Normativity in Action:

How to Explain the Knobe Effect and Its Relatives

Abstract. Intuitions about intentional action have turned out to be sensitive to normative factors: most people say that an indifferent agent brings about an effect of her action intentionally when it is harmful, but unintentionally when it is beneficial. Joshua Knobe explains this asymmetry, which is known as 'the Knobe effect', in terms of the moral valence of the effect, arguing that this explanation generalizes to other asymmetries concerning notions as diverse as deciding and being free. I present an alternative explanation of the Knobe effect in terms of normative reasons. This explanation generalizes to other folk psychological notions such as deciding, but not to such notions as being free. I go on to argue against Knobe that offering a unified explanation of all the asymmetries he discusses is in fact undesirable.

Frank Hindriks (f.a.hindriks@rug.nl), University of Groningen

Forthcoming in *Mind & Language*

Normativity in Action: How to Explain the Knobe Effect and Its Relatives[*]

How should we understand folk psychological notions such as acting intentionally and deciding? Recent empirical evidence reveals that normative factors influence judgments that employ such notions (see Knobe 2010a and Nadelhoffer 2011 for overviews). Joshua Knobe (2010a, 316) defends the view that 'moral considerations actually figure in the competencies people use to make sense of human beings and their actions'. This means that moral considerations legitimately influence attributions of folk psychological notions. His point of departure is the finding that has become known as 'the Knobe effect', which is an asymmetry in our use of the term 'intentional action'. Knobe discovered it when he ran his seminal chairman experiment. Consider a chairman of a company who sets out to implement a profit-maximizing business strategy. As it turns out, the strategy has an unintended effect on the environment. The chairman, however, disavows any interest in the environment. In the help condition of the experiment, the effect on the environment is beneficial; in the harm condition the side effect is harmful. Most people (82%) say that the chairman harms the environment intentionally, even though only a minority (23%) say that he helps the environment intentionally (Knobe 2003a). Apparently it makes a difference to our intentionality attributions whether an effect has moral significance or not, and whether its moral significance is positive or negative.[1]

[*] I would like to thank James Beebe, Michael Bratman, Joshua Knobe, and Paulo de Sousa for helpful discussions on the topic of this paper. The comments from audiences at the Experimental Philosophy Workshop: Attributions of Consciousness in New York (March 2011), the Institute of Cognition and Culture in Belfast (April 2011), the 7th International Symposium on Cognition, Logic and Communication: Morality and the Cognitive Sciences in Riga (May 2011), and the Experimental Philosophy workshop in Eindhoven (October 2011) were also very helpful.

[1] The Knobe effect has been replicated by Adams and Steadman (2004), Feltz and Cokely (2007), Mallon (2008), McCann (2005), Nadelhoffer (2004, 2005, 2006a, 2006b), Nichols and Ulatowski (2007), Wright and Bengson (2009), and others.

Knobe explains the asymmetry in terms of the moral valence of the effect. People attribute intentionality when the effect is bad, but not when it is good. I refer to his explanation as 'the moral valence explanation'. Knobe argues that the moral valence explanation generalizes to other asymmetries, both those concerning other folk psychological notions such as deciding and advocating, and those concerning notions such as freedom and causation. Elsewhere I have argued that the Knobe effect should not be explained in terms of moral valence, but in terms of our beliefs and responses to normative reasons (Hindriks 2008, 2010, 2011).[2] I call this 'the normative reason explanation'. Knobe (2010, 355) has criticized this explanation, claiming that it leaves us 'with a mystery as to why the impact of moral judgment is so pervasive'. The underlying idea is that it is better to offer a single, unified explanation of all asymmetries. I refer to this as 'the argument from unification'. In this paper I answer Knobe's criticism of the normative reason explanation in two steps. First, I argue that the normative reason explanation can be generalized to other folk psychological notions where needed (sections 4.1 and 4.2). Second, I argue that offering a unified explanation of all the asymmetries Knobe sets out to explain is in fact undesirable (section 4.3).

In section 1 I introduce the moral valence explanation and Knobe's argument from unification in more detail. In section 2 I consider the relation between intentional action and moral responsibility, arguing that a proper appreciation of this relation provides the key to an alternative competence explanation of the Knobe effect, my normative reason explanation. In section 3 I discuss how this explanation preserves the idea that acting intentionally is acting with a certain frame of mind. Knobe has to

[2] A normative reason is a consideration that counts in favor of an action or attitude irrespective of the agent's motivation regarding that action or attitude (Scanlon 1998, Schroeder 2008, Smith 1994).

abandon this idea, because the moral valence of an effect is external to the mindset of the acting agent. As I explain the Knobe effect in terms of beliefs about normative reasons, I do not have to pay this price. I call this 'the frame-of-mind argument'. It favors the normative reason explanation over the moral valence explanation. This argument also serves an important role in determining the extent to which an explanation of the Knobe effect should generalize to other asymmetries. Due to the fact that they concern people's frame of mind, it is attractive to explain asymmetries concerning folk psychological notions in terms of attitudes about and responses to normative reasons (sections 4.1 and 4.2). Notions such as causation and freedom do not belong to folk psychology. Hence, there is little reason to restrict explanations of asymmetries concerning these notions to the attitudes and responses of the acting agent. Hence, too, there is no need for the normative reasons explanation to apply to them (section 4.3). The frame-of-mind argument, then, serves not only to reveal the attractions of the normative reason explanation over the moral valence one, but also to rebut the argument from unification that Knobe has used to criticize the normative reason explanation.

1. Knobe's Unifying Competence Theory: Moral Valence

Many commentators regard the Knobe effect as surprising or puzzling, and many believe that it is a bias in our attributions of intentionality.[3] Instead of playing a constructive role, normative factors distort our judgments about intentional action, or

[3] See Adams and Steadman (2004a), Guglielmo, Monroe, and Malle (2009), Knobe (2003a, 2010a), McCann (2005), Nadelhoffer (2004b), Pellizoni, Girotto, and Surian (2010), and Wright and Bengson (2009) for the claim that the Knobe effect is surprising or puzzling. Alicke (2008), Malle and Nelson (2003), McCann (2005), Mele (2001), Nadelhoffer (2004b, 2004c, 2005, 2006a), Nado (2008) are among those who regard the effect as a bias.

so they argue. Many of them also believe that the fact that people blame the chairman for harming the environment leads them to say that he harmed the environment intentionally.[4] As they are not inclined to praise the chairman for helping the environment, people are not motivated to impute intentionality to him. Knobe (2010) refers to this as 'the Motivational Bias Hypothesis'.[5] He criticizes this approach, arguing that no positive evidence for this hypothesis has been produced (ibid., 323).[6] It is important to realize, however, that all those who regard the Knobe effect as puzzling have reason to embrace a bias explanation. Their puzzlement is what motivates them to explain the effect away. The effect consists in an asymmetry concerning our intentionality attributions. The two versions of the scenarios that generate the asymmetrical judgments, however, are symmetrical in all respects except one, which many regard as irrelevant to intentional action. This makes it difficult to avoid the conclusion that it is a bias.

Knobe does not account for the puzzlement. His main reason for not regarding the Knobe effect as a bias is that it is very widespread. The effect is not only displayed by random adult U.S. Americans, it is also displayed by children as young as four (Leslie et al. 2006), by people who speak Hindi (Knobe and Burra 2006), and by people who suffer from an affective deficit (Young et al. 2006). These findings make it attractive, Knobe believes, to opt for what he calls 'a competence theory'. A

[4] Mele (2001) suggested that people assume that blame requires intentionality. He retracted this claim in response to counter-evidence (Mele 2003). All bias theorists mentioned in note 3 invoke blame in one way or another.
[5] In addition to the Motivational Bias Hypothesis, Knobe (2010) discusses what he calls 'the Conversational Pragmatics Hypothesis'. On this hypothesis, the Knobe effect is to be explained in terms of conventions governing conversation. See Knobe (2004, 2010), Nadelhoffer (2006b), and Nichols and Ulatowski (2007) for criticisms.
[6] Sripada and Konrath (2011) present evidence *against* the claim that intentionality attributions are mediated by blame. Pellizoni, Girotto, and Surian (2010) show that participants also attribute intentionality in the harm condition when the agent is deemed to be responsible on independent grounds.

competence theory is a theory on which the Knobe effect reveals something important about the formation of competent or reliable judgments about intentional action.[7]

Knobe (2010, 318) suggests that it would speak in favor of the bias approach if the Knobe effect were an isolated phenomenon restricted only to the notion of intentional action. However, normative factors have an effect on our intuitions concerning other notions in a similar way to the Knobe effect. Our responses concerning notions such as deciding and advocating are also sensitive to the moral valence of the effect at issue. When asked whether the chairman decides to harm (help) the environment, people tend to give a positive (negative) answer (Pettit and Knobe 2009). A similar effect is observed in relation to the notion of 'being in favor of'. People tend to reject the claim that the chairman is in favor of helping the environment, whereas they are neutral about this when the environment is harmed (ibid.). This suggests that the Knobe effect is an instance of a more general phenomenon. The effect of normative factors on our intuitions is pervasive and concerns a wide range of folk psychological notions, including desiring, advocating, and opposing (see section 4).

If these effects are indeed instances of one and the same phenomenon, it should be possible to provide a unified explanation for them all. Knobe has proposed such an explanation: he argues that the moral valence of an effect influences our intuitions concerning the notions at issue. Knobe's (2003a, 2003b, 2006) claim used to be that, when the moral valence of the effect is negative, people will be more inclined to conclude that these notions apply than when the effect is morally neutral. The more sophisticated picture that Knobe (2010, Pettit and Knobe 2009) has developed more recently starts from the idea that people represent an agent as having

[7] See Sripada (2010) and Holton (2010) for other competence theories.

an attitude towards the effect, which can be located on a continuum running from strongly opposed to strongly in favor. When the agent's attitude is deemed to be more towards the relevant extreme of the continuum than the default for the applicability of the notion, that notion is taken to apply to the case at hand. And the default position is sensitive to the moral valence of the effect.

Let me illustrate this for the case of intentional action. When an effect is morally neutral and the agent is indifferent with respect to its occurrence, people will come to the conclusion that he does not bring it about intentionally. This is because, Knobe assumes, a positive attitude is required for the ascription of intentionality. In case of a morally laden effect, things are different. People are expected to have a positive attitude with respect to a morally good effect, and a negative attitude towards a bad effect. This affects the threshold. When the effect is morally good, the default shifts towards the pro side of the continuum; when it is bad, the default shifts towards the con side. As a consequence, an attitude of indifference can be located on different sides of the threshold depending on the moral valence of the object. In case of a good effect, indifference does not reach the threshold, because the threshold is even higher than in morally neutral cases. In case of a bad effect, however, the threshold is rather low. Indifference is in fact above the threshold. This means that indifference does warrant the ascription of intentionality.

Knobe proposes that this psychological process underlies the effects concerning all the notions at issue. He criticizes rival explanations of particular effects, including the normative reason explanation I proposed for intentional action in my (2008), for not being sufficiently general: 'all of them seem to leave us with a mystery as to why the impact of moral judgment is so pervasive' (Knobe 2010b, 355). He maintains that 'what we really need is not a separate theory for each of the

separate concepts but rather unifying theories of the underlying processes' (ibid.). This claim forms the core of Knobe's argument from unification, which can be spelled out in more detail as follows. Many notions exhibit an asymmetry due to some normative factor. An explanation that accounts for all of them is to be preferred to one that explains only some of them. The reason for this is that the former provides more understanding than the latter. This argument is plausible to the extent that the various effects really are instances of one and the same phenomenon. If they are not, then it is artificial to make them fit one and the same explanation. It is not a trivial matter, however, to determine whether all of the asymmetries are instances of one and the same phenomenon. The frame-of-mind argument presented in sections 3 and 4 reveals that they do not. In order to appreciate its force, it will be useful to take a closer look at the normative reason explanation. In section 2 I discuss how normative reasons bear on moral responsibility and how an appreciation of this can contribute to explaining the Knobe effect.

2. Blame, Praise, and Intentionality

Recent findings reveal that moral valence as such cannot explain the asymmetry: the intentionality ratings in the harm condition are substantially lower when the chairman is said to care about the environment.[8] This suggests that the agent's indifference is important. His indifference as such, however, cannot be the explanatory factor either, because in non-moral scenarios an effect with respect to which the agent is indifferent

[8] Guglielmo and Malle (2010) contrast the indifferent chairman to a CEO who regards it as unfortunate that the environment will be harmed. The intentionality ratings go down from 87% to 40% of the participants (see also Mele and Cushman 2007 and Phelan and Sarkassian 2008). See section 3.2 for discussion.

is not brought about intentionally.[9] A natural suggestion to make at this point is that the key to explaining the Knobe effect lies in the fact that *the agent is indifferent with respect to a morally significant effect*.

In order to understand what is so special about indifference with respect to a morally significant effect, it is important to note that the asymmetry in our judgments about intentional action is paralleled by an asymmetry concerning judgments about moral responsibility, which I shall refer to as 'the praise-blame asymmetry'.[10] People blame the chairman for harming the environment, but do not praise him for helping the environment.[11] The thing to appreciate is that how indifference is to be evaluated depends on whether it concerns a harmful or a beneficial effect. Indifference is a flaw in both cases. It only provides ground for blame, however, and not for praise. The reason for this is that an agent is praiseworthy only when she aims at bringing about a good effect, whereas blameworthiness does not require the agent to aim at bringing about a harmful effect (Stocker 1973, 60; Wolf 1990, 80 and 84; Scanlon 1998, 271).

The point can also be formulated in terms of normative reasons, which are considerations that should motivate an agent, regardless of whether they do or not. Responsibility attributions depend on whether what actually motivated an agent was in line with what (she realized) should have motivated her, i.e. with the relevant normative reasons (Arpaly 2002, 231). A misalignment between these two blocks

---

[9] Nadelhoffer (2006) presents a scenario in which a hunter shoots a deer realizing that the sound of gunfire will cause an eagle to fly away. The hunter does not care about this. Only 35% of the participants answer that the hunter intentionally causes the eagle to fly away (ibid., 146).

[10] I introduced the term 'praise-blame asymmetry' in my (2008). Wright and Bengson (2009) also use it in order to explain the Knobe effect. They present evidence for a bi-directional relation between intentionality and responsibility attributions. In my (2008) I argue instead that intentional action and moral responsibility have reasons as a common denominator. This can explain the bi-directionality Wright and Bengson (2009) find.

[11] $M = 1.4$ in the help condition / $M = 4.8$ in the harm condition on a scale from 0 to 6 where 0 designates no and 6 a lot of praise/blame (Knobe 2003a).

praise, but not blame. An agent who is indifferent with respect to a morally significant effect is insufficiently responsive to it and ignores a normative reason. I shall say that a normative reason is negative when it counts against the intended action, and positive when it counts in favor of the intended action. The key idea, then, is that ignoring a negative normative reason licenses blame, whereas ignoring a positive normative reason does not support praise. In light of this the praise-blame asymmetry is perfectly natural.

How is this relevant to the Knobe effect? Information concerning the motives of an agent plays an important role in determining whether someone is worthy of praise or blame. Of particular importance is whether the agent's motives were in line with the normative reasons that applied to the situation. Was the agent motivated to bring about the beneficial effect, or to avoid the harmful effect? In the case of the chairman the answer to both of these questions about motivation is negative. These answers support the ascription of blame, but not praise. In light of this, it is useful to have a notion concerning (a lack of) motivation that is asymmetrical in the same way as our attributions of praise and blame. My hypothesis is that this is our notion of intentional action. This notion flags a discrepancy between normative reasons and motivation only in case of a harmful effect. In this way it facilitates attributions of responsibility. The warrant that intentionality attributions provide is, of course, defeasible. In particular, it might be that an agent is in a position to justify or excuse her action, and avoid being blamed. In light of this, the conclusion that can be drawn is that in moral contexts intentionality judgments serve to indicate that, if an agent does indeed deserve praise, this is due to a proper appreciation of the reasons involved in the situation, and when an agent deserves blame, this is due to a flawed appreciation of those reasons.

To the extent that it makes the feeling of surprise concerning the Knobe effect go away, the analysis presented here speaks in favor of the competence approach. Pre-theoretically moral valence seems to be irrelevant to intentional action. This is why the Knobe effect puzzles so many people (see note 3). On closer inspection, however, what explains the asymmetry is not moral valence as such, but the agent's indifference with respect to it. And it makes a lot of sense to say that someone who is indifferent with respect to a harmful effect that he brings about does so intentionally, even though this would not follow had the effect been beneficial. Putting this in terms of choice can clarify the issue further: the indifferent chairman does not choose to help the environment, but, given that he knows he should do otherwise, he does choose not to refrain from harming the environment. That is why intentionality is attributed only in the harm condition. All in all it is rather natural to explain the Knobe effect in terms of the agent's indifference with respect to a morally significant effect. I use this idea in section 3 to formulate the normative reason explanation in more precise terms. I also continue my defense of it by introducing the frame-of-mind argument.[12]

3. Moral Valence or Normative Reasons?

The chairman should treat the fact that the strategy he favors will harm the environment as a consideration that counts against implementing it. The fact mentioned constitutes a negative normative reason. The chairman foresees the harm,

[12] Nichols and Ulatowski (2007), Cushman and Mele (2008), Cokely and Feltz (2009), and Pinillos et al. (2011) present evidence for the claim that there are two or three concepts of intentional action. Only one of these is the asymmetric concept of intentional action involved in the Knobe effect. The arguments presented in this section reveal that this notion is a very useful one, because it facilitates the attribution of moral responsibility.

is not motivated to avoid it, and ignores the reason mentioned. My proposal is that ignoring a negative normative reason is a sufficient condition for intentional action.[13] It explains why people say the chairman harms the environment intentionally. For reasons explained section 2, ignoring a positive normative condition is not a sufficient condition for intentional action. Hence, people do not attribute intentionality to the chairman in the help condition of the scenario. The normative reason explanation can be made more precise in terms of the Normative Reason account of Intentional Action (*NoRIA* for short; see Lanteri 2009 for a critique). *NoRIA* is formulated in terms of an intended action ψ and an action φ that concerns an effect that is brought about by ψing. The condition of *NoRIA* that explains the Knobe effect is this:[14]

> An agent who intends to ψ, φs by ψing, and expects to φ by ψing φs intentionally if she does not care about her φing by ψing and ψs in spite of the fact that she believes her expected φing constitutes a normative reason against her ψing.

Thus, *NoRIA* describes the agent's ignoring a negative normative reason in terms of her indifference with respect to the effect, against the background of her belief that the effect does constitute a negative normative reason.[15] In other words, the explanatory factor is foreseeing a harmful effect but not caring about it.[16]

---

[13] See Hindriks (2008, 2011). Wible (2009, 176) also suggests that the fact that the agent does not care, even though he should, affects our judgments about intentionality.

[14] I have introduced the terms 'normative reason explanation' and '*NoRIA*' in my 2011. In that paper I also extend *NoRIA* to cases involving luck or a lack of control. *NoRIA* also contains conditions concerning non-normative cases of intentional action. As they do not directly bear on the topics of this paper, I do not present them here.

[15] Scaiffe and Webber (forthcoming) wonder what exactly the ground for intentionality ascription is, in my view (as presented in Hindriks 2008): the fact that

3.1 The Frame-of-Mind Argument

Now, why should the normative reason explanation (NRE) be preferred over Knobe's moral valence explanation (MVE)? A core commitment in our understanding of intentional action is that acting intentionally is acting with a certain frame of mind (Bratman 1987, Velleman 1989, Setiya 2007). As a first approximation, this means that to characterize a behavior as an intentional action is a matter of attributing intentional attitudes to the agent. I refer to the claim that an account of intentional action should honor this core commitment as 'the frame-of-mind condition'. MVE does not satisfy the frame-of-mind condition. This is apparent from Knobe's claim that the moral judgment he invokes 'could be made even in the absence of any information about this specific agent or his behaviors' (2010a, 328). To be sure, attitudes do play a role in his explanation, but only in comparison to a moral standard that is independent of the attitudes of the agent. In Knobe's view, the agent need not even be aware of that moral standard. By violating the frame-of-mind condition Knobe pays a high price: it is far from obvious that the notion he characterizes is our

---

the chairman ought to consider the environment in his deliberation, or the fact that 'the chairman *believes that* he ought to take it into consideration but still does not do so' (ibid., n2). The second fact is the explanatory factor (the first plays a role indirectly due to what I have called 'the Side-Effect Deliberation Norm', ibid., 633).

[16] One might question whether the chairman has the attitudes I ascribe to him and does indeed foresee the harm. Empirical research has uncovered what has become known as 'the epistemic side-effect effect' (Beebe and Buckwalter 2010, Alfano, Beebe, and Robinson 2012): people ascribe foresight – the belief that the intended action will generate the harm – to the agent only in the harm condition, and not in the help condition (the same holds for knowledge). The first thing to note in response is that *NRE* postulates the belief only in the harm condition, which means that it is consistent with the epistemic side-effect effect. Secondly, it might be that the explanatory factor of *NRE* – indifference with respect to harm – could also explain the epistemic side-effect effect. Given that it is hard to see how this could be a genuinely epistemic factor, the effect is probably best regarded as a bias.

commonsense notion of intentional action. In spite of the fact that he presents it as a competence theory, his account is best regarded as deeply revisionist.

NRE explains the Knobe effect without incurring this cost. Consider the indifferent chairman. First of all, the chairman is aware of the effect that his preferred business strategy has on the environment. Second, he will realize at some level that the envisaged harm should make him pause. In other words, he believes that the envisaged effect is a consideration to which he should assign negative weight in his deliberations about which strategy to implement.[17] Third, he ignores this consideration because of his indifference with respect to the environment. Thus, the frame of mind with which the chairman acts can be captured in terms of this indifference and the effect it has on the way the chairman responds to the relevant reasons. According to NRE this is exactly what explains the Knobe effect. Hence, this explanation satisfies the frame-of-mind condition.

Note that normativity enters NRE not as a separate ingredient but as one that features in the attitudes of the agent. Even if it does not motivate her, the agent is aware of the relevant normative reason. MVE invokes both the agent's attitudes and the valence judgments that the attributor makes. In contrast to the agent's attitudes, the attributor's beliefs do not play role in the folk understanding of intentional action (Malle and Knobe 1997). Given that intentional action is a frame-of-mind notion, the agent's beliefs and responses are relevant for competent attributions, while those of

[17] One can have a normative reason without being aware of it (Rosen 2003). If I am right and the chairman ignores the normative reason he has, he is aware of it. I take his awareness to play a crucial role in the attribution of intentionality. What does this imply for the chairman's moral beliefs? On a strong reading, the chairman should be taken to realize at some level that the environment has moral significance. On a weak reading, the fact that many others regard the environment as having moral significance should be reason enough for him to pause and consider the environmental effect that his action has.

attributors are not. This is the frame-of-mind argument in a nutshell. It provides a reason for preferring NRE to MVE.

Rather than being revisionist, *NoRIA* can be seen as a relatively conservative extension of existing theories of intentional action. Whereas existing theories require the agent to assign positive or negative significance to an effect, *NoRIA* also takes a failure to assign negative significance to a harmful effect to be a sufficient condition for intentional action.[18]

3.2 Empirical Evidence for the Frame-of-Mind Condition

The frame-of-mind condition is a conceptual claim about intentional action and other folk psychological notions. A number of empirical findings support the idea that this claim is indeed true. Since MVE violates this condition and NRE satisfies it, these findings favor NRE over MVE. The first finding concerns a case in which the moral evaluations of the attributor and the agent differ from one another. It concerns a chairman in Nazi Germany who decides to make some organizational changes. The vice-president of the company points out to him that by making these changes the company will be conforming to (violating) the racial identification law that is in force. The law served to identify people of certain races so that they could be rounded up and sent to concentration camps. The chairman does not care about conforming to (violating) the law.

Virtually all contemporary attributors will regard conforming to the law as bad and violating it as good. Given these evaluations, Knobe's MVE predicts that people

[18] In Hindriks (2008) I present *NoRIA* as an extension of Bratman's (1987) theory of intentional action. Harman (1976) had something like *NoRIA* in mind. Roughly, *NoRIA* says that an effect is brought about intentionally by an agent if she likes it, dislikes it, or should dislike it but does not. The third disjunct captures NRE.

will say that the chairman conforms to the law intentionally and that he does not violate the law intentionally. As it turns out, however, only a minority (30%) said that the chairman conformed to the law intentionally, whereas a majority (81%) judged that he violated the law intentionally (Knobe 2007). This finding is consistent with the prediction of NRE. On the assumption that the agent's perspective is what matters, the salient feature in the story is that the chairman has a normative reason to conform to the law. By not caring about violating it, he ignores this negative normative reason. This implies that, *pace* Knobe, he violates the law intentionally.[19]

The second finding concerns agents who are not indifferent but care about the effect of the intended action (see section 2 note 8). Mele and Cushman (2007) consider an agent who believes she has to fill in a pond but regrets that by doing so she will make a number of children sad. The mean intentionality rating they report is 3.19 on a seven-point scale, which is down from 5.79 in the chairman scenario (ibid., 194). Only 29% attribute intentionality in a scenario Phelan and Sarkissian (2008) tested concerning a city planner who feels terrible about the fact that his plan for decreasing pollution will lead to an increase in unemployment. Guglielmo and Malle (2010, 1639) contrast the indifferent chairman to a CEO who regards it as unfortunate that the environment will be harmed. The intentionality ratings go down from 87% to 40% of the participants.[20]

---

[19] See also the gay kissing and interracial sex cases in Knobe (2007). Nichols and Ulatowski (2007) observe that the racial identification scenario also constitutes evidence against the Motivation Bias Hypothesis.

[20] In non-moral scenarios most people say that a caring agent brings about the negative effect intentionally (see the sales scenario in Knobe and Mendlow 2004 and the apple tree scenario in Nanay 2010). This is due to the fact that the agent dislikes the effect, but takes this disadvantage to be outweighed by the advantage of the intended effect (see note 18); a more elaborate account of the findings can be given in terms of the negative significance condition of *NoRIA* (see Hindriks 2008 and 2011). Note that Phelan and Sarkassian (2009) present two morally charged scenarios in which there is no significant difference between a caring and an indifferent agent.

These cases reveal that moral valence as such cannot explain the intentionality attributions in morally charged situations. Even though the effects are bad, participants do not attribute intentionality to the agents. In contrast to MVE, NRE accounts for these attributions in a natural way. Given that the agents care about the effects, they do not ignore the fact that they constitute reasons against the intended actions. Apparently the agents assign significance to them without regarding them as having overriding importance. As they do not ignore a negative normative reason, the condition of *NoRIA* that accounts for the Knobe effect is not satisfied in these cases. Hence, attributors legitimately refrain from ascribing intentionality.[21]

Both of these findings undermine Knobe's attempt to explain the data by invoking a factor that is external to the agent's frame of mind. The second finding reveals that intentionality is not attributed in certain cases in which an agent weighs two conflicting normative reasons. The first finding reveals that, rather than in terms of the moral evaluation of the attributor, intentionality attributions are to be explained in terms of the normative reasons that apply to the agent. The upshot is that the frame-of-mind condition favors NRE over MVE on both conceptual (section 3.1) and empirical grounds (section 3.2).

4. A Unifying Theory?

I have argued that the Knobe effect is not as surprising as many have taken it to be (section 2). Moreover, I have criticized Knobe's claim that it is to be explicated in

[21] Knobe (2007) tests a scenario in which a terrorist defuses a bomb in order to save his son. By doing so he saves a number of Americans whom he set out to kill. Participants say that he does not save the Americans intentionally. The terrorist is not indifferent about the Americans. Instead, after 'carefully considering the matter', he regards saving his son as more important than not killing the Americans (ibid., 99-100). Hence, his case can also be explained in terms of weighing pros and cons.

terms of moral valence, arguing that it should instead be explained in terms of normative reasons (section 3). This puts me in a position to address the argument from unification that Knobe (2010b, 355) has voiced against NRE. This argument boils down to the claim that it should be regarded as a disadvantage of an explanation that accounts for the Knobe effect, if it does not also explain the other asymmetries. This presupposes that the Knobe effect is an instance of a more general phenomenon, which should make us wonder what the other instances are. Knobe takes the more general phenomenon to be that moral evaluations (legitimately) influence our judgments about apparently non-normative issues. My argument is that it is more attractive to take the overarching phenomenon to be that the way in which an agent responds to normative reasons (legitimately) influences our judgments about issues one might take to be determined by motivating reasons only.

Knobe argues that the phenomenon he has in mind extends not only to other folk psychological notions such as deciding and advocating, but also to notions outside the realm of folk psychology such as freedom and causation. I do not think the phenomenon extends beyond folk psychology. The reason for this is that the frame-of-mind condition is tailored to folk-psychological notions. Note that I do not claim that normative factors are irrelevant to freedom and causation. This may well be so. The point is just that, if they bear on them, it is unlikely that they do so in the same way.

My position can be illuminated in terms of an analogy concerning family relations. Asymmetries concerning notions such as acting intentionally and deciding are close relatives of the Knobe effect -- siblings, say. Freedom and causation, in contrast, are at best remote relatives of intentional action -- perhaps distant cousins. What they have in common is that the relevant attributions depend on beliefs about normative factors. In line with what I have said about the frame-of-mind condition in

the previous section, I argue that in the case of siblings of the Knobe effect these are the agent's beliefs about normative reasons (section 4.1). In the case of distant cousins, however, they are attributor beliefs about good or bad, or right or wrong (section 4.3). As it turns out, there is a third category. These are notions that appear to be influenced by judgments about normative factors, but which on closer inspection turn out not to be. Desiring, advocating, and opposing fall in this category. I refer to these as 'impostors' (section 4.2).

4.1 Generalizing *NRE*

I use the case of deciding to illustrate that NRE generalizes to other folk psychological notions. As mentioned in section 1, people tend to give a positive answer when asked whether the chairman decides to harm the environment, and a negative answer when asked whether he decides to help it (Pettit and Knobe 2009). If NRE generalizes to the concept at issue, people say that the chairman decides to harm the environment because he ignores a reason that counts against the business strategy he implements. The underlying thesis is that an agent decided to bring about an effect if he ignored the fact that it constituted a negative normative reason. Just as in the case of intentional action, ignoring the environment as a reason that speaks in favor of the strategy does not make people say that he decides to help it.

This explanation derives further support from the way the chairman should deliberate and how this should affect his decision. The chairman should take the environment into account in his deliberations about the business strategy. In the positive case, doing so would not have any effect on his decision. This makes it implausible to attribute the effect to him. After all, the effect did not enter into his

deliberations, and it did not affect his decision. In the negative case, deliberating about the envisaged harm should have changed his decision. His choice not to take it into account, then, had a substantial impact on his decision. This makes it natural to say that he decided to harm the environment. Deciding, then, is a sibling of the Knobe effect.

It is easy to see that NRE could generalize to other folk psychological notions that are sensitive to normative factors, such as advocating and being in favor of. Also in their case, the explanatory factor could be ignoring a negative normative reason. However, it is not obvious that they are siblings of the Knobe effect. Although the findings could be explained in terms of negative normative reasons, it is far from obvious that they should. In section 4.2 I argue that the empirical details of those findings support a different interpretation, according to which they are instead impostors.

4.2 Impostors

Knobe argues that normative factors affect attributions of a substantial number of other notions. These include favoring, advocating, opposing, desiring, and intending (Pettit and Knobe 2009, Knobe 2010). I argue that most of these notions are impostors, in the sense that our attributions of them can easily appear to depend on normative factors, even though they do not in fact do so. At the heart of my argument lies the fact that the asymmetries involved in these notions are less pronounced than the Knobe effect.

Consider the closely related notions favoring and advocating. As it turns out, people tend to reject both the claim that the chairman is in favor of helping the

environment and that he advocates it: on a seven-point scale, which ranges from complete disagreement to complete agreement, the ratings for the favoring and advocating are $M = 2.6$ and $M = 2.8$ respectively (Pettit and Knobe 2009: 593). They are, however, neutral about this when the environment is harmed: the respective ratings are $M = 3.8$ and $M = 4.1$ (ibid.: 593).[22]

Arguably, being in favor of as well as advocating an envisaged event require a pro-attitude regarding that event. The chairman, however, claims not to care about the environment. This suggests that he does not have a pro-attitude towards the environment, and reveals that he is not motivated to benefit it in the help condition of the scenario. In light of this, it is unsurprising that people deny that he is in favor of or advocates helping the environment. Significantly fewer people reject the claim that the chairman is in favor of or advocates harming the environment. In fact, they neither agree nor disagree with the claim (the ratings in the harm conditions do not differ significantly from the midpoint of the scale; Knobe, personal communication). This reveals that the asymmetry under consideration differs from the Knobe effect: whereas a majority does attribute intentionality in the harm condition, most people are neutral about whether or not the chairman is in favor of or advocates harming the environment.[23]

Note that the chairman's claim not to care about the environment is open to interpretation. It could mean that he is indifferent in the sense of being neutral. He does not care whether it is helped or harmed. An alternative interpretation is that he would actually welcome any harm done to the environment, even though he does not

[22] As is to be expected, the data for opposing show a pattern that is the reverse of the notions discussed so far: $M = 2.3$ for the harm version, $M = 3.4$ for the help version (Pettit and Knobe 2009: 600).

[23] Almost all experiments concerning intentionality report forced-choice responses. Mele and Cushman (2007: 193) use a seven-point Likert scale and find a rating of $M = 5.79$ for the harm condition.

set out to harm the environment himself. Indifference would then mask a negative attitude, an attitude that does not play a role in his deliberation and does not bear on his decision. Now, people do not conclude that the chairman is in favor of or advocates harming the environment. This suggests they opt for the neutral interpretation of indifference. Pettit and Knobe discuss data concerning desiring that bear on this issue. Subjects are asked whether the chairman desires to harm or help the environment. The ratings are $M = 3.4$ and $M = 1.6$ respectively.[24] Whereas the rating of the chairman's desire to help is very low, the rating of his desire to harm the environment is intermediate. This confirms my suggestion that people opt for the interpretation of indifference as neutrality.

The asymmetries involved in favoring and advocating (as well as for desiring and opposing; see note 22) can all be explained in terms of indifference as neutrality. Someone who is neutral with respect to the environment obviously does not favor or advocate helping it, because that would require a pro-attitude regarding the environment. It would not be plausible to infer that such a person favors or advocates harming the environment. However, there is also little basis for denying these claims. After all, the person under consideration does not have a pro-attitude towards the environment. In line with this, people remain neutral on this issue. The upshot is that the asymmetries can be explained without invoking normative factors.

What is more, MVE cannot explain these asymmetries. According to MVE, the threshold for applying a notion is higher for good actions than it is for bad ones (and mutatis mutandis for opposing). In case of morally bad actions, even an attitude

[24] Pettit and Knobe (2009) take these data from an as yet unpublished paper by Tannenbaum, Ditto, and Pizarro. Guglielmo and Malle (2010) report similar findings ($M = 3.6$ and $M = 1.5$ for the harm and help conditions respectively). They argue that the desire ratings explain the intentionality attributions. I take the desire ratings in the harm condition to be too low – too close to the midpoint of the scale – for this to be plausible.

that is located slightly towards the con side would justify the verdict that the notion applies. In the case of favoring, this boils down to the claim that someone who is not clearly against harming the environment is actually in favor of it. However, people do not in fact claim that the chairman favors harming the environment. As a consequence, MVE conflicts with what is observed.

The notions discussed thus far in this section exhibit a consistent pattern. The ratings for harm are neutral, and the ratings for help are low (the reverse holds in the case of opposing). That pattern is distinct from the Knobe effect, which involves a high rating for harm and a low one for help. In other words, the contrasts involved in the notions at issue in this section are less pronounced than that involved in acting intentionally. Pettit and Knobe (2009, 593) downplay this difference when they claim that 'the significance of the concept of intentional action is simply that people fall somewhere near the midpoint in their willingness to apply this concept in cases of morally neutral side-effects', which makes the asymmetry 'especially easy to detect'. I have in effect argued against this claim: only those notions with ratings that differ significantly from the midpoint require a special explanation in terms of normative factors. Pettit and Knobe conclude: 'there is now good reason to believe there are no concepts anywhere in folk psychology that enable one to describe an agent's attitudes in a way that is entirely independent of moral considerations. The impact of moral judgments, we suspect, is utterly pervasive' (Ibid., 603). I believe we have good reason to reject this claim. Many of the empirical findings are better viewed as impostors rather than as relatives of the Knobe effect.[25]

[25] In light of this it is unsurprising that very view people claim that the chairman intends to harm or help the environment – only 29% do so in the harm condition (Knobe 2004).

As the frame-of-mind condition holds for all the notions discussed in this section, it could be that NRE explains the relevant asymmetries. However, because the asymmetries are less pronounced there is no reason to appeal to NRE. So there are empirical reasons for not regarding these asymmetries as siblings of the Knobe effect. Hence, a unified explanation is uncalled for. Knobe cannot use the fact that NRE does not explain these asymmetries to his advantage, because MVE does not apply to these asymmetries either. In light of this, the argument from unification fails insofar as these notions are concerned.

4.3 Causation, Freedom, and Moral Evaluation

It is not uncommon to regard the distant cousins of the Knobe effect, in particular causation and freedom, as non-normative notions. Knobe and his colleagues (Knobe and Fraser 2008, Phillips and Knobe 2009), however, have shown that normative factors play a role in how they are used. These notions differ from the ones discussed thus far in that they are not frame-of-mind notions. As a consequence, factors other than the attitudes of the agents could be explanatory. This opens up two new possibilities. Normative factors might play a direct role in the explanation of these asymmetries, or their role could be mediated by the attitudes of the attributor.

As it turns out, causation and freedom are sensitive not to what is good or bad, but to what is right or wrong. The main example used for causation concerns a philosophy department in which administrative assistants are allowed to take pens from the desks of the receptionist, whereas faculty members are supposed to buy their own. In spite of reminders, faculty members sometimes take pens from the receptionist's desk. Now suppose that at some point the receptionist cannot write

down an important message, because there are no pens left on her desk. Earlier that day both an administrative assistant and Professor Smith took a pen from her desk. Who caused the problem? Even though both persons contributed to the problem equally, people tend to say that Professor Smith caused the problem (Knobe and Fraser 2008). Knobe suggests that 'people's judgment that the professor is doing something wrong is somehow affecting their intuitions about whether or not the professor *caused* the events that followed' (Knobe 2010a, 319-20).[26]

Now, I think this claim is on the right track. One thing to note, however, is that the explanation Knobe provides is not a straightforward generalization of his moral valence explanation for the simple reason that moral valence plays no role in it at all. Furthermore, rather than the attitudes of the acting agent, Hitchcock and Knobe (2009) invoke moral evaluations made by the attributor. This makes a lot of sense, because causing is a more objective notion than deciding and acting intentionally. In light of this, it would be implausible if the attitudes of the acting agent bore systematically on what she caused. The upshot is that causing does not meet the frame-of-mind condition.

I take it that the idea is that whether someone caused something depends in part on whether what she did was right or wrong. The moral evaluations of the attributor enter the explanation of our causal judgments because the attributor's best bet will be that things are as she believes them to be. So the attributor's attitudes enter the explanation indirectly. Does any of this reflect negatively on NRE? More specifically, does the fact that NRE does not apply to causation provide a legitimate basis for regarding it as too narrow (Knobe 2010, 355)? No, it does not. One reason

[26] Hitchcock and Knobe (2009) offer a more in-depth analysis of causation in terms of counterfactual reasoning and norm violation. See Hart and Honoré (1985) for an earlier defense of the idea that normative factors bear on causation. Menzies (2010) argues that their account is to be preferred to that of Hitchcock and Knobe.

for this is that the MVE does not apply either. Another reason is that causing is not a frame-of-mind notion, which means that it is to be expected that it needs to be accounted for in a different way. The upshot is that Knobe's argument from unification also fails for this notion.

The scenario with which the impact of moral judgments on judgments about freedom was tested concerns a hospital in which the chief of surgery orders a doctor to prescribe Accuphine to a patient (Phillips and Knobe 2009). The rules of the hospital are such that the doctor has to follow this order. In the help condition of this scenario the doctor dislikes the patient and does not want her to be cured. Realizing that this would result in her recovery, he prescribes the drug nevertheless. In the harm condition of this scenario the doctor really likes the patient and wants her to be cured. Realizing that this would result in her death, he prescribes the drug nevertheless. People are then asked whether the doctor was free or forced to do what he did. As it turns out, people are substantially more inclined to say that the doctor was forced to do what he did in the help condition as compared to the harm condition.

Phillips and Knobe also wanted to know whether moral judgments influenced people's intuitions about whether it was open to the doctor not to prescribe the drug. They presented their subjects with the following claim: "Given the rules of the hospital, the doctor did not really have the option of not prescribing Accuphine." Participants were substantially more inclined to say that the doctor did not have that option in the help condition than the harm condition. Phillips and Knobe appeal to judgments about right and wrong to explain the difference, and conclude: '[O]ne of the factors that can make an option seem more "closed" is that it is regarded as morally wrong, and one factor that can make an option seem more "open" is that it is regarded as morally right' (ibid., 35). Given the connections between the notion of

freedom and those of being forced and being open, Phillips and Knobe suggest that 'people's ordinary understanding of freedom is wrapped up in a fundamental way with *moral* considerations' (ibid.).

Again, I think this is on the right track.[27] To be sure, this is only one of several plausible renderings of the term 'freedom'.[28] However, the evidence Phillips and Knobe present reveals that we do employ the term 'freedom' in this way in certain contexts. Judgments about what is open to an agent or eligible to him are moral evaluations. I agree that such evaluations should play a role in an explanation of the responses that Phillips and Knobe (2009) collected. However, the caveat I made concerning causation needs to be made here as well. The fact that moral evaluations are the best candidate for explaining certain judgments concerning freedom does not mean that folk psychological judgments are to be explained in the same terms. In fact, I take myself to have presented powerful arguments to the contrary. It is eminently plausible that the normative factors that explain our freedom attributions differ from those that account for our intentionality ascriptions. The former concept is more objective than the latter in the sense that intentional action meets the frame of mind condition, whereas freedom does not. An option's being open or closed does not depend on what an agent wants or believes, but on what would be right or wrong for her to do (consequently, attributions will depend on the beliefs that attributors ascribe to the agent and on what attributors believe is right or wrong, respectively). Only the latter notions are to be explained in terms of moral evaluations. This makes causation

[27] The idea that normative factors bear on freedom is not new. Similarly to Phillips and Knobe (2009), Benn and Weinstein (1971) invoke the notion of an option being open to a person. They claim that an attribution of unfreedom 'contains *an evaluation*', which is the claim that it would be unreasonable to expect anyone to perform the relevant action in the circumstances at issue (ibid., 208).

[28] Definitely since Isaiah Berlin's (1969) 'Two Concepts of Liberty', it has been widely recognized that there are several notions of freedom (see also Skinner's 2002 'A Third Concept of Liberty').

and freedom distant cousins of deciding and acting intentionally at best. And again the upshot is that Knobe's argument from unification fails.[29]

5. Conclusion

I have argued that the Knobe effect is to be explained in terms of normative reasons. My proposal is that an agent brings about an unintended effect of her action intentionally if she ignores a normative reason concerning the effect that counts against performing the intended action. This normative reason explanation generalizes to the siblings of the Knobe effect. Deciding may be a sibling of the Knobe effect. Despite initial appearances, however, there is little reason to regard advocating, opposing, and desiring as siblings of the Knobe effect. These asymmetries are less pronounced than that involved in acting intentionally. The moral valence of the effect need not play a role in their explanation. They can be explained in terms of the agent's indifference as such. As normative factors play no role in their explanation, they are best regarded as impostors.

Asymmetries in our judgments about causation and freedom are to be explained in a different way, or so I argue. These distant cousins of the Knobe effect are affected by moral evaluations about right and wrong, rather than about good and bad. A more fundamental difference, however, is that the attitudes of the agent play no role in their attribution. Instead, attributors rely on their own judgments about normative factors when evaluating whether notions such as causation and freedom apply. In contrast to more subjective notions such as intentional action, they do not

[29] Phillips, Misenheimer, and Knobe (2011) report experiments that suggest that asymmetries exist between happiness and unhappiness, between love and lust, and between valuing and thinking good. These may well be distant relatives of the Knobe effect.

satisfy the frame-of-mind condition. The upshot is that Knobe's argument from unification has only limited application.

References

Adams, F. and A. Steadman (2004a) 'Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding', *Analysis* 64: 173-81.

Adams, F. and A. Steadman, (2004b) 'Intentional Actions and Moral Considerations: Still Pragmatic', *Analysis* 64: 264-67.

Alfano, M., J. Beebe, and B. Robinson (2012) 'The Centrality of Belief and Reflection in Knobe-Effect Cases: A Unified Account of the Data', *Monist* 95: 264-89.

Alicke, M.D. (2008) 'Blaming Badly', *Journal of Cognition and Culture* 8: 179-86.

Arpaly, N. (2002) 'Moral Worth', *Journal of Philosophy* 99: 223-45.

Beebe, J. and W. Buckwalter (2010) 'The Epistemic Side-Effect Effect', *Mind & Language* 25: 474-98.

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

Butler, R.J. (1977) '*Analysis Competition* "Problem" NO.16', *Analysis* 37: 97.

Cokely, E.T. and A. Feltz (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 A. Mele (2008) 'Intentional Action: Two-and-a-Half Folk Concepts?' In J. Knobe and S. Nichols (eds), *Experimental Philosophy*. New York: Oxford University Press, 171-88.

Enç, B. (2003) *How We Act: Causes, Reasons, and Intentions*. Oxford: Oxford University Press.

Feltz, A. (2007) 'The Knobe Effect: A Brief Overview', *Journal of Mind and Behavior* 28: 265-77.

Guglielmo, S., A.E. Monroe, and B.F. Malle (2009) 'At the Heart of Morality Lies Folk Psychology', *Inquiry* 52: 449-66.

Guglielmo, S. and B.F. Malle (2010) 'Can Unintended Side Effects Be Intentional? Resolving a Controversy Over Intentionality and Morality', *Personality and Social Psychology Bulletin* 36: 1635-47.

Harman, G. (1976) 'Practical Reasoning', *Review of Metaphysics* 29: 431-63.

Hindriks, F. (2008) 'Intentional Action and the Praise-Blame Asymmetry.' *Philosophical Quarterly* 58 (233): 630-41.

Hindriks, F. (2010) 'Person as Lawyer : How Having a Guilty Mind Explains Attributions of Intentional Agency', *Behavioral and Brain Sciences* 33: 339-40.

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

Holton, R. (2010) 'Norms and the Knobe Effect', *Analysis* 70, 417-24.

Knobe, J. (2003a) 'Intentional Action and Side Effects in Ordinary Language', *Analysis* 63: 190-94.

Knobe, J. (2004a) 'Intention, Intentional Action and Moral Considerations', *Analysis* 64: 181-87.

Knobe, J. (2004b) 'Folk Psychology and Folk Morality: Response to Critics', *Journal of Theoretical and Philosophical Psychology* 24: 270-79.

Knobe, J. (2006) 'The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology', *Philosophical Studies* 130: 203-31.

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

Knobe, J. (2010a) 'Person as Scientist, Person as Moralist', *Behavioral and Brain Sciences* 33: 315-29.

Knobe, J. (2010b) 'The Person as Moralist Account and Its Alternatives', *Behavioral and Brain Sciences* 33: 353-65.

Knobe, J. and A. Burra (2006) 'Experimental Philosophy and Folk Concepts: Methodological Considerations', *Journal of Cognition and Culture* 6: 331-42.

Knobe, J. and G. Mendlow (2004) 'The Good, the Bad, and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology', *Journal of Theoretical and Philosophical Psychology* 24: 252-58.

Lanteri, A. (2009) 'Judgments of Intentionality and Moral Worth: Experimental Challenges to Hindriks', *Philosophical Quarterly* 59: 713-20.

Malle, B. and J. Knobe (1997) 'The Folk Concept of Intentionality', *Journal of Experimental Social Psychology* 33: 101-21.

Malle, B. and S. Nelson (2003) 'Judging Mens Rea: The Tension between Folk Concepts and Legal Concepts of Intentionality', *Behavioral Sciences and the Law* 21: 563-80.

Malle, B.F. and S.E. Nelson (2003) 'Judging *Mens Rea*: The Tension between Folk Concepts and Legal Concepts of Intentionality', *Behavioral Sciences and the Law* 21: 563-80.

McCann, H.J. (2005) 'Intentional Action and Intending: Recent Empirical Studies', *Philosophical Psychology* 18: 737-48.

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, 27-43.

Mele, A. (2003) 'Intentional Action: Controversies, Data, and Core Hypotheses', *Philosophical Psychology* 16: 325-40.

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

Nadelhoffer, T. (2004a) 'The Butler Problem Revisited', *Analysis* 64: 277-84.

Nadelhoffer, T. (2004b) 'Praise, Side Effects, and Intentional Action', *Journal of Theoretical and Philosophical Psychology* 24: 196-213.

Nadelhoffer, T. (2004c) 'Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow', *Journal of Theoretical and Philosophical Psychology* 24: 259-69.

Nadelhoffer, T. (2005) 'Skill, Luck, Control, and Intentional Action', *Philosophical Psychology* 18: 341-52.

Nadelhoffer, T. (2006a) 'Desire, Foresight, Intentions, and Intentional Actions: Probing Folk Intuitions', *Journal of Cognition and Culture* 6: 133-57.

Nadelhoffer, T. (2006b) 'On Trying to Save the Simple View', *Mind and Language* 21: 565-86.

Nadelhoffer, T. (2011) 'Experimental Philosophy of Action', In J. Aguilar, A. Buckareff and K. Frankish (eds.), *New Waves in Philosophy of Action*. New York: Palgrave McMillan.

Nado, J. (2008) 'Effects of Moral Cognition on Judgments of Intentionality', *British Journal for the Philosophy of Science* 59: 709-31.

Nanay, B. (2010) Morality or Modality? What Does the Attribution of Intentionality Depend On?', *Canadian Journal of Philosophy* 40: 25-40.

Nichols, S. and J. Ulatowski (2007) 'Intuitions and Individual Differences: the Knobe Effect Revisited', *Mind & Language* 22: 346-65.

Pellizoni, S., V. Girotto, and L. Surian (2010) 'Beliefs and Moral Valence Affect Intentionality Attributions: the Case of Side Effects', *Review of Philosophy and Psychology* 1: 201-09.

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

Phelan, M. and H. Sarkissian (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: 291-98.

Phelan, M. and H. Sarkissian (2009) 'Is the 'Trade-off Hypothesis' Worth Trading For?', *Mind & Language* 24: 164-80.

Phillips, J., L. Misenheimer, and J. Knobe (2011) 'The Ordinary Concept of Happiness (and Others Like It)', *Emotion Review* 3: 320-22.

Pinillos, N.A., N. Smith, G.S. Nair, P. Marchetto, and C. Mun (2011) 'Philosophy's New Challenge: Experiments and Intentional Action', *Mind & Language* 26: 115-39.

Rosen, G. (2003) 'Culpability and Ignorance', *Proceedings of the Aristotelian Society* 103: 61-84.

Scaiffe, R. and J. Webber (forthcoming) 'Intentional Side-Effects of Action', *Journal of Moral Philosophy*.

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

Scanlon, T. (2008) *Moral Dimensions. Permissibility, Meaning, and Blame.* Cambridge (MA): Harvard University Press.

Schroeder, M. (2008) *Slaves of the Passions*. New York: Oxford University Press.

Setiya, K. (2007) *Reasons without Rationalism*. Princeton: Princeton University Press.

Sher, G. (2006) *In Praise of Blame*. Oxford: Oxford University Press.

Smith, M. (1994) *The Moral Problem*. Oxford: Blackwell.

Sripada, C. (2010) 'The Deep Self Model and Asymmetries in Folk Judgments about Intentional Action', *Philosophical Studies* 151: 159-76.

Sripada, C.S. and S. Konrath (2011) 'Telling More Than We Can Know About Intentional Action', *Mind & Language* 26: 353-80.

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

Velleman, J.D. (1989) *Practical Reflection*. Princeton: Princeton University Press.

Wible, A. (2009) 'Knobe, Side Effects, and the Morally Good Business', *Journal of Business Ethics* 85: 173-78.

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

Wright, J.C. and J. Bengson (2009) 'Asymmetries in Judgments of Responsibility and Intentional Action', *Mind & Language* 24: 24-50.

Young, L., F. Cushman, R. Adolphs, D. Tranel, and M. Hauser (2006) 'Does Emotion Mediate the Relationship Between an Action's Moral Status and Its Intentional Status? Neuropsychological Evidence', *Journal of Cognition and Culture* 6: 291-304.