A New Angle on the Knobe Effect: Intentionality Correlates with Blame, not with Praise

Frank Hindriks
Faculty of Philosophy, University of Groningen
f.a.hindriks@rug.nl

Igor Douven
Sciences, Normes, Décision (CNRS)
Paris-Sorbonne University
igor.douven@paris-sorbonne.fr

Henrik Singmann
Department of Psychology, University of Zurich
singmann@psychologie.uzh.ch

Abstract

In a celebrated experiment, Joshua Knobe showed that people are much more prone to attribute intentionality to an agent for a side effect of a given act when that side effect is harmful than when it is beneficial. This asymmetry has become known as "the Knobe Effect." According to Knobe's Moral Valence Explanation (as we call it), bad effects trigger the attributions of intentionality, whereas good effects do not. Many others believe that the Knobe Effect is best explained in terms of the high amount of blame attributed in the harm condition, and the low amount of praise attributed in the help condition. This Blame Hypothesis (as we call it) explains the high number of intentionality attributions in the harm condition in terms of the high degree of blame people ascribe, and the low number of intentionality attributions in the help conditions in terms of the low degree of praise people attribute. We replicated Knobe's original experiment and conducted a logistic regression on the results to probe more deeply into the relationship between attributions of intentionality and responsibility. The statistical analysis revealed a hitherto unknown interaction effect: intentionality correlates with blame, but not with praise. This effect is consistent with the Moral Valence Hypothesis, but inconsistent with the Blame Hypothesis, as well as with two of the three other hypotheses discussed here.

1. Introduction. According to the Simple View of intentional action, an effect of an action is brought about intentionally only if the agent intended to bring it about (Adams 1986). Those who reject the Simple View accept that someone can also intentionally bring about an effect that she did not want to bring about. Gilbert Harman (1976)

provides a famous example of a sniper who alerts the enemy by shooting a soldier. He maintains that the sniper alerts the enemy intentionally, because he takes the benefit of his intended action to outweigh the cost of the side effect (see also Bratman 1987). This entails that an intentional action can concern something the agent does not favor. Joshua Knobe (2003) has made the striking discovery that people sometimes attribute intentionality to an agent who expresses indifference about a side effect of her action. Furthermore, they do so only when the side effect is harmful and not when it is beneficial. This asymmetry in intentionality attributions has become known as "the Knobe Effect." As Knobe (2006) explains it in terms of the moral valence of the side effect, we refer to his account as "the Moral Valence Hypothesis" (MVH). According to MVH, people tend to attribute intentionality when a side effect is bad, but not when it is good.

Knobe also measured attributions of moral responsibility. He discovered what one of us has called "the Praise–Blame Asymmetry" (Hindriks 2008:630): people attribute a lot of blame to an agent who is indifferent about a harmful side effect, but they hardly attribute any praise when the agent is indifferent about a beneficial side effect.[1] Knobe made a further observation that came to play a central role in the explanation of the Knobe Effect, to wit, that the moral responsibility attributions people make correlate with their intentionality ascriptions.[2] This observation has inspired many to defend what we call "the Blame Hypothesis" (BH), the claim that the responsibility attributions explain the intentionality ascriptions people make in scenarios in which the Knobe Effect is observed (Mele 2001, Malle and Nelson 2003, Nadelhoffer 2004, 2006, Nado 2008).[3] BH boils down to the claim that the Praise–Blame Asymmetry explains the Knobe Effect.

As responsibility is a gradable property, BH can easily be reformulated in terms of degrees. According to the most straightforward graded formulation (call it "BH*"), the higher the level of responsibility attributed, the higher the chance that people ascribe intentionality. When tested in terms of average values of praise or blame, most experiments fit the hypothesis reasonably well. Such tests, however, all but ignore people who for some reason attribute either little blame or a lot of praise. Still, BH* does make a prediction about them: the chance with which they ascribe intentionality will be low and high respectively. This prediction has thus far not been tested directly. Chandra Sripada and Sara Konrath (2011) fit a structural path model that includes the agent's moral status, which they regard as an indicator of moral responsibility. The question they ask is whether the agent's status is moral or immoral (on a 7-point Likert scale with "Very moral" and "Very immoral" as anchors). They find that the agent's moral status has no significant effect on intentionality. To the extent that status does indeed track responsibility, this does not bode well for the aforementioned prediction.

[1] Specifically, $M = 4.8$ versus $M = 1.4$ on a scale from 0 to 6, where 0 stands for no blame/praise and 6 for a lot of blame/praise (Knobe 2003).

[2] Knobe observed that "the total amount of praise or blame that subjects offered was correlated with their judgments about whether or not the side effect was brought about intentionally, $r(120) = .53$, $p < .001$" (2003:193).

[3] Phelan and Sarkissian (2008) refer to this as the blameworthiness model.

After all, BH takes intentionality and responsibility attributions to be correlated. For all we know, however, people have something other than praise- or blameworthiness in mind when they rate the agent's moral status.[4] The upshot is that BH can be put to a more severe test than it has faced thus far.

BH is symmetrical in the sense that it regards both praise and blame as factors due to which people are prone to attribute intentionality. The asymmetry in intentionality ascriptions is due not to the fact that praise does not trigger the attribution of intentionality, but to the fact that people hardly attribute any praise (see Section 4.2 for more on this). MVH, in contrast, is asymmetric, as it regards only bad effects as triggers and not good effects. As moral valence is a gradable property, just like responsibility, MVH can also be reformulated in terms of degrees. This version (call it "MVH*") has it that the worse someone takes a side effect to be, the more likely it is that she attributes intentionality. The part of MVH* that applies to the help condition remains the same: irrespective of the degree to which it is good, the goodness of a side effect is not a triggering factor. In contrast to BH*, MVH* entails an interaction effect. Assuming that the degree of blame people attribute correlates with how bad they deem a side effect to be, attributions of blame and intentionality should correlate, whereas attributions of praise and intentionality should not.

In Section 3, we investigate experimentally the relation between intentionality and responsibility attributions in order to establish which prediction is corroborated, that of MVH* or that of BH*. In Section 4, we also discuss three more recent hypotheses about the Knobe Effect in the light of the empirical results presented below (Hindriks 2008, 2011, Holton 2010, Sripada 2010, Sripada and Konrath 2011). We start by taking a more fine-grained look at how people's intentionality judgments relate to their attributions of moral responsibility, which allows for a more detailed assessment of the various explanations of the Knobe Effect that have been proffered in the literature.

2. The Moral Valence Hypothesis and the Blame Hypothesis. The Knobe Effect is usually accounted for at the level of percentages and averages: the asymmetry that is explained consists of the fact that a large majority of the participants attribute intentionality in the harm condition, whereas a large majority does not do so in the help condition (82 % versus 23 % in Knobe 2003). The Praise–Blame Asymmetry consists of the fact that the average amount of blame attributed in the harm condition is high, whereas the average amount of praise attributed in the help condition is low (see note 1). In spite of the efforts of Sripada and Konrath (2011), an in-depth investigation into the correlations between intentionality and responsibility attributions remains to be conducted. In the present paper, we set out to fill this lacuna. The idea that underlies

[4] Sripada (2011) finds that people blame the agent not only in the harm condition but also in the help condition, albeit to a smaller degree. He concludes that it is unlikely that the blameworthiness of the agent explains the intentionality attributions (ibid., 236). In spite of the fact that he conducts a number of mediation analyses in this paper, he has not, however, conducted any tests that bear directly on the relation between intentionality and responsibility.

our plea for degrees is that gradual differences are more informative than averages. They promise to substantially advance the debate about the explanation of the Knobe Effect by narrowing the range of explanations that still deserve to be taken seriously. In this section, we make precise the predictions that degree versions of MVH and BH support.

MVH purports to explain the Knobe Effect in terms of moral valence attributions, that is, in terms of whether the side effect that the agent brings about is good or bad. The graded version MVH* has the following two sub-theses:

MVH*–Harm: The higher the degree of badness people assign to the side effect in the harm condition, the higher the chance that they attribute intentionality to the agent.

MVH*–Help: The chance with which people attribute intentionality to the agent is independent of the degree of goodness they assign to the side effect in the help condition.

MVH* predicts that intentionality and moral valence correlate when the side effect is bad, but not when it is good. To the extent that degree of badness correlates with extent of blame, MVH* predicts an interaction effect between intentionality and moral responsibility.

BH purports to explain the Knobe Effect in terms of moral responsibility attributions, specifically in terms of the Praise–Blame Asymmetry. According to BH*, the degree of responsibility attributed correlates with the chance with which intentionality is attributed.[5] BH* has the following two sub theses:

BH*–Harm: The more blame people assign to the agent in the harm condition, the higher the chance that they attribute intentionality to him.

BH*–Help: The more praise people assign to the agent in the help condition, the higher the chance that they attribute intentionality to him.

In contrast to MVH*, BH* implies that intentionality and responsibility correlate both in the harm and the help conditions. These two hypotheses can be tested using the very same experimental setup as Knobe (2003) used originally, although doing so requires a statistical analysis different from those Knobe and others have carried out so far.

3. Experiment. We reran what is basically Knobe's original experiment concerning the chairman scenario. The purpose of this was not to verify the reproducibility of Knobe's result. As discussed in Knobe (2010), sufficiently many other studies have

[5] Mele's (2001) version of BH cannot be reinterpreted in terms of degrees. He suggested that people attribute intentionality in the harm condition because they assign blame and mistakenly assume that blame presupposes intentionality. However, Mele (2003) abandoned his hypothesis when it became clear that people do not in fact make this assumption.

done that, with consistently positive outcomes. Instead we were particularly interested in conducting an analysis of the data that might shed light on the question of whether the relationship between intentionality and blame/praise ascriptions interacts with the harm/help variable, as is suggested by some explanations of the Knobe Effect (e.g., MVH and also NRH to be discussed below).

3.1. Method

Participants

There were 347 participants in our study. The participants were recruited via CrowdFlower (http://www.crowdflower.com), which directed them to the Qualtrics platform (http://www.qualtrics.com) via which the survey was administered. Repeat participation was blocked via the CrowdFlower interface. All participants were from Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. Nonnative speakers (N = 34) were excluded from the analysis. From the remaining 313 participants, we removed the 5 % fastest and 5 % slowest responders, which left us with 282 participants on whose responses the analysis was based. These participants spent on average 182 seconds on the survey (SD = 152 s). Of them, 154 were female (55 %). The mean age of the participants was 34 (SD = 13).

Materials and Procedure

The materials were all in English, the participants' native language, and shown on screen.

At the beginning of the survey, participants were told that they would be presented a fictional scenario involving a decision that they would be asked to assess. Participants were then randomly assigned to one of two conditions, to which we refer as "the harm condition" and "the help condition." Participants in the harm condition were first presented the following scenario, which is taken verbatim from Knobe (2003):

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

Participants in the help condition were shown the same scenario except that in that scenario the environment was said to be *helped* as a side effect of starting the new program and the chairman expressed his indifference towards *helping* the environment. On a second screen, all participants were then asked whether they thought that the chairman brought about the effect on the environment intentionally, the answer options being "Yes" and "No." And on the third and last screen, they were asked how blameworthy or praiseworthy they thought the chairman was, given that his decision affected the environment, where the answer had to be given on a 7-point Likert scale with "Very blameworthy" and "Very praiseworthy" as anchors and "Neither blameworthy nor praiseworthy" as the midpoint.

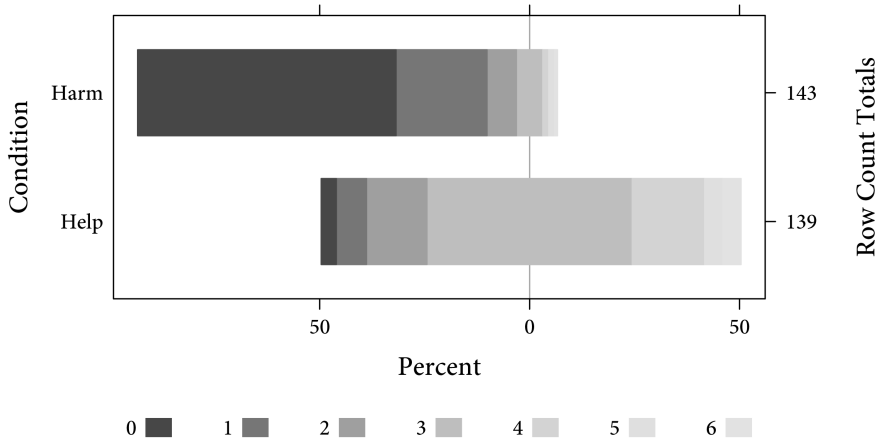

Figure 3.1: The attribution of blame/praise (from 0 = "Very blameworthy" to 6 = "Very praiseworthy") considered separately for the harm and help conditions in a diverging stacked bar chart (Heiberger and Robbins 2014).

3.2. Results

In conformity with Knobe's and others' findings, a vast majority (84 %) of the participants in the harm condition attributed intentionality to the chairman, whereas only a relatively small minority (23 %) of the participants in the help condition attributed intentionality: a clearly significant difference using Fisher's exact test, $p < .0001$.

We also found further confirmation of the Praise–Blame Asymmetry. We coded the "Very blameworthy" endpoint of our 7-point scale as 0, the "Very praiseworthy" endpoint as 6, and the intermediate points in the obvious way. As Figure 3.1 shows, participants in the harm condition on average attributed a lot of blame but participants in the help condition on average did not attribute any praise (nor blame). Specifically, in the harm condition $M = 0.71$ ($SD = 1.18$), which a one-sample t-test showed to be significantly different from the neutral "Neither blameworthy nor praiseworthy" midpoint coded as 3 ($t(142) = -23.20$, $p < .0001$), while in the help condition $M = 2.99$ ($SD = 1.22$), which a one-sample t-test showed to be not statistically significant from the midpoint ($t(138) = -0.07$, $p = .94$). The two conditions also significantly differed from each other ($t(287.76) = -15.92$, $p < .0001$).

Also like Knobe (2003), we found a moderately strong to strong correlation between the amount of praise or blame that subjects offered and their judgments concerning whether the side effect was brought about intentionally, $r(280) = -.52$, $p < .0001$. This correlation is a direct consequence of the differences in both attributed intentionality and attribution of praise or blame. What has not been done so far, to our knowledge, is considering the relationship for the harm and help conditions separately. For this,

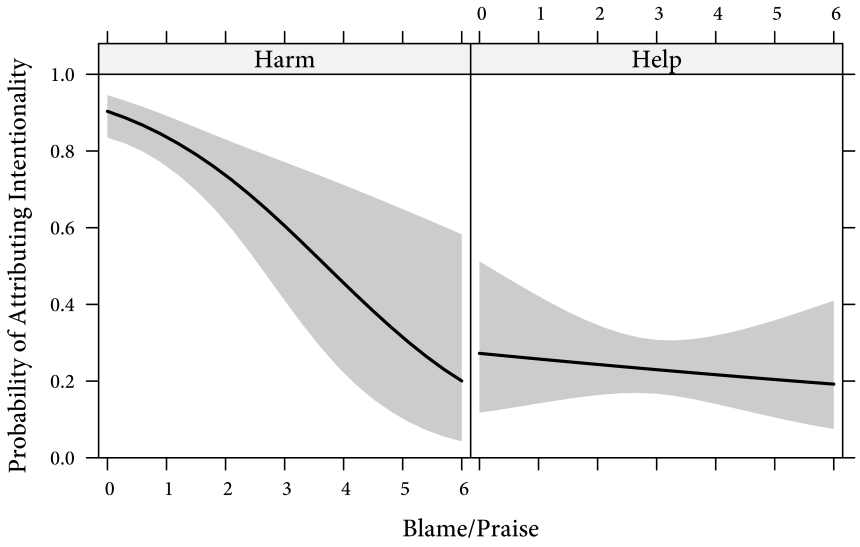

Figure 3.2: The probability of attributing intentionality as a function of the attribution of blame/praise considered separately for the harm and help conditions. The grey areas depict 95% confidence bands.

it is important to note that this entails comparing the relationship of attributions of praise or blame with intentionality across two conditions, where intentionality is the dependent variable. As intentionality is binary ("yes" versus "no"), we employed a logistic regression for this analysis, a statistical procedure appropriate for binary dependent variables (Agresti 2002).

We estimated a logistic regression with intentionality as dependent variable and condition, the amount of praise or blame, and their interaction as independent variables. To facilitate the interpretation of the main effects, we centered the amount of praise and blame at the midpoint of the scale (Cohen, Cohen, Aiken, and West 2002). Results revealed a significant interaction, $\chi^2(1) = 4.93$, $p = .03$, indicating that the relationship of amount of praise and blame with intentionality differs between conditions. Furthermore, we again found the main effects of both condition, $\chi^2(1) = 12.17$, $p = .0005$, and the amount of praise and blame, $\chi^2(1) = 8.29$, $p = .004$. The results of the logistic regression are displayed in Figure 3.2. As can be seen, the relationship between the amount of blame and praise and intentionality was only present in the harm condition but absent in the help condition. This was further confirmed by a post-hoc trend test: only in the harm condition was the effect of amount of praise and blame significant, $\beta = -0.60$, $p = .0005$, but not in the help condition, $\beta = -0.08$, $p = .65$.[6]

[6]When running separate correlations of attributions of praise or blame with attributions of intentionality per condition the same pattern emerges, $r(141) = -.33$, $p < .0001$, in the harm condition and

3.3. Discussion

Following Knobe (2003), we have investigated intentional action and moral responsibility in conjunction. However, instead of resting content with reporting the overall correlation between intentionality and responsibility attributions, we have distinguished between the correlation in the help condition and the correlation in the harm condition. The main findings of our study are HARM and HELP:

HARM: The more blame that participants attribute in the harm condition, the higher the chance that they take the indifferent agent to have acted intentionally.

HELP: The chance that participants attribute intentionality to the indifferent agent in the help condition is low, irrespective of the praise or blame they assign.

Jointly, HARM and HELP reveal a new asymmetry, one that pertains to the correlations between intentionality and responsibility attributions: the degree of responsibility ascribed correlates with the chance with which intentionality is attributed in the harm condition, but not in the help condition.[7]

4. General discussion. How can it be that intentional action (IA) and moral responsibility (MR) are independent in the help condition, but correlated in the harm condition? This finding reveals that, in order to be remotely plausible, the hypothesis that explains the Knobe Effect (1) involves a gradable property that (2) is treated asymmetrically in the help and harm conditions. HARM can, after all, only be explained in terms of a property that admits of degrees. And a symmetric hypothesis can explain either HARM or HELP, but not both. As is discussed below in some detail, MVH* has this structure, whereas BH* does not. MVH* will turn out to be problematic on independent grounds, however. We further discuss three other hypotheses that have been put forward in recent years, finding one of these to be better poised to explain our experimental findings than rival hypotheses.

4.1. The Moral Valence Hypothesis

According to Knobe's MVH, the badness of a side effect explains the attribution of intentionality, whereas the goodness of a side effect does not. When reformulated in terms of degrees, this hypothesis predicts a correlation between the degree to which a side effect is bad and the chance with which people attribute intentionality. MVH* does not predict such a correlation in the help condition. Due to the fact that (1)

$r(137) = -.04$, $p = .65$, in the help condition.

[7] The chance that participants attribute intentionality to the agent in the help condition is far from zero. The data on this might be nothing more than noise. The salient alternative explanation is that there are multiple concepts of intentionality only one of which harbors an asymmetry. On such a pluralist interpretation, some people treat foresight as a sufficient condition for intentionality, and others treat desire as a necessary condition (Nichols and Ulatowski 2007, Cushman and Mele 2008, Cokely and Feltz 2009, and Pinillos et al. 2011).

the explanatory factor in the harm condition is a gradual property and (2) plays an asymmetric role in the hypothesis, MVH* fits HARM as well as HELP.

Both MVH and MVH* explain the Knobe Effect in terms of the moral judgments that participants make. According to these hypotheses, what matters is whether attributors regard an effect as good or bad. The evidence suggests, however, that the moral perspective that explains the intentionality attributions is that of the agent instead. Knobe (2007) tests a scenario involving a CEO of a company in Nazi Germany. In this scenario, a racial identification law is in force that requires all companies to identify the race of their employees. Just as in the chairman vignettes, the CEO of this company expresses his indifference with respect to this law, claiming that he only cares about profit. Knobe used this scenario because he expected that the moral evaluations of the CEO will be the reverse of those of (US American) participants. The CEO thinks that it is in principle a good thing to conform to the racial identification law. He just cannot be bothered. In contrast, the participants will regard conforming to this law a bad thing. As it turns out, the majority attributes intentionality in the condition that the agent regards as bad—the one in which he violates the norm—and not in the condition that the participant regards as bad. Thus, the intentionality attributions turned out to be sensitive to the agent's hypothesized valuations and not to those of the attributors. As they invoke moral judgments of the attributors, this finding is bad news for both MVH and MVH*. This finding fits seamlessly with the idea that intentional action is a folk psychological notion and is as such concerned with the perspective of the agent. Intentional action is a notion that qualifies the frame of mind with which the agent acts (Bratman 1987, Hindriks 2014).

Knobe (2010; see also Pettit and Knobe 2009) has recently proposed a version of MVH that is sensitive to the fact that intentional action is a frame of mind notion. What we refer to as "MVH†" features not only the perceived goodness or badness as an explanatory factor, but also the pro-attitudes and thereby the frame of mind of the agent. According to MVH†, an effect requires a pro-attitude for it to be brought about intentionally. The intensity of the pro-attitude—that is, the degree to which it favors the effect—has to surpass a certain threshold value in order for it to warrant the attribution of intentionality. The threshold value depends on the perceived moral valence of the effect. Morally neutral effects provide for the default. When the effect is good, the threshold for intentionality is higher than the default value. When the effect is bad, the threshold is lower. This is because we expect people to have a favorable attitude toward a good effect and an unfavorable attitude toward a bad effect. Indifference exceeds the threshold when the effect is bad, but not when it is good. This last feature of MVH† accounts for the Knobe Effect.[8] In contrast to MVH*, MVH† cannot account for both HARM and HELP. Even though it involves a gradable property, MVH† is a threshold

[8] MVH† served to account for asymmetries concerning notions such as desiring, favoring, and being happy. People tend to apply these notions in the harm condition. In the help condition, however, they are more or less neutral about their applicability. Knobe (2010) argues that this can be explained by using a more flexible framework that involves the pro-attitudes of the agent as well as threshold values concerning those attitudes (see Hindriks 2014 for a critical discussion).

account, which means that it cannot explain HARM. Furthermore, the hypothesis treats the help and harm conditions symmetrically. This means that a version that does not rely on thresholds can explain only HARM and not HELP. Note that, just as MVH and MVH*, MVH† cannot account for valence reversals such as the one involved in the Nazi Germany case. This can be explained only by accounts that give priority to the agent's perspective over that of the attributor (Uttich and Lombrozo 2010).

4.2. The Blame Hypothesis

Knobe believes that the asymmetry in the attributions of intentionality is due to the conceptual competences people have regarding the concept of intentional action. Judgments in which this concept is applied will be correct if they are due only to those competences. As it explains the asymmetry only in terms of people's competences, Knobe offers a competence account. Many others hold, in contrast, that the Knobe Effect is due to a factor that interferes with those competences. That factor distorts or biases the judgments to which it contributes. As they invoke a biasing factor in their explanation, we refer to such accounts as "bias accounts."[9] The vast majority of those who do so hold that responsibility attributions explain the Knobe Effect. As blame is a gradable property, BH* can explain HARM. The problem is that praise is a gradable property as much as blame is. Hence, BH* predicts that praise also correlates with intentionality. HELP reveals that it does not. The upshot is that BH* does not have the requisite structure for explaining both HARM and HELP. It only explains HARM, and not HELP.

At this point, one might wonder whether BH* should perhaps have been formulated in asymmetric terms. Perhaps supporters of BH have postulated a mechanism that treats praise and blame differently. There is, however, no reason to interpret existing proposals in this way. A proposal that was put forward early on is that emotions connect responsibility and intentionality (Malle and Nelson 2003). Both positive and negative affect can be more or less intense. Hence, there seems to be no reason why an affect-based mechanism would treat praise and blame asymmetrically.[10]

Jennifer Nado (2008) argues that responsibility attributions *tacitly* interfere with the ascription of intentionality. Because people are not aware of their influence, the bias arises even when it conflicts with consciously held beliefs.[11] Nado (ibid., 721) suggests that this mechanism is at work both in the harm condition and in the help condition. This means that this second version of BH* is symmetric as well.[12]

[9] Nado contrasts competence accounts to performance accounts, where performance accounts "recast the effects as performance errors" (2008:713). As "performance" does not wear the inadequacy of the attributions on its sleeve, we prefer the term "bias," which is more commonly used in order to mark distorted responses.

[10] Nadelhoffer (2006) regards both moral valence and emotions related to responsibility as factors that explain the Knobe Effect. This makes it difficult to determine how exactly his account would generalize.

[11] This is meant to defuse the problem from which Mele's (2001) proposal suffered (see note 5).

[12] Just as Nado does, Alicke (2008:185) also allows for tacit influences. In addition to this, he appeals to the outcome bias in support of the claim that blame distorts intentionality attributions more than praise

At this point, the prospects for BH and BH* are pretty bleak.[13] To be sure, it cannot be ruled out that a proper motivation can be given in support of an asymmetrical formulation of BH. We cannot think of one, however, and turn to three hypotheses that appear more promising.

4.3. Deep selves, norms, and reasons

The remaining hypotheses to be discussed are, like MVH, competence accounts. In contrast to MVH, however, they explain the Knobe Effect in terms of the agent's perspective, and not in terms of the attributor's perspective. Specifically, we discuss Sripada's Deep-Self Hypothesis (DSH), Holton's Norm-Violation Hypothesis (NVH), and Hindriks' Normative Reason Hypothesis (NRH). We argue that DSH and NVH cannot explain both HARM and HELP, whereas NRH can.

Sripada (2010; see also Sripada and Konrath 2011) zooms in on the goodness or badness of the side effect as perceived by the agent, and he relates it to the agent's evaluative attitudes. He takes the relevant values and attitudes to constitute the agent's "deep self." Sripada finds that, when asked, people attribute anti-environment values and attitudes to the chairman (e.g., profit is more important than environment). According to DSH, people attribute intentionality only to agents whose attributed deep self is concordant with the moral valence of the effect she brings about. In the scenarios in which the Knobe Effect is observed, the agent's deep self is concordant with a harmful effect, and not with a beneficial effect.

It is in fact quite natural to interpret DSH in graded terms. Presumably an effect can be more or less concordant or discordant with the agent's deep self. However, the problem is that DSH is symmetric. This means that when interpreted in categorical terms, DSH can explain only HELP and not HARM. In contrast, when it is interpreted in graded terms, the hypothesis can explain only HARM and not HELP. The upshot is that a graded version of DSH cannot account for the interaction effect we have found.

According to Holton's (2010) NVH, violating a norm is something people do intentionally, whereas conforming to a norm need not be done intentionally.[14] In the harm condition, the agent violates a norm. In the help condition, the agent conforms to it. Not only does the agent conform to a norm in the help condition, it is clear that

does. According to the outcome bias, negative outcomes arbitrarily trigger stronger moral evaluations than positive outcomes. However, this line of reasoning works only if there is reason to believe that the responsibility attributions at issue are biased. This is not the case. The indifference of the agent supports blame, but not praise. In this connection it is interesting that, just like Sripada (2011), we find that on average people do not praise the agent in the help condition at all. Without an appeal to the outcome bias, Alicke's account collapses into Nado's.

[13] Independent evidence against BH is provided by Pellizzoni, Girotto and Surian (2010). They present the participants in their experiment with a scenario in which two agents attribute responsibility to the chairman, but disagree about whether he affected the environment intentionally. The underlying idea was that this would do away with any bias people might have to attribute intentionality in order to support their responsibility attributions. However, they still find the Knobe Effect: 15 % attributes intentionality in the help condition, 84 % in the harm condition.

[14] See Uttich and Lombrozo (2010) for another explanation of the Knobe Effect that invokes norms.

his conforming to the help norm is incidental to his action. Hence, NVH explains the intentionality attributions in both conditions. As it turns out, however, NVH faces problems insofar as the correlations between responsibility and intentionality are concerned. Violating a norm—the explanatory factor that Holton invokes—is not a gradable property. Hence, NVH can only account for HELP and not for HARM. Perhaps it makes sense to talk of more or less conforming to a norm—one could help more, for instance. However, it is difficult to see what more or less violating a norm would mean in this context—any harm constitutes a violation of the norm. Hence, we conclude that NVH cannot explain the correlational asymmetry we found.

According to Hindriks' (2008, 2011, 2014) NRH, the indifference of the agent plays a central role in the explanation of the Knobe Effect. Due to his indifference, the agent fails to be motivated by an effect that he should care about. In other words, he ignores a normative reason. An asymmetry surfaces, however, once it is recognized that the side effect counts against performing the intended action of maximizing profit only in the harm condition. Now why would this matter to the ascription of intentionality? Because intentionality attributions serve as input for responsibility attributions.[15] In order for it to be useful for this purpose, the notion of intentional action has to mark the motivation of the agent in an insightful manner. The question that arises is how the agent's motivation should be marked when an agent is indifferent about the outcome. No special indication is needed when the outcome is helpful, as the agent's indifference blocks praise. When, by contrast, the outcome is harmful, the agent's indifference betrays the fact that he was not motivated to avoid it, as he should. A purpose is served by marking this and attributing intentionality to the agent, as it indicates that, insofar as his motivation is concerned, there is reason to blame him.[16]

How can NRH be reformulated in terms of degrees? The thing to observe is that an agent can be more or less indifferent with respect to something. In common parlance, indifference is a propositional attitude we sometimes think of in a categorical way—as when we say categorically that someone is indifferent (or not indifferent) with respect to a given issue—and sometimes think of in a graded way—as when we say that someone is indifferent to some issue to some degree. This is nothing out of the ordinary: the same is true for belief (Foley 1992) and various other propositional attitudes (Over, Douven, and Verbrugge 2013). When conceived of in gradual terms, complete indifference is a matter of neutrality. The extent to which an agent cares can be plotted along a continuum with neutrality as one extreme and some maximum degree of caring as the other extreme. When the agent's indifference is treated as a gradual matter, NRH is the hypothesis that the more indifferent an agent is with respect to a normative reason

[15] Guglielmo and Malle (2010:1635) observe that the factors that typically feed into intentionality attributions also bear on the attribution of moral responsibility.

[16] The idea that lies at the basis of this line of reasoning is that praise requires an intention to bring about a good effect, whereas blame does not require an intention to bring about a bad effect (Stocker 1973:60; Scanlon 1998:271; Wolf 1990:80). This implies that, insofar as bad effects are concerned, there are cases in which the agent's overall motivation is bad, even though this is not apparent from his intention. In order to mark this, people say that he brings about the bad side effect intentionally.

that counts against her intended action, the higher the chance that intentionality is attributed to her.

Even though it plays no role in MVH, Knobe recognized the relevance of indifference early on when he observed: "Even when the effect itself was clearly bad, people only regarded it as intentional when the agent was indifferent, not when the agent was reluctant or trying to prevent it" (Knobe 2004:277; see also Nadelhoffer 2004). More recent evidence suggests that the degree to which someone cares matters. Steve Guglielmo and Bertrand Malle (2010) contrast the indifferent chairman to a CEO who says that it would be unfortunate if the environment got harmed, but his primary concern is to increase profits. The intentionality ratings go down from 87 % to 40 % of the participants. Al Mele and Fiery Cushman (2007) as well as Mark Phelan and Hagop Sarkassian (2008) find even more extreme results concerning agents who regret bringing about the side effect. As it turns out, hardly anyone attributes intentionality to a caring agent. This provides independent evidence in support of the claim that the agent's attitude toward the side effect is what matters.[17]

Apart from caring, NRH harbors a second property that can be conceived of in terms of degrees: normative reasons. The weight of a normative reason indicates the degree to which an agent should care. Mele and Cushman (2007) are on to the relevance of this factor when they suggest that participants might take the agent to be justified in accepting the cost of bringing about the harmful side effect. In other words, a regretful agent might be right to perform the intended action, as the value of the intended effect could indeed outweigh the side effect. When this is the case, guilt will not be in order, because the action is permissible. Regret is in order, because it reflects the appropriate concern. Even if what you do is acceptable, you may lament part of it. In light of this, Hindriks (2011:799) proposes that it matters whether the harmful effect provides sufficient reason for not performing the intended action. This fits well with how NRH conceptualizes the relation between intentionality and responsibility. When the agent is justified in bringing about an effect and recognizes this, his motivation is not such as to warrant blame. Hence, there is no reason for qualifying his behavior as intentional (it would in fact only be confusing to say that it is intentional but not blameworthy). Thus, there is also independent support for the relevance of the weight of the normative reason. Reformulating NRH explicitly in terms of degrees with respect to the two gradable properties it features results in the following hypothesis in the harm condition:

NRH*–Harm: The larger the discrepancy between how much the agent should care about the harmful side effect and how much she actually cares about it, the higher the chance that people attribute intentionality.

[17] Whereas regret decreases the tendency to attribute intentionality, there is reason to believe that disfavoring the side effect increases it. In non-moral scenarios most people say that an agent who disfavors the negative effect he brings about does so intentionally (Knobe and Mendlow 2004). This is in line with what Harman (1976) says about the sniper case mentioned in the introduction and with what we said about disfavoring an outcome above.

NRH does not provide any basis for ascribing intentionality in the help condition (at least not in the absence of a pro-attitude from the agent). Hence, the second half of the hypothesis is simply this:

NRH*–Help: The chance someone attributes intentionality to the agent is independent of how much she should care about the beneficial side effect.

The next question we face is how to test NRH*. After all, we do not have any data on how much participants take the agent to care, and how much weight they attribute to the harmful side effect. Note also that, in the chairman scenario, the agent claims not to care about the environment *at all*. This could mean that those who participated in the experiment placed his attitude toward the extreme point (although people's perceptions of this might differ). But we have neither data on the degree of indifference attributed to the agent, nor on the weight of the relevant normative reasons. Therefore, we cannot empirically evaluate NRH* directly.

Our data on blame turn out to be useful here. It seems plausible to say that, *ceteris paribus*, the less someone cares about a harmful side effect, the more she will be blamed. It also seems unobjectionable to say that, *ceteris paribus*, the worse the effect is, the more blameworthy the agent is. And it stands to reason that this is reflected in the amount of blame people actually ascribe to the agent. Given these two claims, the idea that comes in sight is that the amount of blame people attribute depends on the extent to which they see a discrepancy between how much the agent *should* care and how much she *actually* cares—between the weight of the reason and the agent's appreciation of it.

The upshot is that NRH* can account for HELP and HARM in a manner that is not ad hoc. Moreover, in contrast to MVH*, NRH* is sensitive to the evaluative perspective of the agent (which means that it can account for the Nazi Germany scenario). The other hypotheses considered in this paper cannot explain the asymmetry we uncovered, either because they cannot plausibly be formulated in gradual terms (NVH), or because they are symmetrical (the graded versions of BH and DSH).

In sum, the main empirical finding reported in Section 3 is that blame attributions correlate with the chance with which intentionality is attributed (HARM), whereas praise attributions do not (HELP). In order to account for HARM, the explanatory factor needs to be a gradable property. In order to account for HELP, this property should not bear on the attribution of intentionality in the help condition. NRH* is the only viable hypothesis that exhibits this structure. As such, it is the only hypothesis we know of that can account in a principled manner for HELP as well as HARM. Further research is needed in order to put NRH* to more stringent tests. A salient question is whether it survives when the extent to which an agent cares about a harmful side effect is varied independently from the extent to which the side effect might be perceived as bad.

References

Agresti, A. (2002) *Categorical Data Analysis*, New York: Wiley-Interscience.

Adams, F. (1986) "Intention and Intentional Action: The Simple View," *Mind & Language* 1:281–301.

Alicke, M. D. (2008) "Blaming Badly," *Journal of Cognition and Culture* 8:179–186.

Bratman, M. (1987) *Intention, Plans, and Practical Reason*, Cambridge MA: Harvard University Press.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2002) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, London: Routledge.

Cokely, E. T. and Feltz, A. (2009) "Individual Differences, Judgment Biases, and Theory-of-Mind: Deconstructing the Intentional Action Side Effect Asymmetry," *Journal of Research in Personality* 43:18–24.

Cushman, F. and Mele, A. (2008) "Intentional Action: Two-and-a-Half Folk Concepts?" in J. Knobe and S. Nichols (eds) *Experimental Philosophy*, New York: Oxford University Press, pp. 171–88.

Foley, R. (1992) "The Epistemology of Belief and the Epistemology of Degrees of Belief," *American Philosophical Quarterly* 29:111–124.

Guglielmo, S. and Malle, B. F. (2010) "Can Unintended Side Effects Be Intentional? Resolving a Controversy Over Intentionality and Morality," *Personality and Social Psychology Bulletin* 36:1635–1647.

Harman, G. (1976) "Practical Reasoning," *Review of Metaphysics* 29:431–463.

Heiberger, R. and Robbins, N. (2014) "Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications," *Journal of Statistical Software* 57 (available at `http://www.jstatsoft.org/v57/i05`).

Hindriks, F. (2008) "Intentional Action and the Praise–Blame Asymmetry," *Philosophical Quarterly* 58:630–641.

Hindriks, F. (2011) "Control, Intentional Action, and Moral Responsibility," *Philosophical Psychology* 24:787–801.

Hindriks, F. (2014) "Normativity in Action: How to Explain the Knobe Effect and Its Relatives," *Mind & Language* 29:51–72.

Holton, R. (2010) "Norms and the Knobe Effect," *Analysis* 70:417–424.

Knobe, J. (2003) "Intentional Action and Side Effects in Ordinary Language," *Analysis* 63:190–194.

Knobe, J. (2004) "Folk Psychology and Folk Morality: Response to Critics," *Journal of Theoretical and Philosophical Psychology* 24:270–279.

Knobe, J. (2006) "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *Philosophical Studies* 130:203–231.

Knobe, J. (2007) "Reason Explanation in Folk Psychology," *Midwest Studies in Philosophy* 31:90–106.

Knobe, J. (2010) "Person as Scientist, Person as Moralist," *Behavioral and Brain Sciences* 33:315–329.

Knobe, J. and Mendlow, G. (2004) "The Good, the Bad, and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology," *Journal of Theoretical and Philosophical Psychology* 24:252–258.

Malle, B. F. and Nelson, S. E. (2003) "Judging Mens Rea: The Tension between Folk Concepts and Legal Concepts of Intentionality," *Behavioral Sciences and the Law* 21:563–580.

Mele, A. (2001) "Acting Intentionally: Probing Folk Notions," in B. Malle, L. Moses, and D. Baldwin (eds) *Intentions and Intentionality: Foundations of Social Cognition*, Cambridge MA: MIT Press, pp. 27–43.

Mele, A. (2003) "Intentional Action: Controversies, Data, and Core Hypotheses," *Philosophical Psychology* 16:325–340.

Mele, A. and Cushman, F. (2007) "Intentional Action, Folk Judgments, and Stories: Sorting Things Out," *Midwest Studies in Philosophy* 31:184–201.

Nadelhoffer, T. (2004) "Praise, Side Effects, and Intentional Action," *Journal of Theoretical and Philosophical Psychology* 24:196–213.

Nadelhoffer, T. (2006) "Desire, Foresight, Intentions, and Intentional Actions: Probing Folk Intuitions," *Journal of Cognition and Culture* 6:133–157.

Nado, J. (2008) "Effects of Moral Cognition on Judgments of Intentionality," *British Journal for the Philosophy of Science* 59:709–731.

Nichols, S. and Ulatowski, J. (2007) "Intuitions and Individual Differences: The Knobe Effect Revisited," *Mind & Language* 22:346–365.

Over, D. E., Douven, I., and Verbrugge, S. (2013) "Scope Ambiguities and Conditionals," *Thinking & Reasoning* 19:284–307.

Pellizzoni, S., Girotto, V., and Surian, L. (2010) "Beliefs and Moral Valence Affect Intentionality Attributions: The Case of Side Effects," *Review of Philosophy and Psychology* 1:201–209.

Pettit, D., and Knobe, J. (2009) "The Pervasive Impact of Moral Judgments," *Mind & Language* 24:586–604.

Phelan, M. and Sarkissian, H. (2008) "The Folk Strike Back: Or, Why You Didn't Do It Intentionally, Though It Was Bad For You and You Knew It," *Philosophical Studies* 138:291–298.

Pinillos, N. A., Smith, N., Nair, G. S., Marchetto, P., and Mun, C. (2011) "Philosophy's New Challenge: Experiments and Intentional Action," *Mind & Language* 26:115–139.

Scanlon, T. (1998) *What We Owe To Each Other*, Cambridge MA: Harvard University Press.

Sripada, C. (2010) "The Deep Self Model and Asymmetries in Folk Judgments about Intentional Action," *Philosophical Studies* 151:159–176.

Sripada, C. (2011) "Mental State Attributions and the Side-Effect Effect," *Journal of Experimental Social Psychology* 48:232–238.

Sripada, C. and Konrath, S. (2011) "Telling More Than We Can Know About Intentional Action," *Mind & Language* 26:353–380.

Stocker, M. (1973) "Act and Agent Evaluations," *Review of Metaphysics* 27:42–61.

Uttich, K. and Lombrozo, T. (2010) "Norms Inform Mental State Ascriptions: A Rational Explanation for the Side-Effect Effect," *Cognition* 116:87–100.
Wolf, S. (1990) *Freedom Within Reason*, Oxford: Oxford University Press.