considerations surely do act expansively on folk ascriptions of intentional action … ideally they ought not have this effect' (p. 214).

We find Nadelhoffer's claim that it is possible that affect may sometimes have a biasing effect on intentionality attributions to be perfectly reasonable. Nevertheless, there are several reasons to reject the view that an affect-driven bias provides a general explanation of the Knobe effect. First, an appeal to an affect-driven bias is *unnecessary*, since the asymmetry in judgments of positive/negative responsibility, coupled with the observation that responsibility typically implies intentionality, by itself provides an adequate explanation of the Knobe effect that does not reference affect. Second, such an appeal appears *insufficient*, since neuropsychological research conducted on VMPC participants (i.e. participants with dysfunctional emotional processing) by Young *et al.* (2006) and related research reported in Hauser (2006) suggests that intentionality attributions are influenced by evaluative considerations even in the absence of a robust affective reaction.

A more basic worry is that bias accounts in general appear to be *unmotivated*, since the standard line of reasoning offered in support of the claim that participants' judgments are biased, or unjustified, is problematic. This reasoning, which relies heavily on the fact that cases which elicit the Knobe effect are at first glance similar in all respects that are relevant to intentionality attributions, was briefly outlined in §1. In a defense of the bias account, Nadelhoffer (2006) employs this reasoning in his discussion of the following two scenarios:

> THIEF: Imagine that a thief is driving a car full of recently stolen goods. While he is waiting at a red light, a police officer comes up to the window of

[19] Incidentally, because Alicke's model is restricted to blame, the bias account seems unable to explain the results of HARM/HELP and the corresponding manipulations, in which participants' attributions of intentionality were significantly correlated with (and, moreover, were influenced by) judgments of *praise*, not only blame. More generally, adopting Alicke's model seems inconsistent with Nadelhoffer's own view, expressed in Nadelhoffer, 2004b, that *both* praise and blame influence intentionality attributions.

> the car while brandishing a gun. When he sees the officer, the thief speeds off through the intersection. Amazingly, the officer manages to hold on to the side of the car as it speeds off. The thief swerves in a zigzag fashion in the hopes of escaping—knowing full well that doing so places the officer in grave danger. But the thief doesn't care; he just wants to get away. Unfortunately for the officer, the thief's attempt to shake him off is successful. As a result, the officer rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.
>
> DRIVER: Imagine that a man is waiting in his car at a red light. Suddenly, a car thief approaches his window while brandishing a gun. When he sees the thief, the driver panics and speeds off through the intersection. Amazingly, the thief manages to hold on to the side of the car as it speeds off. The driver swerves in a zigzag fashion in the hopes of escaping—knowing full well that doing so places the thief in grave danger. But the driver doesn't care; he just wants to get away. Unfortunately for the thief, the driver's attempt to shake him off is successful. As a result, the thief rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

In a study involving THIEF and DRIVER, participants routinely made dissimilar judgments regarding the two cases; participants given THIEF said that the thief intentionally brought about the death of the police officer significantly more often (37%) than participants given DRIVER said that the driver intentionally brought about the death of the thief (10%). Nadelhoffer (2006, p. 210) contends that since 'the cases are identical in terms of the cognitive and conative considerations of the thief [in THIEF] and the driver [in DRIVER],' the cases ought to have been treated similarly. He concludes that participants' judgments were biased, and that the source of this bias was affect.

However, THIEF and DRIVER do not lend support to the claim that participants' judgments are biased. For reflection upon these cases calls into question the claim that 'the cases are identical in terms of the cognitive and conative considerations of the thief and the driver.' The two scenarios, while similar, are different in (at least) one crucial respect. Consider: while both scenarios involve a driver of a car being approached by a man brandishing a gun, in THIEF the approaching man is a *police officer*, while in DRIVER he is a *thief*. This is a crucial difference. People would typically take themselves to have reason to not speed off (but instead to cooperate) when approached by a police officer, whereas they would have no such reason when approached by a thief. As this reveals, the cases are *not* identical in terms of the cognitive and conative considerations of the thief and the driver. Thus, we lack reason to think that asymmetrical judgments about THIEF and DRIVER must be inappropriate—in which case positing a bias is unmotivated.

This point bears emphasis. In THIEF, the actor knowingly brought about a side-effect which he had reason to not bring about. He was, consequently, blameworthy for having so acted. Presumably, this led participants to make intentionality attributions in THIEF, just as it did in HARM: a judgment of a

blameworthy actor and a bad action led to an intentionality attribution. On the other hand, DRIVER involves distinct cognitive and conative considerations. For in DRIVER, the actor had reason to engage in the action which he knowingly performed. He was, consequently, *not* blameworthy for having so acted. Presumably, this led participants to refrain from making intentionality attributions in DRIVER: it was because participants did not judge the actor to be blameworthy that they did not judge him to have acted intentionally in this case.

As this makes clear, the asymmetry in participants' judgments of negative responsibility—and, as a result, in intentionality attributions—between cases like THIEF and DRIVER appears to make good psychological sense. If this is correct, then reflection on cases such as these provides reason to believe that the Knobe effect is *not* due to an (affect-driven) bias.[20]

3.4 Summary

We have discussed several currently prominent alternative accounts of the Knobe effect and found them wanting. In spite of this, we believe that they identify factors that are relevant to a complete explanation of the Knobe effect. For example, the badness account rightly observes that the Knobe effect is somehow related to judgments of badness, which figure into factor (ii) of our account. And the bias account appears to be correct in claiming that judgments of responsibility, which figure into both factors (i) and (ii), are the primary influence on intentionality attributions. As this illustrates, our account appears to have the resources to explain in a systematic way the appeal of alternative accounts while avoiding their difficulties. For reasons that should be obvious, we take this to be an additional consideration in its favor.

4. A Non-moral Knobe Effect?

While existing accounts of the Knobe effect commonly assume that the phenomenon of interest is to be explained by reference to specifically *moral* considerations (Knobe, 2003a, 2003b, 2004, 2005, 2006; Knobe and Mendlow, 2004; Nadelhoffer,

[20] Adams and Steadman (2004a, 2004b, 2007) propose an alternative sort of bias, or error, account. They appeal to pragmatic connections between judgments of positive/negative responsibility and intentionality attributions in order to explain the Knobe effect as the result of false or fallacious attributions of intentional action. However, empirical research conducted by Knobe (2004), Nadelhoffer (2006), Adams and Steadman (2007), and Nichols and Ulatowski (2007) provides forceful evidence against this explanation. Adams and Steadman (2007) have responded that this evidence does not challenge their view, but is rather simply an indication of the insidiousness of participants' 'pragmatic programming'. We find this response unconvincing, since, as we have just seen, the main line of reasoning in favor of the view that participants' intentionality attributions in these cases are biased, or in error, is problematic. Consequently, positing a malfunction in participants 'programming' to account for the empirical inadequacy of a pragmatic explanation is, at this point, unacceptably *ad hoc*.

2004a, 2004b, 2006; Malle, 2006; Adams and Steadman, 2007; Pizarro *et al.*, 2007), there are reasons to think that the Knobe effect could be elicited by non-moral considerations as well (Turner, 2004; Machery, 2008). Since our account does not invoke specifically moral considerations, it would be entirely consistent with this result. All judgments of positive/negative responsibility, whether moral or not, are subject to the asymmetry discussed in §1. So, our view allows that it is possible to elicit the Knobe effect even in the absence of moral considerations.

This is a virtue of our account. Recall the New Jersey sales cases, in which participants' judgments of responsibility influenced their intentionality attributions (see §3.1). Although DECREASE2/INCREASE elicited the Knobe effect, they are putatively non-moral scenarios. Insofar as these scenarios do not involve explicitly moral considerations, it is likely that participants are not attributing moral responsibility to the actor; presumably, they are attributing a sort of non-moral responsibility. Viewing the chairperson in DECREASE2 as negatively responsible (blameworthy) led participants to say that the chairperson acted intentionally; whereas viewing the chairperson in INCREASE as not positively responsible (not praiseworthy) prevented them from saying that the chairperson acted intentionally. This remains so despite the fact that the relevant sort of responsibility was not moral.[21]

So construed, the results of the New Jersey sales cases indicate that the Knobe effect arises in at least some non-moral cases. Accounts of the Knobe effect which appeal to the alleged influence of specifically moral judgments (e.g. judgments of *moral* badness, *moral* transgressions, or *moral* blame) on intentionality attributions are unable to explain why or how this is so. Yet, while the results of the New Jersey sales cases pose a serious challenge to these other accounts of the Knobe effect, they offer further support in favor of our account.

5. Conclusion

We have argued for a particular account of the asymmetry in folk judgments of intentional action (the Knobe effect). On this account, the asymmetry is best explained by appeal to another asymmetry: namely, the asymmetry in judgments of positive/negative responsibility. Bringing about a foreseen bad outcome is typically sufficient for negative responsibility (e.g. blameworthiness, criticizability), regardless of one's reasons. On the other hand, positive responsibility (e.g. praiseworthiness, laudability) typically requires more, namely, bringing about a foreseen good outcome *for the right reasons*. This asymmetry, coupled with the fact that intentionality commonly connects the evaluative status of actions to the responsibility of actors, accounts for the asymmetry in intentionality attributions.

As noted at the outset, it is clear that if x intentionally acts to bring about a bad outcome, we may form different judgments about x or x's behavior than if that

[21] The smoothie vignettes discussed in §3.2 also count as non-moral cases.

same outcome is simply an accident or the result of (non-willful) ignorance. Our account allows that, in addition, if *x* is responsible for bringing about a good/bad outcome, we may form different judgments about the intentionality of *x*'s action than if that action is *not* good/bad and, in particular, *not* one for which *x* is responsible. Our account is largely neutral regarding the justification of such an influence of evaluative considerations on attributions of intentionality (though we have argued that the primary line of reasoning for the conclusion that this influence is unjustified is problematic). Nevertheless, our account acknowledges the psychological reality of this influence. In effect, it recognizes that there is a *bi*-directional relation between judgments of intentionality and evaluative considerations—in particular, judgments about the responsibility of actors.[22]

We anticipate that this consequence may be met with skepticism. It is tempting to maintain that the relation between intentionality attributions and judgments of responsibility is *uni*-directional: the former influence the latter, but not *vice versa*. But, in light of the observations in §1 and the manipulations reported in §2, this view strikes us as implausible. Indeed, it seems clear that on certain occasions we may attribute responsibility absent a judgment about whether or not an agent acted intentionally, and then reason that because typically an agent who is responsible for her action acted intentionally, the agent acted intentionally. Recall the student, described in §1, who infers that her professor humiliated her intentionally on the grounds that he is criticizable for having done so. In that case, the student is able to (rightly or wrongly) attribute responsibility absent a judgment about whether or not her professor acted intentionally, and then infer from the professor's responsibility to his action's intentionality. Given the relative paucity of direct information regarding whether her professor's action was intentional, it is extremely useful to the student to be able to reason thus.

Of course, an attribution of intentionality which follows from a judgment of responsibility may (and, presumably, in many cases should) be revised in light of further information. Still, the fact that an attribution of responsibility might initially influence an intentionality attribution must be acknowledged. Because of the complexity of folk psychology—in particular, the largely Neurathian character of judgments regarding goodness/badness, responsibility, and intentionality—we believe that skepticism regarding the psychological reality of a bi-directional relation between intentionality attributions and judgments of responsibility is unwarranted. It has been argued that intentionality must play a useful folk psychological role in

---

[22] The implications of this bi-directional relation for jury reasoning have already been noted (e.g. Nadelhoffer, 2006). We believe that it may have important implications for other areas, such as theory of mind, perspective-taking, and other facets of social cognition, as well. For example, research on the 'hostile attribution bias' (Dodge, 1980) may benefit from the recognition that responsibility judgments can influence intentionality attributions—and thus to the extent that some people are more inclined to form responsibility judgments, they may be more inclined to view ambiguous actions as intentional. Finally, it is worth noting that if future research reveals asymmetries in other folk psychological judgments (e.g. causal, dispositional, emotional, or epistemic judgments), then our account may be extended to explain these asymmetries as well. Of course, the success of such extensions must be judged case-by-case.

evaluative judgments, and that this is inconsistent with the view that judgments of responsibility influence intentionality attributions, for this view makes the psychological representation of intentionality a 'pointless mechanism' (Knobe and Mendlow, 2004). But using the presence of responsibility to infer the presence of intentionality on *some* occasions clearly does not make the psychological representation of intentionality a pointless mechanism.[23] On the contrary, it reveals just how complex—and interesting—this element of folk psychology really is.

Department of Psychology
College of Charleston

Department of Philosophy
University of Texas at Austin

References

Adams, F. 1986: Intention and intentional action: the simple view. *Mind & Language*, 1, 281–301.

Adams, F. and Steadman, A. 2004a: Intentional action and moral considerations: core concept or pragmatic understanding? *Analysis*, 64, 173–181.

Adams, F. and Steadman, A. 2004b: Intentional action and moral considerations: still pragmatic. *Analysis*, 64, 264–267.

Adams, F. and Steadman, A. 2007: Folk concepts, surveys, and intentional action. In C. Lumer (ed.). *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy*. Aldershot: Ashgate Publishers.

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

Anscombe, G.E.M. 1957: *Intention*. Ithaca, NY: Cornell University Press.

Bengson, J., Moffett, M. and Wright, J. forthcoming: The folk on knowing how. *Philosophical Studies*.

Bratman, M. 1984: Two faces of intention. *Philosophical Review*, 93, 375 –405.

Dodge, K.A. 1980: Social cognition and children's aggressive behavior. *Child Development*, 51, 162–170.

Duff, R.A. 1982: Intention, responsibility, and double effect. *The Philosophical Quarterly*, 32, 1–16.

Harman, G. 1973: *Thought*. Princeton, NJ: Princeton University Press.

Harman, G. 1976: Practical reasoning. *Review of Metaphysics*, 79, 431–63.

Hauser, M. 2006: *Moral Minds: The Unconscious Voice of Right and Wrong*. New York: Harper Collins.

Knobe, J. 2003a: Intentional action and side effects in ordinary language. *Analysis*, 63, 190–193.

[23] See Nadelhoffer, 2004a for further criticisms of Knobe and Mendlow's argument.

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: 81–187.

Knobe, J. 2005: Theory of mind and moral cognition: exploring the connections. *TRENDS in Cognitive Science*, 9, 357–359.

Knobe, J. 2006: The concept of intentional action: a case study in the uses of folk psychology. *Philosophical Studies*, 130, 203–231.

Knobe, J. 2007: Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31, 90–106.

Knobe, J. and Mendlow, G. 2004: The good, the bad and the blameworthy: understanding the role of evaluative considerations in folk psychology. *The Journal of Theoretical and Philosophical Psychology*, 24, 252–258.

Machery, E. 2008: The folk concept of intentional action: philosophical and experimental issues. *Mind & Language*, 23, 165–189.

Malle, B. 2006: Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, 6, 87–112.

Mallon, R. 2008: Knobe versus Machery: testing the trade-off hypothesis. *Mind & Language*, 23, 247–255.

McCann, H. 1986: Rationality and the range of intention. *Midwest Studies in Philosophy*, 10, 191–211.

McCann, H. 2005: Intentional action and intending: recent empirical studies. *Philosophical Psychology*, 18, 737–748.

Nadelhoffer, T. 2004a: The Butler Problem revisited. *Analysis*, 64, 277–284.

Nadelhoffer, T. 2004b: Blame, badness, and intentional action: a reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology*, 24, 259–269.

Nadelhoffer, T. 2004c: On praise, side effects, and folk ascriptions of intentional action. *The Journal of Theoretical and Philosophical Psychology*, 24, 196–213.

Nadelhoffer, T. 2006: Bad acts, blameworthy agents, and intentional actions: some problems for jury impartiality. *Philosophical Explorations*, 9, 203–219.

Nichols, S. and Knobe, J. 2007: Moral responsibility and determinism: the cognitive science of folk intuitions. *Noûs*, 41, 663–685.

Nichols, S. and Ulatowski, J. 2007: Intuitions and individual differences: the Knobe effect revisited. *Mind & Language*, 22, 346–365.

Phelan, M. and Sarkissian, H. 2008: The folk strike back: or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138, 291–298.

Pizarro, D., Knobe, J. and Bloom, P. 2008: College students implicitly judge interracial sex and gay sex to be morally wrong. Unpublished Manuscript. Available at: http://www.unc.edu/%7Eknobe/pkb_implicit.pdf.

Sosa, E. 2007: Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132, 99–107.

Turner, J. 2004: Folk intuitions, asymmetry, and intentional side effects. *Journal of Theoretical and Philosophical Psychology*, 24: 214–219.

von Wright. G.H. 1983: *Practical Reason: Philosophical Papers, vol. 1*. Oxford: Blackwell.

Wolf, S. 1990: *Freedom Within Reason*. Oxford: Oxford University Press.

Young, L., Cushman, F., Adolphs, R., Tranel, D. and Hauser, M. 2006: Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, 6, 265–278.