Asymmetries in Judgments of Responsibility and Intentional Action*

Jennifer Lyn Wright and John Bengson

Draft—posted 5/4/07

Abstract. Recent experimental research 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, coined the ‘Knobe effect’, which invokes an intuitive asymmetry in assessments of responsibility (e.g., praise and blame) and 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 simultaneously 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 action influence evaluative considerations, but evaluative considerations also influence attributions of intentional action (Knobe, 2003a, 2003b, 2004, 2005; Knobe and Mendlow, 2004; Nadelhoffer, 2004a, 2004b, 2004c, 2006a; McCann, 2005; Young, et al., 2006; Adams and Steadman, forthcoming; Machery, forthcoming; Nichols and Knobe, forthcoming; Nichols and Ulatowski, forthcoming; Phelan and Sarkissian, forthcoming).

While we believe that these sorts of experimental results must be treated with care in a

---

* Earlier versions of this paper were presented at the 2006 Mountain-Plains Philosophy Conference, 2006 SSP Philosophy Conference, and the 2007 Central APA. Thanks to audiences at those events, as well as Marc Moffett, Shaun Nichols, Mark Phelan, George Sher, Ed Sherline, Liane Young, and especially Joshua Knobe for helpful comments and discussion. Thanks also to Christin Covello, Jerry Cullum, and Piper Grandjean for assistance with data collection/entry.

\(^1\) 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 of intending, as defenders of the so-called ‘simple view’ claim (Adams, 1986; McCann, 1986, 2005; for recent critical discussion, see Nadelhoffer, 2006b). Since our primary focus is the explanation of attributions of intentionality, not intention, we shall set this issue to the side here.
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, forthcoming). In what follows, we defend a novel account of the Knobe effect which exploits two factors: an intuitive asymmetry in assessments of responsibility (e.g., praise and blame) and 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 of the Knobe effect. In §2, we provide empirical evidence that strongly supports this account. In §3, we critically discuss several currently prominent alternative accounts. Then, in §4, we consider whether or not the Knobe effect arises in non-moral cases, explaining why there is good reason to think, pace the vast majority of existing accounts, that it does. 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, forthcoming).

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 (see, e.g., Knobe, 2003a, 2003b, 2004, 2005, 2006; Knobe and Mendlow, 2004; Nadelhoffer, 2004a, 2004b, 2004c, 2006a; Malle, 2006). 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. That is, it does not tell us what is it about evaluative considerations that enable them to influence intentionality attributions in an asymmetrical fashion. Our account, which we will explicate in this section and defend in the following sections, is designed to do just this. In short, we believe that there is an asymmetry in assessments of responsibility (e.g., praise and blame) that, because of the putative connection between responsibility and intentionality, generates the Knobe effect.

We begin with the first component of our account, namely, that there is an asymmetry in assessments of responsibility. 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 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, 84). No corresponding requirement appears to hold for negative responsibility.\(^2\)

---

\(^2\) Here and throughout, we restrict ourselves to cases in which \(x\) has a known reason to \(\varphi\) or not \(\varphi\), \(\varphi\)-ing or not \(\varphi\)-ing is an action properly attributed to \(x\) (Scanlon, 1998), 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 assessments of positive/negative responsibility, as well as assessments 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 assessments of praise/blame (laudability/criticizability,
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 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 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,

\[
\text{good/bad action} + \text{intentional action} = \text{positively/negatively responsible actor.}
\]

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

\[
\begin{align*}
A \text{ was good/bad.} \\
A \text{ was done intentionally.} \\
\text{The actor was positively/negatively responsible for } A\text{-ing.}
\end{align*}
\]

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

However, the diagram 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 etc.), the Doctrine of Double Effect and various other considerations suggest that responsibility and praise/blame may come apart. Specifically, assessments of responsibility may sometimes, but not always, manifest in assessments of praise/blame. If this is the case, it would, on our view, remain an open question whether intentionality attributions are being influenced by assessments of responsibility or, more specifically, praise/blame (laudability/criticizability, etc.). Either would be consistent with the account offered here. See note 11 for related discussion.
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 diagram above, the goodness/badness of an action and the responsibility of an actor can be used to defeasibly infer the presence/absence of intentionality.

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 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 seems to her to provide the ‘best total explanation’ (Harman, 1973) of his criticizability. The student thus concludes,

---

3 This inference is most likely inductive or abductive in nature. In effect, the analogy with a simple mathematical inference is merely an analogy. Plausibly, the ‘+’ and ‘=’ in the diagram, for instance, ought to be interpreted as marking transitions in what Harman (1973) calls ‘inference to the best total explanation’ or some other non-deductive inference.
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 (criticizable, etc.), 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 (laudable, etc.), 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 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 3), it might 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, such a 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.

![Diagram](image)

Figure 2: Majority reasoning in HARM.

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

Figure 3: Majority reasoning in HELP.

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

i. assessments 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 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% vs. 44%; $\chi^2(121) = 3.7, p = .054, \phi = .18$. HELP: 19% vs. 2%; $\chi^2(120) = 9.8, p = .002, \phi = .29$). 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% vs. 29%; $\chi^2(122) = 8.9, p = .003, \phi = .27$. HELP: 22% vs. 1%; $\chi^2(121) = 17.5, p < .001, \phi = .38$). 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% vs. 78%; $\chi^2(122) = .72, p > .05, \phi = -.08$. HELP: 4% vs. 8%; $\chi^2(121) = .59, p > .05, \phi = -.07$). In short, participants’ judgments of goodness/badness were not significantly associated with their intentionality attributions, whereas assessments of both praise and blame were.

Further evidence that assessments of responsibility played the central role comes from considering partial correlations between assessments of badness, blame, and intentionality. In HARM, when the variance explained by assessments of badness was controlled for, assessments 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 assessments of blame was controlled for, assessments 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% vs. 44%; $\chi^2(121) = 18.4, p < .001, \phi = .39$). It is just that judgments of badness became relevant to intentionality attributions only when coupled with assessments of blame.
These results provide strong support for our account. Because of the asymmetry in assessments 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, an assessment 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 assessments of blame than praise.

Our account predicts that a similar asymmetry would result from more frequent assessments of praise than blame: specifically, when participants more frequently praise than blame, participants’ intentionality attributions will likewise be asymmetrical. To test this prediction, we conducted a second study in which 59 participants were given the following two scenarios, modeled after cases described by Mele and Sverdlik (1996):

DOCHARM: A patient is suffering from a potentially terminal disease. A doctor judges that death is very likely, but that an operation has some chance of saving the patient’s life. The doctor also knows that the operation itself has a good chance of killing the patient. The doctor operates and this kills the patient.

DOCHELP: A patient is suffering from a potentially terminal disease. A doctor judges that death is very likely, but that an operation has some chance of saving the patient’s life. The doctor also knows that the operation itself has a good chance of killing the patient. The doctor operates and this saves the patient.

Participants were asked (a) whether the doctor’s action was good, bad, or neither, (b) whether the

---

While the majority of participants (62%) made asymmetrical intentionality attributions, there was of course a minority (38%) who did not. Nichols and Ulatowski (forthcoming) suggest that this indicates that there are multiple correct interpretations of ‘intentional’ (cf. Sosa, 2007). This ‘interpretative diversity hypothesis’ acknowledges that the majority response demonstrates 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 (Bengson, Moffett, and Wright, 2007), we are content to note that since our account provides an explanation of the asymmetry in intentionality attributions demonstrated by the majority response, an explanation that remains neutral on the question of whether or not there is a single correct interpretation of ‘intentional’, it is entirely consistent with Nichols and Ulatowski’s suggestion.
doctor deserved any praise, blame, or neither for that action, and (c) whether or not the doctor intentionally brought about the patient’s death/saved the patient.\(^5\)

In DOCHARM, while nearly half of the participants (44%) judged that the death of the patient was bad, none judged that the doctor deserved blame and almost none judged that the doctor brought about the death of the patient intentionally (2%). DOCHHELP elicited very different judgments: most participants (83%) judged that saving the patient was good, that the doctor deserved praise (78%), and that the doctor saved the patient intentionally (73%). Most of the intentionality attributions in DOCHHELP involved judgments of both a good action and a praiseworthy doctor: participants were significantly more likely to judge that the doctor acted intentionally when they stated both that the action was good and that the doctor was praiseworthy than when they only agreed to one or neither of these (83% vs. 48%; \(\chi^2(61) = 8.4, p = .004, \varphi = .37\)). Again, a close look at the data reveals which of these two judgments played the central role. Participants were significantly more likely to judge that the doctor acted intentionally when they stated that the doctor was praiseworthy than when they did not (81% vs. 39%; \(\chi^2(61) = 9.3, p = .002, \varphi = .39\)). Although they were also significantly more likely to judge that the doctor acted intentionally when they stated that the action was good than when they did not (80% vs. 46%; \(\chi^2(60) = 5.4, p = .02, \varphi = .30\), when the variance explained by assessments of praise was controlled for, the association between judgments of goodness and intentionality attributions disappeared (partial \(r = .17, p > .05\), demonstrating the centrality of assessments of praise. Still, participants were significantly more likely to praise the doctor when they considered the doctor’s action to be good than when they did not (86% vs. 46%; \(\chi^2(60) = 8.6, p = .003, \varphi = .38\)). This suggests that judgments of goodness became relevant to intentionality attributions only when coupled with assessments of praise.

These results confirm our prediction: when participants more frequently praise than blame, participants’ intentionality attributions are likewise asymmetrical. Because participants more frequently praised the DOCHHELP doctor than blamed the DOCHARM doctor, participants more frequently ascribed intentionality to the DOCHHELP doctor’s action than to the DOCHARM doctor’s action.

Of course, the statistical analyses we employed in the above studies provide results that are strictly speaking neutral regarding the directionality of the relation between assessments of responsibility and intentionality attributions. Nevertheless, these analyses establish that there is a

---

\(^5\) Obviously, DOCHARM/DOCHHELP does not exactly mirror HARM/HELP. But since the point of DOCHARM/DOCHHELP was simply to test the above prediction by determining the presence or absence of certain statistical relations between assessments of positive responsibility (praise) and intentionality attributions, this is incidental to the purpose of the study.
strong relation between these judgments. Given the conceptual points advanced in §1 concerning
the asymmetry in assessments of responsibility, on the one hand, and the putative connection
between responsibility and intentionality, on the other, we have good reason to believe that in
both HARM/HELP and DOCHARM/DOCHHELP intentionality attributions were influenced by
assessments of responsibility, and not the other way around.⁶ Because the conditions under which
assessments 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. So, while the empirical
data is not by itself sufficient to establish that assessments of responsibility led to an asymmetry
in intentionality attributions in these cases,⁷ these data together with the conceptual points
advanced in §1 provide compelling reasons to believe that this is so.⁸

3. Alternative Accounts of the Knobe Effect

We have thus far articulated an 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 consider the view that judgments of badness explain the Knobe effect,
the view that affective bias generates the Knobe effect, and the view that the Knobe effect is
merely due to conversational pragmatics.

3.1 The Badness Account

Knobe and Mendlow (2004) and Phelan and Sarkissian (forthcoming) have reported preliminary
research which suggests that a judgment of blameworthiness is not generally required to elicit
intentionality attributions. In two pilot studies (⁸N < 25), participants were given the following

⁶ We should note, in addition, that a study (in progress) involving a variety of manipulations provides
prima facie empirical evidence that psychological representations of responsibility may in fact influence
psychological representations of intentionality. In this study (⁸N = 86), a variation in input regarding
blameworthiness resulted in a corresponding variation in output regarding intentionality in the majority of
participants (⁸76%).

⁷ More generally, they could not establish that any factor, including judgments of badness (see §3.1) or
cost (see note 14), influence intentionality attributions, and not the other way around. So, all extant
accounts are subject to the worry expressed here. We hope that future research, such as the research
reported in note 6, will address this issue.

⁸ 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 cases:
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, forthcoming; and §3.1 below), and the free cup and extra dollar cases (see
Machery, forthcoming; and note 11 below), among others.
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.’

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., 2007).

We believe that 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), pace 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 were not significantly associated with their intentionality attributions. Moreover, when the variance explained by assessments of blame was controlled for, judgments of badness and intentionality became negatively correlated. These
results clearly demonstrate the empirical inadequacy of the badness account.9

Let us return, then, to the results of DECREASE, which might be interpreted as challenging our account. There are good reasons to think that they do not. For one, our account does not offer a generally necessary condition for intentionality attributions.10 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.11

Indeed, it is not difficult to understand why DECREASE elicited intentionality attributions in the absence of assessments 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, 

---

9 Indeed, Knobe (forthcoming) now retracts the badness view partly in response to an earlier version of the present paper.
10 In this and other ways, our account differs markedly from the view that Phelan and Sarkissian (forthcoming) refer to as ‘BLAM’.
11 Machery (forthcoming) presents very interesting findings which also might be interpreted as challenging our account. In his ‘free cup’ case, the majority of participants judged that the actor did not receive a free cup intentionally, where receiving the free cup was an unintended consequence of the actor’s purchase of a smoothie. In the ‘extra dollar’ case, the majority of participants judged that the actor did pay an extra dollar intentionally, where paying an extra dollar was required in order to purchase the smoothie. In neither case did the majority judge that the actor was praiseworthy or blameworthy for the relevant action. Machery concludes that assessments of praise/blame cannot explain the Knobe effect. This research is quite provocative. Nevertheless, we are not convinced that it challenges our account. For one, while receiving a free cup may be a side-effect of the intended outcome, it is not clear that paying an extra dollar is: the latter, but not the former, was done in order to receive the smoothie (see Knobe, 2006 for a discussion of the relevance of the phrase ‘in order to’ to intentionality attributions). Further, since in some cases responsibility may be present in the absence of praise/blame (see note 2), it remains possible that the asymmetry in intentionality attributions is due to an asymmetry in assessments of responsibility, in which case Machery’s findings would be entirely consistent with our account. To test this hypothesis, we conducted a study in which 78 participants were given the free cup and extra dollar cases and asked to judge (a) whether the actor should be held responsible for the relevant outcome and (b) whether the actor performed the relevant action intentionally. 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. In line with our account, in both cases participants’ judgments about responsibility and intentionality were significantly positively correlated (free cup: φ = .35, p = .002; extra dollar: φ = .21, p = .073).
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, a 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:

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% vs. 41%; \( \chi^2(120) = 7.4, p = .007, \phi = .25 \). INCREASE: 47% vs. 10%; \( \chi^2(121) = 13.6, p < .001, \phi = .34 \)). 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% vs. 34%; \( \chi^2(120) = 14.9, p < .001, \phi = .35 \). INCREASE: 41% vs. 11%; \( \chi^2(121) = 10.8, p = .001, \phi = .30 \)). 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% vs. 43%; \( \chi^2(120) = 3.5, p > .05, \phi = .17 \). INCREASE: 17% vs. 11%; \( \chi^2(121) = .70, p > .05, \phi = .08 \)). 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% vs. 18%; $\chi^2(120) = 49.8, p < .001, \phi = .64$. INCREASE: 18% vs. 5%; $\chi^2(121) = 3.3, p = .069, \phi = .17$). This indicates that judgments of goodness/badness became relevant to intentionality attributions only when coupled with assessments 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, DECREASE2/INCREASE actually support it.

### 3.2 The Bias Account

In a recent discussion of the potential implications of the Knobe effect for the problem of jury impartiality, Nadelhoffer (2006a) has proposed that the Knobe effect is due to an affect-driven bias. Nadelhoffer concedes the central contention of our account—that assessments 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 (2006a).

In explicating this view, which we will call the bias account, Nadelhoffer invokes Alicke’s (2000, 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 assessment of the actor and his action]’ (2006a). As a result, participants in HARM, for instance, are led to attribute intentionality as a result of their affect-driven attribution of blame.12 Nadelhoffer concludes, ‘even though moral considerations surely do act expansively on folk ascriptions of intentional action…ideally they ought not have this effect’ (2006a).

---

12 Incidentally, because Alicke’s model is restricted to blame, the bias account seems unable to explain the results of HARM/HELP and DOCHARM/DOCHELP, in which participants’ attributions of intentionality were significantly correlated with assessments of praise. More generally, adopting Alicke’s model seems inconsistent with Nadelhoffer’s own view, expressed in his (2004b), that both praise and blame influence intentionality attributions.
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, we believe that 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 assessments 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 (forthcoming) suggests that intentionality attributions are influenced by evaluative considerations even in the absence of an 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 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 (2006a) 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 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 (2006a) 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, reflection on 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. We typically have reason to not speed off (but instead to cooperate) when approached by a police officer, whereas we have no such reason when approached by a thief. Thus, while the thief had reason to not speed off (but instead to cooperate), the driver had no such reason—and, plausibly, had reason to speed off (and/or otherwise refuse to cooperate). What is more, it is reasonable to assume that both drivers were aware that they possessed or lacked these reasons: the thief knew he had reason not to speed off, and the driver knew he had no such reason.

Interestingly, this suggests that while THIEF and DRIVER are not relevantly similar, THIEF and HARM may be. For in both THIEF and HARM, the actors knowingly brought about a side-effect which they had reason to not bring about. They were, consequently, blameworthy for having so acted. Presumably, this led participants to make intentionality attributions in THIEF as in HARM: a judgment of a blameworthy actor and a bad action typically leads to an intentionality attribution. On the other hand, DRIVER involves cognitive and conative considerations more in line with DOCHARM. For in both DRIVER and DOCHARM, the actors had reason to engage in the action which they knowingly performed. They were, consequently, not blameworthy for having so acted. Presumably, it was because participants did not judge the actors to be blameworthy that they did not judge that the actors acted intentionally in these cases.

As this makes clear, the asymmetry in participants’ assessments 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 no reason to believe that the Knobe effect is due to an affect-driven bias.

3.3 The Pragmatic Account

We now turn to what we will call the pragmatic account, which appeals to pragmatic connections between assessments 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 (Adams and Steadman, 2004a, 2004b, 2006; Wasserman, 2007). On this view, the Knobe effect is
the result of participants’ desire to avoid an unwanted implicature associated with the denial of intentionality. Roughly, the idea is that while participants’ intentionality attributions are strictly speaking false, they convey something true: e.g., that the actor is properly considered positively/negatively responsible for the side-effects of his/her action. Were participants to state that the actor’s action was not intentional, this would imply that the actor is not properly considered positively/negatively responsible for those effects. Because participants desire to imply no such thing, they are led to attribute intentionality where none is called for.

One way to test this hypothesis would be to remove the tension created by this alleged implicature. As Adams and Steadman (forthcoming) and Wasserman (2007) both suggest, a study could be run in which participants are given an alternative way to respond to scenarios such as HARM/HELP. For example, participants could have the option to state that the relevant action was not intentional, but performed with full awareness of the consequences. This would allow participants to imply that the actor was properly considered positively/negatively responsible for the effects of the action, without feeling forced to state that the actor brought about those effects intentionally. Consequently, such a study would enable us to test the empirical adequacy of the pragmatic account.

Yet in just such a (pilot) study, conducted by Adams and Steadman (forthcoming), 80% of the participants stated that the chairman in HARM harmed the environment ‘knowingly and intentionally’ while only 20% stated that he acted ‘knowingly, but not intentionally’. (Similarly, in the studies discussed in §2 and §3.1, participants attributed intentionality even though it would have been quite easy for them to assess the actor as positively/negatively responsible while denying that the actor performed the relevant action intentionally.) So, even when participants were given the opportunity to assess the actor as positively/negatively responsible for the effects of the relevant action without stating that the actor brought about those effects intentionally, participants continued to respond as they did in earlier studies—namely, with an intentionality attribution. These results disconfirm the predictions of proponents of the pragmatic account. (See also the results reported in Knobe (2004), Nadelhoffer (2006a), and Nichols and Ulatowski (forthcoming) for further empirical evidence against pragmatic accounts.)

Adams and Steadman (forthcoming) have responded that such results do not challenge their account, but are simply evidence for the insidiousness of participants’ ‘pragmatic programming’.13 We find this response unconvincing. Of course, it does not follow from anything
we have said here that participants’ intentionality attributions in Knobe effect cases are justified. Still, we have argued (in §3.2) that the main line of reasoning in favor of the view that participants’ intentionality attributions in these cases is unjustified is unmotivated. In the absence of some other reason to think that these attributions are unjustified, we lack reason to convict participants in the relevant studies of widespread error. Consequently, positing a malfunction in participants ‘programming’ to account for the empirical inadequacy of the pragmatic account is, at this point, unacceptably ad hoc.

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 each identifies a factor that is 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.14 And the bias and pragmatic accounts appear to be correct in claiming that assessments 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, 2004a, 2004b, 2006a; Malle, 2006; Pizarro, et al., 2007; Adams and Steadman, forthcoming), Turner (2004) has argued that the Knobe effect could be elicited by non-moral considerations as well. He invites us to consider two putatively non-moral scenarios in which a producer knowingly incurs criticism/praise from the media as a
side-effect of making a profitable movie. The producer ‘doesn’t care at all’ about such criticism/praise: he ‘just wants to make as much profit as [he] can’. Turner predicts that the folk will judge that the producer incurred the criticism intentionally, but did not incur the praise intentionally (cf. Wasserman, 2007). If this (yet untested) prediction is correct, then insofar as these scenarios do not involve moral considerations, the asymmetry which they elicit in intentionality attributions cannot be explained by appealing to the alleged influence of judgments regarding the moral status of actions and/or actors on intentionality attributions.

Since our account does not invoke specifically moral considerations, it—unlike most other explanations of the Knobe effect—sits happily with this prediction. We believe that assessments of positive/negative responsibility, though evaluative through and through, can be both moral and non-moral. Since all assessments of positive/negative responsibility, whether moral or not, are subject to the asymmetry discussed in §1, it would follow that it is possible to elicit the Knobe effect even in the absence of moral considerations.

Recall DECREASE2/INCREASE, in which participants’ assessments of responsibility influenced their intentionality attributions. 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.

So construed, the results of DECREASE2/INCREASE (reported in §3.1) support Turner’s hypothesis 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 or moral blame) on intentionality attributions are unable to explain why or how this is so. Yet, while these results pose a serious challenge to other accounts of the Knobe effect, they offer further support in favor of our account.

5. Conclusion

For instance, an athlete may be praised or criticized for his performance. An assessment 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 contribution is being evaluated.

The free cup and extra dollar cases discussed in note 11 also count as non-moral cases.
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 assessments 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 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. Although our account is largely neutral regarding the justification of such an influence of evaluative considerations on attributions of intentionality, it 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.

We are well aware that this consequence will be met with skepticism. It is tempting to maintain that the relation between intentionality attributions and assessments of responsibility is uni-directional: the former influence the latter, but not vice versa. But this view strikes us as implausible, for 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 §3, 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 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 an assessment 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 assessments of responsibility is unwarranted. It has been argued that intentionality must play a useful folk psychological role in evaluative judgments, and that this is inconsistent with the view that assessments of responsibility influence intentionality attributions, for this view makes the psychological representation of intentionality a ‘pointless mechanism’ (Knobe and Mendlow, 2004). But, as we hope to have shown, using the presence of responsibility to infer the presence of intentionality on some occasions does not make the psychological representation of intentionality a pointless mechanism. On the contrary, it reveals just how complex—and interesting—this element of folk psychology really is.

References
Bengson, J., M. Moffett, and J. Wright. 2007: The folk on knowing how. Unpublished manuscript. Available at: webspace.utexas.edu/jtb538/Research.htm.

17 See Nadelhoffer (2004a) for further criticisms of Knobe and Mendlo’s argument.


Phelan, M. and H. Sarkissian. Forthcoming: The folk strike back: Or, why you didn’t do it intentionally, though it was bad and you knew it. *Philosophical Studies*.


