Asymmetries in Folk Judgments of Responsibility and Intentional Action[*]

Jen Wright[†] and John Bengson[‡]

Draft—9/15/06

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, it is appropriate to 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 other words, it makes sense that whether or not a given action is intentional matters to us when we assess the action's or actor's evaluative status.

This uni-directional relation between judgments of intentionality[1] and judgments about the goodness/badness of actions or responsibility (e.g., praiseworthiness/blameworthiness) of actors seems relatively straightforward. What is surprising is that recent experimental research strongly suggests a *bi*-directional relation between these judgments in the "folk". 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.[2]

While we believe that these sorts of experimental results must be treated with care in a philosophical setting, it remains incumbent on philosophers of action and moral psychologists to provide a descriptively correct understanding of this phenomenon, coined the "Knobe effect" (Nichols & Ulatowski, 2006). In what follows, we explore the Knobe effect in an attempt to provide a conceptually and empirically sound understanding of the relation between evaluative considerations and folk attributions of intentionality. Specifically, we present the results of three new studies that, we shall argue, provide reasons to (1) reject several currently prominent accounts of the Knobe effect and (2) accept an account which exploits two factors: an asymmetry

[*] Thanks to Christin Covello, Jerry Cullum, and Piper Grandjean for assistance with data collection/entry and Marc Moffett, Shaun Nichols, Mark Phelan, George Sher, Ed Sherline, and especially Joshua Knobe for comments and discussion.

[†] University of Wyoming, Departments of Psychology and Philosophy: *narvik* (at) *uwyo* (dot) *edu*.

[‡] University of Texas at Austin, Department of Philosophy: *jsteele* (at) *mail* (dot) *utexas* (dot) *edu*.

[1] This use of the term 'intentionality' differs from that found in philosophy of mind. Here, it signifies a particular property of actions—namely, the property of having been done *intentionally* (i.e., intentional action). It remains an open question whether intentional action requires the specific mental state of *intending*, as defenders of the so-called "simple view" claim (see, e.g., Adams, 1986; cf. McCann, 2005 and Wasserman, 2006). Since our primary focus is the explanation of folk attributions of intentionality, not intention, we shall set this issue aside.

[2] See, e.g., Adams & Steadman (forthcoming); Knobe (2003a, 2003b, 2004, 2005); Knobe & Mendlow (2004); Machery (2006); McCann (2005); Nadelhoffer (2004a, 2004b, forthcoming); Nichols & Knobe (forthcoming); Nichols & Ulatowski (2006); Phelan & Sarkissian (forthcoming).

between assessments of positive and negative responsibility (e.g., praise vs. blame) and the fact that intentionality commonly connects the goodness/badness of actions to the responsibility of actors.

In §1, we articulate what we take to be the most plausible explanation for the so-called Knobe effect—what we will call the *two-factor account*, so-named because it exploits the two factors mentioned above. In §2, we provide empirical evidence that strongly supports the two-factor account. In §3, we critically discuss several currently prominent alternative accounts of the Knobe effect, arguing that each is empirically or conceptually inadequate. 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, with a brief discussion of how the two-factor account may explain in a systematic way the appeal of alternative accounts while avoiding their pitfalls.

1. The 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 & Ulatowski, 2006). But, to many, this 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).

A popular explanation for this asymmetry posits the bi-directional relation between evaluative considerations and intentionality attributions described above (e.g., Knobe, 2003a,

2003b, 2004, 2005; Knobe & Mendlow, 2004; Malle, 2006; Nadelhoffer, 2004a, 2004b, forthcoming). While this response is well-motivated, it seems to us that an appeal to a bi-directional relation alone cannot explain the asymmetry. For such an appeal does not by itself explain *how* evaluative considerations 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 view, which we will explicate in this section and defend in the following sections, is designed to answer this question in a descriptively accurate manner.[3] In short, we believe that there is an asymmetry in judgments of positive and negative responsibility (e.g., praise vs. blame) that, because of a putative connection between responsibility and intentionality, generates the Knobe effect.

We begin with the first component of our view, which consists in the claim that there is an asymmetry in judgments of responsibility. Intuitively, people are to some extent blameworthy or criticizable—that is, *negatively responsible*—when they engage in an action that they know will generate a bad outcome, even if that bad outcome is simply a side-effect of an intended outcome. On the other hand, people are not praiseworthy or laudable—that is, *positively responsible*—for bringing about a good outcome only as a side-effect, even if that side-effect is foreseen. In order to be to some extent positively responsible, intuitively a person must do the right thing *for the right reasons*. No corresponding requirement appears to hold for negative responsibility.[4]

This asymmetry clearly emerges 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, the chairman presumably knew that he had a reason not to implement the new program—namely, that it would harm the environment—yet he still implemented the

[3] See Machery (2006) for a defense of the view that the primary aim of philosophers of action and moral psychologists should be to provide a descriptively accurate understanding of the Knobe effect.

[4] Here and throughout, 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* (Scanlon, 1998), and *x* is normal (i.e., possesses the general capacities presupposed by agency). We also ignore complications that may arise from the Doctrine of Double Effect (see note 5). Given these restrictions, we can say with a bit more precision that *x* is to some extent negatively responsible (e.g., blameworthy or criticizable) for φ-ing iff *x* fails to appropriately respond to reason(s) to not φ, and *x* is to some extent positively responsible (e.g., praiseworthy or laudable) for φ-ing iff *x* appropriately responds to reason(s) to φ. The asymmetry noted in the text becomes apparent when we observe that while *x* fails to appropriately respond to reason(s) to not φ iff (i) *x* φ-s and (ii) *x* has reason(s) to not φ, *x* appropriately responds to reason(s) to φ iff (a) *x* φ-s, (b) *x* has reason(s) to φ, and (c) *x* φ-s <u>because</u> *x* has reason(s) to φ. Of course, one might hold that praise is even more demanding than this: e.g., perhaps it must be the case that, in addition, (d) φ-ing is especially difficult—or supererogatory—for *x*. But this further debate does not affect the present point, since the necessity of (d) would only reinforce our suggestion that praise requires more than blame, thereby giving rise to an asymmetry in assessments of positive and negative responsibility. Thanks to George Sher for discussion on these points.

program anyway. On the other hand, in HELP the chairman failed to bring about a good outcome—namely, helping the environment—for the right reasons, thereby making praise seem unwarranted.

This brings us to the second component of our view: namely, the putative connection between responsibility and intentionality. Typically, people who intentionally act to bring about a bad outcome are negatively responsible and people who intentionally act to bring about a good outcome are positively responsible. In this way, intentionality commonly connects the evaluative status of actions to the responsibility of actors.[5] We can diagram this connection in the following manner: typically,

good/bad action + intentionality = positive/negative responsibility.

Normally, determining both goodness/badness of action and presence/absence of intentionality is required to infer positive/negative responsibility.[6] But this diagram makes it clear how other inferences are possible. For just as we can solve the equation "2 + ? = 9" for the answer "7", it is possible to solve the equation "good/bad action + ? = positive/negative responsibility" for the answer "intentionality".

(Of course, the analogy with a simple mathematical inference is merely an analogy. The '+' and '=' in the diagram, for instance, ought to be interpreted somewhat liberally, perhaps as marking transitions in what Harman (1973) calls "inference to the best total explanation". This is simply a reflection of the fact that the inference in the intentionality case is potentially more complicated than standard, simple mathematical inferences. But while the math analogy makes

[5] While intentionality may commonly connect the goodness/badness of actions to the responsibility of actors, it is clearly not necessary: other factors, like negligence or willful ignorance, can also connect them. Incidentally, while our discussion runs together judgments of responsibility and assessments of praise/blame, the Doctrine of Double Effect suggests that responsibility and praise/blame may come apart. 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 praise/blame. (The results of Machery's (2006) smoothie case suggest that it is the former; but, of course, further research is needed.) Either would be consistent with the coarse-grained version of the two-factor account presented here. Thanks to Ed Sherline for discussion.

[6] Indeed, the standard interpretation of the diagram above is that intentionality attributions influence assessments of responsibility, but not *vice versa*. Commitment to this standard interpretation has motivated Knobe (2003a, 2003b) and Knobe & Mendlow (2004) to posit an asymmetry in folk judgments of goodness/badness in order to explain the Knobe effect (see §3.1 for discussion). Nichols & Ulatowski (2006), on the other hand, posit an "interpretative diversity" in the term 'intentional' to explain the Knobe effect (see §5 for discussion). Pragmatic accounts (discussed in §3.3) also locate the asymmetry in the term 'intentional' (*via* associated implicatures). We disagree with all of these approaches; in the absence of compelling conceptual and empirical evidence to the contrary, the most straightforward explanation of the Knobe effect would appear to be one that posits an asymmetry in a place where, intuitively, one already exists: namely, in assessments of positive/negative responsibility. (Nadelhoffer (forthcoming) also locates the asymmetry here; but, as we argue in §3.2, he mistakenly construes it as an affect-driven bias.)

the target inference look clean and simple, it need not be in order to be psychologically realistic, or even rational.)

Presumably, the inference to intentionality is driven by the putative relation between responsibility and intentionality. Given that actors are typically held to be positively/negatively responsible for actions which are (morally or non-morally) good/bad and only if they act intentionally, the responsibility of actors—and, derivatively, the goodness/badness of actions—serves as an indicator for intentionality.[7] As a result, goodness/badness of action and responsibility of actor can be used to determine the presence/absence of intentionality. Of course, the inference to intentionality is defeasible (and, again, most likely proceeds in a non-deductive fashion, *à la* Harman). Nevertheless, if someone who knowingly acts to bring about a (bad) outcome is blameworthy, it is possible to infer that he/she brought about that outcome intentionally. Likewise, if someone who knowingly acts to bring about a (good) outcome is praiseworthy, it is possible to infer that he/she brought about that outcome intentionally.

It is tempting to maintain that this inference, like the fallacy of affirming the consequent, is not philosophically defensible. In §§3.2-3.3, we consider two accounts which give in to this temptation. Though it is not our goal to establish that this inference is in fact justified, it is worth noting that it has been argued that to the extent that one knowingly *A*-s and is properly held responsible for *A*-ing, one *A*-s intentionally (Bratman, 1984; Duff, 1982; Harman, 1976). If this view is correct, then the inference to intentionality may be justified.[8]

Whether or not it is philosophically defensible, the inference clearly emerges in HARM and HELP. Viewing the HARM chairman as blameworthy for a bad outcome inclines one to say that he acted intentionally, whereas viewing the HELP chairman as not praiseworthy not only disinclines one to say that he acted intentionally, it actually creates a reason not to 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 we judge him to be not praiseworthy, it follows by a *modus tollens*-esque inference—again, perhaps an inference to the best total explanation—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 (both of which, as noted above, may be philosophically defensible):

---

[7] Given that responsibility typically implies foreknowledge of the relevant outcome, presumably such foreknowledge, like goodness/badness, would to some extent derivatively serve as an indicator of intentionality. Thanks to Mark Phelan for discussion on this point.

[8] Of course, it is possible to understand the folk judgments in question as providing support for the philosophical view, rather than the philosophical view as providing support for the folk judgments.

i. assessments of positive/negative responsibility are asymmetrical;

ii. intentionality commonly connects the evaluative status of actions to the responsibility of actors, the latter of which alone typically implies intentionality.

Once we note that a judgment of a responsible actor and, derivatively, a good/bad action typically generates an intentionality attribution we are able to account for the Knobe effect in a relatively straightforward manner. Participants in these 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 an attribution of intentionality in HARM but not in HELP. Since it exploits the aforementioned two factors in folk judgments of intentionality, let us call this the *two-factor account* of the Knobe effect.

2. Empirical Support for the Two-factor Account

We conducted two studies designed to test the empirical adequacy of the two-factor account. In the first study, 122 participants were given slightly revised HARM and HELP scenarios. In addition to being asked whether or not the chairperson[9] acted intentionally, participants were asked: (a) whether the harming/helping of the environment was "good", "bad", or "neither" and (b) whether the chairperson deserved any "praise", "blame", or "neither" for harming/helping the environment.

Most of the (many) intentionality attributions in HARM accompanied judgments of both a *bad* action and a *blameworthy* chairperson; likewise, most of the (few) intentionality attributions in HELP accompanied judgments of both a *good* action and a *praiseworthy* chairperson. In both cases, participants were significantly[10] 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.[11] 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 he/she was praiseworthy/blameworthy than when they did not.[12] However, in both cases, participants were no more likely to judge that the chairperson acted

[9] 'Chairperson' was substituted for 'chairman' to eliminate the possibility of gender bias.

[10] Throughout, 'significant' denotes *statistical* significance. Traditionally, a given relation r is statistically significant iff r possesses a p value of less than .05 ($p < .05$), which means that there is more than a 95% chance that the relation is genuine (i.e., is true of the general relevant population, and not merely a quirk of the actual data sample).

[11] HARM: 68% vs. 44%; $\chi^2(121) = 3.7$, $p = .054$, $\varphi = .18$. HELP: 19% vs. 2%; $\chi^2(120) = 9.8$, $p = .002$, $\varphi = .29$.

[12] HARM: 69% vs. 29%; $\chi^2(122) = 8.9$, $p = .003$, $\varphi = .27$. HELP: 22% vs. 1%; $\chi^2(121) = 17.5$, $p < .001$, $\varphi = .38$.

intentionally when they stated that the action was good/bad than when they did not.[13] In line with this, Pearson chi-squares showed that participants' judgments of goodness/badness were not significantly associated with their intentionality attributions, whereas assessments of *both* praise and blame were.

Further support for the claim that assessments of responsibility play 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 (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. Of course, this does not mean that judgments of badness played no role at all; participants were significantly more likely to blame the chairperson when they considered his/her action to be bad than when they did not.[14] 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 the two-factor 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, which alone typically implies intentionality, explains why participants were 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 generate 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 his/her action, they ascribed intentionality. In this way, assessments of positive/negative responsibility led to an asymmetry in intentionality attributions in HARM/HELP.

Further support for the two-factor account—this time in the form of evidence that assessments of *praiseworthiness* influence intentionality attributions—comes from 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

[13] HARM: 63% vs. 78%; $\chi^2(122) = .72$, $p > .05$, $\varphi = -.08$. HELP: 4% vs. 8%; $\chi^2(121) = .59$, $p > .05$, $\varphi = -.07$.
[14] 92% vs. 44%; $\chi^2(121) = 18.4$, $p < .001$, $\varphi = .39$.

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.

In addition to being asked whether or not the doctor intentionally brought about the patient's death/saved the patient, participants were asked: (a) whether the doctor's action was "good", "bad", or "neither" and (b) whether the doctor deserved any "praise", "blame", or "neither" for that action.[15]

In DOCHARM, while nearly half of the participants (44%) stated that the death of the patient was bad, none stated that the doctor deserved blame and very few stated that the doctor brought about the death of the patient intentionally (2%). DOCHELP elicited very different judgments: most participants (83%) stated 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 DOCHELP 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.[16] 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 he/she was praiseworthy than when they did not.[17] 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,[18] when the variance explained by assessments of praise was controlled for, the association between judgments of goodness and intentionality attributions disappeared, demonstrating the centrality of praise assessments. Nevertheless, once again, participants were significantly more likely to praise

[15] Obviously, DOCHELP does not exactly mirror HELP because in DOCHELP the doctor plausibly intends to save the patient's life, and so it may not be considered a side-effect of his action. Indeed, it is difficult—if not impossible—to craft a case in which an actor does not intend to bring about a good side-effect for which he/she is properly held positively responsible. Cases like DOCHELP do not mirror HELP for the reason cited in §1: praise requires knowingly bringing about a good outcome *for the right reasons*. That said, the purpose of DOCHARM/DOCHELP was simply to determine the presence or absence of certain statistical relations between assessments of positive/negative responsibility and intentionality attributions. Despite the fact that DOCHARM/DOCHELP differ from HARM/HELP in several ways, the statistical relations we found (reported below) clearly support the two-factor account.

[16] 83% vs. 48%; $\chi^2(61) = 8.4$, $p = .004$, $\varphi = .37$.

[17] 81% vs. 39%; $\chi^2(61) = 9.3$, $p = .002$, $\varphi = .39$.

[18] 80% vs. 46%; $\chi^2(60) = 5.4$, $p = .02$, $\varphi = .30$.

the doctor when they considered his/her action to be good than when they did not,[19] suggesting that judgments of goodness became relevant to intentionality attributions only when coupled with assessments of praise. Thus, in line with the two-factor account, judgments of responsibility and, derivatively, goodness/badness explain the asymmetry in intentionality attributions in these cases.

3. Alternative Accounts of the Knobe Effect

We have thus far articulated the two-factor account and presented empirical research which confirms it. In this section, we critically discuss several currently prominent alternative accounts (see note 6 for a rough taxonomy of positions in the debate). 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 an assessment of the positive/negative responsibility of actors is not generally required to elicit intentionality attributions. In two pilot studies ($N < 25$), participants were given the following vignette, which we will call DECR:

> DECR: 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 negative responsibility of the actor is generating participants' intentionality attributions. Knobe and Mendlow (2004) conclude that the perceived moral status of actions—*not* the responsibility of actors—influenced participants intentionality attributions.[20] Specifically, they claim that the

[19] 86% vs. 46%; $\chi^2(60) = 8.6$, $p = .003$, $\varphi = .38$.

[20] Knobe and Mendlow (2004) further argue, on *a priori* grounds, that intentionality must play a useful (folk psychological) role in moral judgments, and that this is inconsistent with the view that assessments of responsibility influence intentionality attributions, since this view makes the concept of intentionality a "pointless mechanism". Because of the complexity of folk psychology, we find this argument unconvincing, though we will not address it here. We refer the interested reader to Nadelhoffer's (2004a) criticisms of this argument.

perceived moral badness of the relevant action—decreasing sales in New Jersey—influenced participants' intentionality attributions. On this view, which we will call the *badness account*, judgments of moral badness, but not goodness (nor positive/negative responsibility), lead participants to attribute intentionality (see also Knobe, 2003a, 2003b; Pizarro et al., 2006).

The badness account faces a number of difficulties. First, it is a one-factor account: it attempts to explain the Knobe effect by appealing only to judgments of moral badness. As suggested by the discussion in §1, one-factor accounts are generally problematic because they appear unable to explain exactly *how* the relevant considerations lead to an asymmetry in intentionality attributions. That is, they fail to explain what it is about the relevant considerations which enable them to influence intentionality attributions in an asymmetrical fashion. It is not enough to say that judgments of moral badness, for instance, influence intentionality attributions. One must also explain *how* such judgments have this effect. In other words, in order to achieve a conceptually sound understanding of the Knobe effect, one must first locate the source of the asymmetry of interest; then, one must explain how this source connects up with intentionality. It is not clear how one factor could play both of these roles simultaneously. In the absence of a clear explanation of how this is possible, a one-factor account remains conceptually inadequate.

The second reason to be skeptical of the badness account is that it is not obvious that it is supported by the results of DECR. In their study, Knobe and Mendlow did not actually ask participants whether or not decreasing sales in New Jersey was bad. Participants in Phelan and Sarkissian's study were asked whether this action was bad; they judged that it was *not*. Consequently, Phelan and Sarkissian conclude that (at least in these sorts of cases), *pace* the badness account, something other than the perceived badness (moral or otherwise) of the action is generating participants' intentionality attributions.

Third, and most importantly, the badness account is clearly disconfirmed by the findings reported above (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.[21]

Still, one might wonder whether the results of DECR challenge the two-factor account. There are good reasons to think that they do not. For one, the two-factor account does not state a

[21] Indeed, Knobe (forthcoming) now retracts the badness view in response to an earlier version of the present paper (as well as the results reported in, e.g., Phelan & Sarkissian (forthcoming)).

generally necessary condition for intentionality attributions (and thus it differs significantly from the view that Phelan and Sarkissian refer to as 'BLAM'). Surely other factors, such as intending to *A* and successfully *A*-ing, *modulo* deviant causal chains, may be sufficient for holding that *A* was done intentionally. So, the fact that participants attributed intentionality to Susan while failing to judge her blameworthy for her action is entirely consistent with our account.

Indeed, once we recognize that factors other than responsibility may be sufficient for intentionality attributions, it is not difficult to understand why DECR elicited such attributions. 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 addition, because DECR involves explicit deliberation, it is too dissimilar from HARM/HELP to function as a genuine comparison. In order to test the empirical adequacy of the two-factor (or, for that matter, any other) account of the Knobe effect, scenarios much more similar to HARM/HELP are needed. With this in mind, we gave 122 participants the following variants on DECR:

> DECR2: 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.
>
> INCR: 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 DECR, the main character (Susan) engaged in explicit deliberation. When explicit deliberation was removed, as in DECR2 and INCR, participants' responses became strikingly similar to those found in HARM/HELP.

First, most of the (many) intentionality attributions in DECR2 accompanied judgments of both a *bad* action and a *blameworthy* chairperson; likewise, most of the (few) intentionality attributions in INCR 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.[22] Second, participants were significantly more likely to judge that the chairperson acted intentionally when they stated that he/she was praiseworthy/blameworthy than when they did not.[23] 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.[24] 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.[25] This suggests 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 generates an intentionality attribution—and that, of the two, the responsibility of the actor plays the central role. Consequently, rather than challenging the two-factor account, DECR2/INCR 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 (forthcoming) has proposed that the Knobe effect is due to an affect-driven bias. Nadelhoffer concedes the central contention of the two-factor account—that assessments of responsibility (in particular, praise/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 (forthcoming, 22).

In explicating his view, 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]" (forthcoming, 16-17). As a result,

[22] DECR2: 68% vs. 41%; $\chi^2(120) = 7.4$, $p = .007$, $\varphi = .25$. INCR: 47% vs. 10%; $\chi^2(121) = 13.6$, $p < .001$, $\varphi = .34$.
[23] DECR2: 70% vs. 34%; $\chi^2(120) = 14.9$, $p < .001$, $\varphi = .35$. INCR: 41% vs. 11%; $\chi^2(121) = 10.8$, $p = .001$, $\varphi = .30$.
[24] DECR2: 60% vs. 43%; $\chi^2(120) = 3.5$, $p > .05$, $\varphi = .17$. INCR: 17% vs. 11%; $\chi^2(121) = .70$, $p > .05$, $\varphi = .08$.
[25] DECR2: 83% vs. 18%; $\chi^2(120) = 49.8$, $p < .001$, $\varphi = .64$. INCR: 18% vs. 5%; $\chi^2(121) = 3.3$, $p = .069$, $\varphi = .17$.

participants in HARM, for instance, are led to attribute intentionality as a result of their affect-driven attribution of blame.[26] Nadelhoffer concludes, "even though moral considerations surely do act expansively on folk ascriptions of intentional action…ideally they ought not have this effect" (forthcoming, 22).

While we are open to the possibility that affect may sometimes have a biasing effect on intentionality attributions, we believe that there are several reasons to reject Nadelhoffer's contention that an affect-driven bias explains 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 a judgment of a responsible actor and, derivatively, a good/bad action typically generates an intentionality attribution, 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. Third, the view is otherwise unmotivated, since the standard line of reasoning for the claim that participants' judgments are biased is problematic. Let us explain.

This reasoning is adopted by Nadelhoffer's 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

[26] Incidentally, because Alicke's model is restricted to blame, Nadelhoffer's account seems unable to explain the results of the DOCHARM/DOCHELP study, 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.

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, Nadelhoffer found that 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%). But since, according to Nadelhoffer (forthcoming, 14), "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. Nadelhoffer concludes that participants' judgments were biased, and that the source of this bias was affect.

However, reflection on these cases reveals that it is false that "the cases are identical in terms of the cognitive and conative considerations of the thief and the driver." The two vignettes are only superficially identical; in truth, they 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 a reason to not speed off (but instead to cooperate), the driver had no such reason—and, plausibly, had a positive 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 a reason not to speed off, and the driver *knew* he had no such reason.

This suggests that while THIEF and DRIVER are not cognitively and conatively identical, THIEF and HARM may be. For in both THIEF and HARM, the actors knowingly brought about a side-effect which they had a reason to not bring about. They were, consequently, blameworthy for having so acted. Presumably, this lead participants to make intentionality attributions in THIEF as in HARM; on the two-factor account, a judgment of a blameworthy actor and, derivatively, a bad action typically generates 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 negatively responsible (blameworthy) that they were disinclined to say that the actors acted intentionally in these cases.

As this makes clear, the standard line of reasoning for the claim that participants' judgments are biased is problematic. In fact, 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 (and perhaps philosophical) sense. If this is correct, then there is no reason to believe that the Knobe effect is due to an affect-driven bias.

3.3 Pragmatic Accounts

We now turn to pragmatic accounts, which appeal 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. According to Adams and Steadman (2004a, 2004b, 2006) and Wasserman (2006), the Knobe effect is the result of participants' desire to avoid an unwanted implicature associated with the denial of intentionality. Roughly, both accounts maintain 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 he/she 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 (2006) 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 his/her action, without feeling forced to state that he/she brought about those effects intentionally. Consequently, such a study would enable us to test the empirical adequacy of the pragmatic account.

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". So, even when participants were given the opportunity to assess the chairman as positively/negatively responsible for the effects of his action without stating that he brought about those effects intentionally, participants continued to respond as they did in the original studies—namely, with an intentionality attribution. These results disconfirm the predictions of proponents of the

pragmatic account. (See also the studies reported in Knobe (2004) and Nadelhoffer (2006) for further empirical evidence against pragmatic accounts.)

Adams and Steadman (forthcoming) have responded that these results do not undermine their account, but are simply evidence for the insidiousness of participants' "pragmatic programming".[27] We find this response unconvincing. To be sure, it does not follow from anything we have said above that participants' intentionality attributions in these cases are justified. As stated in §1, it is not our goal to defend this claim here. Nevertheless, the fact remains that pragmatic accounts, which entail that participants' intentionality attributions in these cases are unjustified, appear to be empirically inadequate.

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,[28] Turner (2004) has argued that the Knobe effect could be elicited by non-moral considerations as well. To illustrate, Turner invites us to consider the following putatively non-moral scenarios:

> CRITICISM: A director came to a Hollywood producer with a script and said "I want to make this movie. If you produce it, your studio will earn millions of dollars but you personally will be criticized by the media." The producer answered "I don't care at all about being criticized by the media. I just want to make as much profit as I can. Let's make the movie." They made the movie. Sure enough, the producer was criticized by the media.
>
> PRAISE: A director came to a Hollywood producer with a script and said "I want to make this movie. If you produce it, your studio will earn millions of dollars and you personally will be praised by the media." The producer answered "I don't care at all about being praised by the media. I just want to make as much profit as I can. Let's make the movie." They made the movie. Sure enough, the producer was praised by the media.

Turner predicts that the folk will judge that the producer incurred the criticism intentionally, but did not incur the praise intentionally (cf. Wasserman 2006). If this yet untested prediction is correct, then insofar as these scenarios do not involve moral considerations, the asymmetry which

[27] They defend this response by appealing to preliminary results which suggest that participants were more willing to judge that the chairman acted "knowingly, but not intentionally", rather than "knowingly and intentionally", in a putatively non-moral variant on HARM. In this variant, the side-effect of implementing the new program was tipping off a competitor, rather than harming the environment. Since participants were not asked whether the chairman was (negatively) responsible for his action, it is unclear exactly how to interpret these results. Thus, while we do not take this to be compelling evidence in favor of Adams and Steadman's view, we do consider it to be an invitation to future research exploring this and related issues.

[28] See, e.g., Adams & Steadman (forthcoming); Knobe (2003a, 2003b, 2004, 2005, forthcoming), Knobe & Mendlow (2004); and Nadelhoffer (2004a, 2004b, forthcoming).

they elicit in intentionality attributions cannot be explained by invoking a connection between assessments of the moral status of actions and/or actors and intentionality attributions.

Since the two-factor 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 responsibility, though evaluative through and through, can be both moral and *non-moral*, and that all assessments of positive/negative responsibility are subject to the asymmetry discussed in §1 (see also Harman, 1976). If this is correct, then an asymmetry in assessments of positive/negative responsibility, albeit *non-moral* positive/negative responsibility, could be responsible for the asymmetry between intentionality attributions in CRITICISM and PRAISE (assuming that there would be one). Our explanation proceeds as follows.

Recall DECR2/INCR, in which participants' assessments of responsibility influenced their intentionality attributions. Although DECR2/INCR elicited the Knobe effect, they are putatively non-moral scenarios. Insofar as these scenarios do not involve explicitly moral considerations, it is implausible to suppose that participants are attributing *moral* responsibility to the actor; presumably, they are attributing a sort of *non-moral* responsibility. Viewing the chairperson in DECR2 as negatively responsible (blameworthy) inclined participants to say that he/she acted intentionally; whereas viewing the chairperson in INCR as not positively responsible (not praiseworthy) disinclined them from saying that he/she acted intentionally. This remains so despite the fact that the relevant sort of responsibility was not moral.

An application of this point to CRITICISM/PRAISE is relatively straightforward. Plausibly, many would be inclined to say that the producer was to some extent criticizable for incurring the media's wrath since arguably he had at least some reason not to do so. Conversely, many would be reluctant to say that the producer was to any extent laudable for incurring the media's praise—after all, he made the movie simply to "make as much profit as [he could]." Regardless, we believe that to the extent that one would be willing to attribute responsibility to the producer for the media's reaction, one would be inclined to say that the producer incurred the media's reaction intentionally. That is, viewing the CRITICISM producer as negatively responsible would incline one to say that he/she incurred the media's criticism intentionally, whereas viewing the PRAISE producer as not positively responsible would disincline one to say that he/she incurred the praise intentionally. If people are more likely to say that the producer in CRITICIZE was negatively responsible for incurring the media's criticism than that the producer in PRAISE was positively responsible for incurring the media's praise (since he/she did not incur that praise *for the right reasons*), it would be unsurprising if the prediction of Turner was confirmed.

This particular prediction aside, the results of DECR2/INCR appear to support the hypothesis that the Knobe effect arises in at least some non-moral cases. Accounts of the Knobe effect which appeal to the 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. While these results pose a serious challenge to, e.g., the badness account, they offer strong support in favor of the two-factor account (as shown in §3.1).

5. Conclusion

We have argued for a particular account of the asymmetry in folk judgments of intentional action (the Knobe effect). On this two-factor 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 sufficient for negative responsibility (e.g., blameworthiness, criticizability), regardless of one's reasons. On the other hand, positive responsibility (e.g., praiseworthiness, laudability) requires more: it requires bringing about a foreseen good outcome *for the right reasons*. This asymmetry, coupled with the fact that intentionality commonly connects the goodness/badness of actions to the responsibility of actors, which alone typically implies intentionality, accounts for the influence of evaluative considerations on intentionality attributions.

While we have suggested that this account may render the asymmetry in folk judgments of intentional action justified, our primary goal has been to show that the two-factor account provides an empirically and conceptually sound understanding of the Knobe effect. In particular, we have argued that it is entirely consonant with the results of a number of empirical studies that probe folk judgments concerning the relation between evaluative considerations and intentionality.

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 instance, the badness account rightly observes that the Knobe effect is somehow related to judgments of badness, which figure into factor (ii) of the two-factor account. 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.

The same is true of various alternative accounts which were not discussed in §3, such as Nichols and Ulatowski's (2006) *interpretative diversity hypothesis*. According to these authors, the best explanation of the asymmetry found in HARM/HELP scenarios is an "interpretative

diversity" (e.g., ambiguity, polysemy, or semantic underspecification) in the term 'intentional'. They argue that some participants use the word 'intentionally' to roughly mean *with foreknowledge*, whereas others impose stricter requirements: for them, 'intentionally' roughly means *with motive*. Note that this asymmetry corresponds to the asymmetry between assessments of positive and negative responsibility invoked in factor (i). While positive responsibility is commonly taken to require a *motive* (i.e., that the action be performed for the right reasons), negative responsibility is commonly taken to require only *foreknowledge* (i.e., that the action be performed with knowledge of the outcome). It seems, then, that we may explain what is right about Nichols and Ulatowski's hypothesis without invoking an "interpretative diversity", since the alleged diversity appears to reduce to a principled difference between the conditions under which assessments of positive and negative responsibility are intuitively appropriate.

As this illustrates, the two-factor account appears to have the resources to explain in a systematic way the appeal of alternative accounts while avoiding their pitfalls. For reasons that should be obvious, we take this to be an additional consideration in its favor.

References

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

Adams, F. & Steadman, A. Forthcoming. Folk concepts, surveys, and intentional action. In C. Lumer (ed.). *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy*. Aldershot: Ashgate Publishers.

--------. 2004a. Intentional action and moral considerations: Core concept or pragmatic understanding? *Analysis*, 64: 173-181.

--------. 2004b. Intentional action and moral considerations: Still pragmatic. *Analysis*, 64: 264-267.

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.

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

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

Duff, R.A. 1982. Intention, responsibility, and double effect. *The Philosophical Quarterly*, 32(126): 1-16.

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

--------. 1973. *Thought*. Princeton, NJ: Princeton University Press.

Hauser, M. Forthcoming. *Moral minds: the unconscious voice of right and wrong*. Harper Collins.

Knobe, J. Forthcoming. Reason explanation in folk psychology. *Midwest Studies in Philosophy*.

--------. 2005. Theory of mind and moral cognition: Exploring the connections. *TRENDS in Cognitive Science*, 9(8): 357-359.

--------. 2004. Intention, intentional action, and moral considerations. *Analysis*, 64: 81-187.

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

--------. 2003b. Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16(2): 309-324.

Knobe, J. & 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. 2006. Understanding the folk concept of intentional action: Philosophical and experimental issues. Unpublished manuscript.

Malle, B. 2006. Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, 6(1-2): 87-112.

McCann, H. 2005. Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, 18(6): 737-748.

Mele, A. & Sverdlik, S. 1996. Intention, intentional action, and moral responsibility. *Philosophical Studies*, 82: 265-87.

Nichols, S. & Knobe, J. Forthcoming. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*.

Nichols, S. & Ulatowski, J. 2006. Intuitions and individual differences: The Knobe effect revisited. Unpublished Manuscript. Available at: http://www.unc.edu/~knobe/ Nichols-Ulatowski.pdf.

Nadelhoffer, T. Forthcoming. Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*.

--------. 2006. Saving the simple view. Unpublished Manuscript. Available at: http://garnet.acns.fsu.edu/%7Etan02/Papers/Simple%20View%20Revise%20and%20Resubmit.pdf.

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

--------. 2004b. On praise, side effects, and folk ascriptions of intentional action. *The Journal of Theoretical and Philosophical Psychology*, 24: 196-213.

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

Pizarro, D., Knobe, J. & Bloom, P. 2006. 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.

Scanlon, T. 1998. *What we owe to each other*. Cambridge, MA: Harvard University Press.

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.

Wasserman, R. 2006. Intentional action and the unintentional fallacy. Unpublished Manuscript. Available at: http://myweb.facstaff.wwu.edu/wasserr/research.html.

Young, L., Cushman, F., Adolphs, R., Tranel, D. & 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(1-2): 265-278.