RESEARCH ARTICLE

Intentional action and the frame-of-mind argument: New experimental challenges to Hindriks

(final draft, to appear in *Philosophical Explorations*)

Florian Cova[a*]
[a]*Swiss Centre for Affective Sciences, University of Geneva, Geneva, Switzerland*

Based on a puzzling pattern in our judgments about intentional action, Knobe (2003) has claimed that these judgments are shaped by our moral judgments and evaluations. However, this claim goes directly against a key conceptual intuition about intentional action – the 'frame-of-mind condition', according to which judgments about intentional action are about the agent's frame-of-mind and not about the moral value of his action. To preserve this intuition, Hindriks (2008, 2014) has proposed an alternate account of the Knobe Effect. According to his 'Normative Reason account of Intentional Action' (NoRIA), a side-effect counts as intentional only when the agent thought it constituted a normative reason not to act but did not care. In this paper, I put Hindriks' account to test through two new studies, the results of which suggest that Hindriks' account should be rejected. However, I argue that the key conceptual insight behind Hindriks' account can still be saved and integrated in future accounts of Knobe's results.

Keywords: folk psychology; Frank Hindriks; Knobe effect; intentional action; NoRIA

[*] Email: florian.cova@gmail.com

1. The Knobe Effect and the frame-of-mind argument

Studying the way people think about intentional action, Joshua Knobe (2003) uncovered a very puzzling phenomenon. Consider the following thought experiment:

> *Harm Case* – 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*.
>
> Did the chairman intentionally harm the environment?

In this case, Knobe found that 82% of the people he surveyed judged that the chairman intentionally harmed the environment.

Now, consider the following, very similar case:

> *Help Case* – 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, and it will also *help* the environment." The chairman of the board answered, "I don't care at all about *helping* the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was *helped*.
>
> Did the chairman intentionally help the environment?

In this case, only 23% of participants judged that the chairman intentionally helped the environment.

This surprising asymmetry, which people have come to call the 'Knobe Effect'[1], seems to directly contradict a common presupposition about the folk concept of intentional action: that this concept serves a purely descriptive function within folk psychology, and only describes a particular relation between

an agent's mental states and his action. Indeed, in both the *Help* and *Harm* cases, the chairman's mental states seem identical: in both cases, he knowingly brings about an outcome he did not specifically intend to cause and does not care about. Thus, it seems that the difference between the two cases cannot be explained by a mere difference in the chairman's mental states. Rather, it seems that what makes a difference is a *moral* (or, more broadly, an *evaluative*) factor: either the value of the outcome itself (harming the environment is bad, helping it is good) or the chairman's moral responsibility (the chairman deserves blame for harming the environment, but no praise for helping it).

1.1. Two kinds of accounts: evaluative accounts and attitudinal accounts

Evaluative accounts. Based on these results, some philosophers have developed *evaluative* accounts of our concept of intentional action, according to which judgments about intentional action are shaped by evaluative (or normative considerations). This suggests that the folk concept of intentional action not only has a descriptive role, but also serves a moral and evaluative function (Knobe 2006). For example, Joshua Knobe initially argued that people's judgments about intentional action were shaped by their evaluation of the outcome as good or bad (Knobe 2004). Since then, he has refined his account to claim that what shapes judgments about intentional action are expectations (including moral expectations) about the attitudes one should have towards one side-effect: an agent intentionally brings about a given side-effect only if his desire to bring about this side-effect is equal or superior to the expected desire (Pettit and Knobe 2009; Knobe 2010). For example, we think that people should desire to help the environment, which means that we (morally) expect the agent's attitude to be above indifference in the *Help* case. But we think people should be reluctant to harm the environment, which also means that we morally expect the agent's attitude to be below indifference in the *Harm* case (on scale ranging from reluctance to strong desire). This is why the same attitude on the chairman's behalf (his indifference) leads people to judge his action as intentional in the *Harm* case (in which indifference is a greater desire than the expected reluctance) but not in the *Help* case (in which indifference is a lesser desire than the expected enthusiasm).

Other evaluative accounts include accounts that refer to judgments of moral responsibility (blame and praise) and accounts that refer to norms in general (beyond moral norms). In the first category, we can find accounts that derive the asymmetry in people's judgments about intentional action from a more basic asymmetry in people's attributions of moral responsibility (blame and praise). Because praise for a given outcome requires that the agent specifically intended to bring about that outcome, while blame for an outcome does not have such requirements, people judge that the chairman deserves blame for harming the environment (in the *Harm* case) but does not deserve praise for helping the environment (in the *Help* case). This asymmetry then feeds in people's judgments of intentional action (for different reasons, depending on the account), thus causing the asymmetry (Alicke 2008; Nadelhoffer 2004a, 2004b, 2006).

In the second category, we find Richard Holton's account of the asymmetry, according to which the asymmetry in ascriptions of intentionality is directly caused by another asymmetry in norms. According to Holton, there is a fundamental asymmetry concerning norms: to intentionally violate a norm, all one needs to do is to knowingly violate it, whereas to intentionally conform to a norm one needs to be counterfactually guided by it. And since whether a norm was intentionally violated or conformed influence our ascriptions of intentionality, we have another explanation for the asymmetry in judgments about intentional action (Holton 2010).

Attitudinal accounts. Evaluative accounts of intentional action are highly revisionary. This is the reason why many have resisted such ambitious conclusions. Attitudinal accounts of the Knobe Effect are thus accounts that reject all references to evaluative factors and argue that the asymmetry in people's judgments about intentional action can be fully explained by traditional factors, such as the agent's desire to bring about the relevant outcome (Guglielmo and Malle 2010; Sripada 2010), or the agent's belief that his action will indeed bring about the relevant outcome (Alfano, Beebe, and Robinson 2012).

These accounts always start by challenging one widespread assumption about Knobe's original cases: that people actually consider the chairman to be indifferent to the side-effect in both cases. Indeed, several studies have shown that people consider that the chairman actually desire to harm the

environment (in the *Harm* case) but not to help the environment (in the *Help* case) (Pettit and Knobe 2009; Guglielmo and Malle 2010). While some consider these results as a confirmation of the fact that the Knobe Effect extends beyond judgments of intentional action and also affects attributions of desire (Knobe 2010), others see them as the ground for a simple, non-revisionary account of the Knobe Effect: people simply judge the chairman to have intentionally harmed the environment (in the *Harm* case) because they see him as actually *desiring* to harm the environment (Guglielmo and Malle 2010; Uttich and Lombrozo 2010).

However, despite their simplicity and elegance, attitudinal accounts face serious empirical challenges. Indeed, the main prediction of these accounts is that the asymmetry should disappear when the agent's attitudes (such the agent's desire to bring about the outcome) are perfectly matched between the *Harm* and *Help* case. However, this is hardly the case. For example, Guglielmo and Malle (2010) have modified Knobe's original cases to design a case in which the chairman joyfully helps the environment (*Joyful Help*). This allowed desire ratings in this case to be matched to desire ratings in the original *Harm* case: 3.67 (*SD*=1.15) in *Joyful Help* and 3.55 (*SD*=1.61) in the original *Harm* case. Still, judgments about intentional action were still higher in the original *Harm* case: 56% of participants judged the action to be intentional in the *Joyful Help* case, against 87% in the original *Harm* case. These results have been replicated in subsequent studies (Cova and Naar 2012b; Cova, Lantian, and Boudesseul, in press), with the same conclusion: even when agent's attitudes are matched between the *Harm* and the *Help* case, there still is an asymmetry in participants' judgments about intentional action, and the effect of condition on participants' judgments is mediated by participants' moral expectations.

Thus, though simple and elegant, attitudinal accounts of the Knobe Effect are unfortunately empirically inadequate. Does this necessarily mean that we should endorse evaluative accounts of the phenomenon?

1.2. The Frame-of-Mind Argument

In a recent paper, Frank Hindriks (2014) gives us one reason to reject evaluative accounts of the Knobe Effect. According to Hindriks (2014, 58), it is "a core commitment in our understanding of intentional action is that acting intentionally is acting with a certain frame of mind" and "as a first approximation, this means that to characterize a behaviour as an intentional action is a matter of attributing intentional attitudes to the agent." He refers to this as the 'frame-of-mind condition'.

The 'frame-of-mind condition' is both intuitive and attractive. Indeed, one reason the Knobe Effect has given rise to so many research and controversies (see Cova 2016 for a review) is because it is a very surprising phenomenon. When we hear for the first time about the asymmetry between the *Harm* and the *Help* case and begin to reflect about it, we tend to find it unexpected and intriguing, even if most of us share participants' intuitions about these cases. And this reaction is not limited to professional philosophers or psychologists: telling a non-specialist audience about the 'Knobe Effect' will trigger the same puzzled reactions.

These reactions do not only speak against attitudinal accounts of the Knobe Effect (according to which there is nothing surprising about the asymmetry), they also pose a serious threat to evaluative accounts. Indeed, let's think about what makes the Knobe Effect surprising: it is the fact that our judgments about intentional action differ from one case to another while we attribute the same attitudes (indifference) to the chairman in both case. It is also the fact that this asymmetry seems to be driven by moral considerations. However, something can be surprising only if it violates some previous expectations. Thus, this means that we expect judgments about intentional action not to be sensitive with factors that have nothing to do with the agent's frame-of-mind or control upon his actions. What makes the Knobe Effect (or Knobe's interpretation of this effect) surprising is that it seems to directly fly in the face of this expectation. Without it, the Knobe Effect would not be that surprising, and it probably wouldn't have stirred so much interest and debate.

According to Hindriks, the existence of such a widespread expectation (the 'frame-of-mind condition') directly speaks against evaluative accounts such as Knobe's. Indeed, the very point of these accounts is to claim that judgments about intentional action are dependent on the participants' moral evaluations, which have nothing to do with the agent's frame-of-mind. Of course, Knobe's accounts (as

well as Holton's account and other evaluative accounts) do take into account the agent's frame-of-mind: they acknowledge that the agent's attitudes and desires do matter when it comes to determine whether he intentionally brought about a given outcome. However, Hindriks seems to adopt a strong reading of the frame-of-mind condition, according to which this condition requires judgments about intentional action to depend only of the agent's mental states, and not to integrate the mental states and moral values of those who make them. Because Knobe's account involves moral judgments the agent does not even need to be conscious of (because they are the participants'), it fails to satisfy this requirement. Thus, according to Hindriks, by violating the frame-of-mind condition, "Knobe pays a high price: it is far from obvious that the notion he characterizes is our commonsense notion of intentional action. (…) his account is best regarded as deeply revisionist" (Hindriks 2014, 59).

1.3. Hindriks' NoRIA and other 'internalizing' accounts of the Knobe Effect

If attitudinal accounts of the Knobe Effect are empirically inadequate, and if evaluative accounts fail to satisfy the frame-of-mind condition, then what are we supposed to do? To answer this question, Hindriks (2008, 2011, 2014) suggests a bold move: take the best of both kinds of accounts by making judgments about intentional action dependent on the agent's evaluative judgments (rather than on the evaluative judgments of an external observer). Let's call this third kind of accounts '*internalizing* accounts' of the Knobe Effect, since they take the kinds of judgments evaluative accounts are interested in, but shift to judgments that are internal to the agent, rather than held by an external observer (the participant).

Hindriks is not the first to advance an 'internalizing account' of the Knobe Effect. Indeed, Jason Turner (2004) already had proposed an account of the Knobe Effect along the following lines:

> The following conditions are jointly sufficient for a side effect E, produced by S's action A, being intentional: (i) S knows that E will (or is likely to) occur as a result of A-ing, (ii) bringing about E counts against A-ing (from the S's perspective), and (iii) S does not try to keep E from occurring. (Turner 2004, 214).

Note the "from S's perspective" that is the mark of a proper internalizing account: while Knobe's account focus on participants' evaluative attitudes, Turner's account focus on the evaluative attitudes participants attribute *to the agent*. Thus, what makes the chairman's harming the environment intentional in the *Harm* case is not that participants consider harming the environment to be bad (and thus a reason not to start the new program) but that the chairman himself considers harming the environment to be bad (and thus a reason not to start the new program), even though he ultimately decides to start it.

Another example of internalizing account might be Edouard Machery's "trade-off hypothesis", according to which participants judge a side-effect as a cost when they conceptualize it "as a foreseen cost that the agent described in the probe incurs in order to reap a foreseen benefit" (Machery 2008, 187). Machery's account has been criticized for being ambiguous: are intentional side-effect those that count as a cost from the agent's perspective, of from the participants' (or both)? (Phelan and Sarkissian 2009). But this means there is indeed an internalized interpretation of it.

However, both accounts have faced decisive counter-examples. Indeed, in response to Turner, Knobe and Mendlow (2004) have designed the following case:

> *Terrorist* – A terrorist discovers that someone has planted a bomb in a nightclub. There are lots of Americans in the nightclub who will be injured or killed if the bomb goes off. The terrorist says to himself, "Whoever planted that bomb in the nightclub did a good thing. Americans are evil! The world will be a better place when more of them are injured or dead."
>
> Later, the terrorist discovers that his only son, whom he loves dearly, is in the nightclub as well. If the bomb goes off, his son will certainly be injured or killed. The terrorist then says to himself, "The only way I can save my son is to defuse the bomb. But if I defuse the bomb, I'll be saving those evil Americans as well... What should I do?"
>
> After carefully considering the matter, he thinks to himself, "I know it is wrong to save Americans, but I can't rescue my son without saving those Americans as well. I guess I'll just have to defuse the bomb."
>
> He defuses the bomb, and all of the Americans are saved.

Did the terrorist intentionally save the Americans?

In this case, very few people answered 'YES' (23% in Cova 2013). However it is clear that the terrorist considers saving the Americans as a reason not to defuse the bomb, and as a cost he incurs for doing so. Thus, the accounts we presented should have predicted that people would answer that the terrorist intentionally saved the Americans. But it is not the case. Can we find an internalizing account that overcomes this difficulty?

In a series of paper, Hindriks (2008, 2011, 2014) has developed a complete account of intentional action, which he dubs the 'Normative Reason account of Intentional Action' (*NoRIA*). In its first appearance (Hindriks 2008), this account also failed to account for the *Terrorist* case. However, the latest version can accommodate this purported counter-example. In its most recent version, the full account can be stated in the following way (Hindriks 2010, 2011):

An agent S φs intentionally if she intends to ψ, φs by ψing, and

(a) S expects to φ by ψing, and intends to ψ because she expects thereby to φ, or

(b) S expects to φ by ψing, and ψs in spite of the fact that she does not want to φ, or

(c) S expects to φ by ψing, does not care about her φing by ψing, and ψs in spite of the fact that she believes her expected φ-ing constitutes a normative reason against her ψing, or

(d) S hopes to φ by ψing, and ψs because (in spite of the fact that) she believes the φ-ing constitutes a normative reason in favor of (against) her ψing.

Hindriks also has advanced a shortened version of *NoRIA* for cases involving normative reasons. Here is how he summarizes it:

> 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. (Hindriks 2014, 58)

Let's now apply this account to Knobe's original cases: the chairman in the *Harm* case believes that harming the environment constitutes a normative reason not to start the new program, but does not care. Thus, he fulfils the aforementioned condition, and his harming the environment is intentional. However, this is not the case in the *Help* case, since he surely does not believe that helping the environment constitutes a normative reason not to start the new program.

What about the *Terrorist* case now? Surely, in this case, the terrorist sees saving the Americans as a reason not to defuse the bomb. So, why don't we judge that he intentionally saved the Americans? Here is how Hindriks explain cases similar to *Terrorist*:

> [Such] 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. (…) 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. (Hindriks 2014, 61)

And:

> [the] 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 to the Americans. Instead, after 'carefully considering the matter', he regards saving his son as more important than not killing the Americans. (Hindriks 2014, 61)

Thus, what explains that the terrorist does not intentionally save the Americans is that, despite the fact that he considers saving the Americans as a reason *not to* defuse the bomb, he does so regretfully, and *not indifferently*. Contrary to Turner and Machery's, Hindriks' account clearly dissociates the agent's conative attitudes towards the side-effect (desire, reluctance) from their evaluative attitudes (considering the side-effect as a reason to, or as a reason not to act). This allows him both to present the fact that the agent considers the side-effect as a reason not to act as a factor that increases intentionality ascriptions, and the fact that the agent is reluctant as a factor that decreases intentionality ascriptions. That the agent must be indifferent and that his reluctance can decrease intentionality judgments is the condition Hindriks has added to his account since its original formulation, precisely to account for cases such as *Terrorist.*

So, Hindriks' *NoRIA* seems empirically adequate. It can thus be considered superior to attitudinal accounts. Moreover, it also has an advantage over evaluative accounts: it fulfils the 'frame-of-mind condition'. This seems obvious since, according to the very formulation of *NoRIA*, judgments about intentional action only depend on the agent's attitudes and normative beliefs, and are not to be explained by reference to the moral values of those who reference to the moral values of observers who make those judgements. But this provides an advantage to *NoRIA* over evaluative accounts, such as Knobe's. In fact, all internalizing accounts, by their very definition, benefit from this advantage.

This suggests that, everything else being equal, internalizing accounts of the Knobe Effect, including *NoRIA*, should be preferred to evaluative accounts. Of course, whether everything else is equal is precisely the question: have we as much or more empirical reasons to endorse *NoRIA*, or internalizing accounts in general, than we have to endorse evaluative accounts of the Knobe Effect? This is not clear, since *NoRIA* has rarely empirically been put to test.[2] Hindriks himself has never provided original evidence in favour of his account, and only has argued that *NoRIA* is the theory that can accommodate most of the existing results. This is why the current paper aims to systematically put NoRIA to empirical test, and will try by the same occasion to evaluate the prospects of internalizing accounts in general.

2. Study 1: Dissociating participants from agents' norms

2.1. Materials, participants and methods

As stated by Hindriks (2014), one way to compare evaluative accounts of the Knobe Effect (such as Knobe's) and internalizing accounts (such as *NoRIA*) is to find a pair of cases in which participants' judgments and the agent's judgments about what counts as a reason *not* to act are diametrically opposed. The *Terrorist* case is such a case, but presents the default of having the agent reluctantly bring about the side-effect, which introduces an additional factor that might mask the impact of other considerations. To correct for this default, I designed the following pair of cases:[3]

> *Racist Bystander* – Joe is a young man who has been raised in a very racist family. His parents have taught him that only white persons should live, and that non-white persons should be eliminated. Though Joe has come to share his parents' beliefs, and think that it is a duty to rid the world of non-white people, he has never been able to share his parents' hatred of non-white people. In fact, he is completely indifferent to non-white people: neither does he care about them, nor does he particularly want to do them harm.
>
> One day, while Joe is walking through the countryside along trolley tracks, he sees that an empty runaway trolley is speeding down a set of tracks toward five white railway workmen. There is also a set of tracks branching off to the right of the main tracks. On this set of tracks is a black railway workman.
>
> If nothing is done, the trolley will proceed down the main tracks and cause the deaths of the five white workmen. It is possible to avoid these five deaths. Joe happens to be near a switch that can turn the trolley onto the side tracks. Joe sees that he can avoid the deaths of the five white workmen by hitting the switch, which will turn the trolley onto the side tracks. But the trolley will now collide with the black workman, thus causing his death.

Joe says to himself: “I definitely have to hit this switch and turn the trolley onto the side tracks. This way I will save the five workmen. Of course, it will also kill this [expletive deleted]. I should be glad about that, but I just don't care. All that matters is saving the life of these white persons.”

Joe hits the switch, and the trolley is turned onto the side tracks, thus saving the lives of the five white workmen. However, the black workman dies in the collision.

Racist Saviour – [Same first paragraph] One day, while Joe is walking through the countryside along trolley tracks, he sees that an empty runaway trolley is speeding down a set of tracks toward five white railway workmen. There is also a set of tracks branching off to the right of the main tracks. On the main tracks, working with the five white workmen, is a black railway workman.

If nothing is done, the trolley will proceed down the main tracks and cause the deaths of the five white workmen, as well as the death of the black workman. It is possible to avoid these six deaths. Joe happens to be near a switch that can turn the trolley onto the side tracks. Joe sees that he can avoid the deaths of the five white workmen by hitting the switch, which will turn the trolley onto the side tracks. But this will also save the life of the black workman.

Joe says to himself: “I definitely have to hit this switch and turn the trolley onto the side tracks. This way I will save the five workmen. Of course, it will also save this [expletive deleted]. I should be sad about that, but I just don't care. All that matters is saving the life of these white persons.”

Joe hits the switch, and the trolley is turned onto the side tracks, thus saving the lives of the five white workmen, as well as the life of the black workman.

In the *Racist Bystander* case, we morally expect Joe to be reluctant to cause the black workman’s death, however he himself thinks he has no normative reason not to cause it. Rather, he thinks he should be happy to cause such an outcome, though he doesn’t care. In the *Racist Saviour* case, on the contrary, we morally expect Joe to be happy to save the black workman’s life, though he himself thinks he has normative reason not to save his life. Moreover, despite thinking he has normative reason not to bring about this outcome, he does not care. In these conditions, *NoRIA* (and internalizing accounts of the Knobe Effect) should predict that participants will judge that Joe intentionally saved the black

workman's life in the *Racist Saviour* case, but will judge that Joe did not intentionally cause the black workman's death in the *Racist Bystander* case. On the contrary, evaluative accounts (such as Knobe's) should predict that people will judge the side-effect intentional in the *Racist Bystander* case, but not in the *Racist Saviour* case (provided that, hopefully, participants are not all in favour of the elimination of non-white people).

To determine which prediction was correct, 100 participants located in United States (M_{age}=36.6, SD_{age}=12.6; 47.0% women and 53.0% men) were recruited through Amazon Mechanical Turk and paid \$0.6 for their participation. Each participant received either the *Racist Bystander* or *Racist Saviour* case, read it, then had to rate their agreement on a 7-point scale (1="Strongly disagree", 4="Neither agree, nor disagree", 7="Strongly agree") with the 13 following statements (presented in a random order):

1) INTENTIONALLY: Joe intentionally caused the black worker's death [saved the black worker's life].
2) DESIRE: Joe wanted to cause the black worker's death [save the black worker's life].
3) RELUCTANCE: Joe was reluctant to cause the black worker's death [save the black worker's life].
4) EXPECTED DESIRE: Joe should have wanted to cause the black worker's death [save the black worker's life].
5) EXPECTED RELUCTANCE: Joe should have been reluctant to cause the black worker's death [save the black worker's life].
6) REASON FOR: Joe thought that causing the black worker's death [saving the black worker's life] counted as a reason to hit the switch.
7) REASON AGAINST: Joe thought that causing the black worker's death [saving the black worker's life] counted as a reason NOT to hit the switch.
8) SHOULD: Joe thought that he should cause the black worker's death [save the black worker's life].
9) SHOULD NOT: Joe thought that he should NOT cause the black worker's death [save the black worker's life].

10) GOOD: Joe thought that causing the black worker's death [saving the black worker's life] was a good thing.

11) BAD: Joe thought that causing the black worker's death [saving the black worker's life] was a bad thing.

12) PRAISE: Joe thought that causing the black worker's death [saving the black worker's life] would make him praiseworthy.

13) BLAME: Joe thought that causing the black worker's death [saving the black worker's life] would make him blameworthy.

Question 1 is a standard measure of intentionality. Questions 2 and 3 control for the agent's attitudes (desire and reluctance), while questions 4 and 5 measures the kind of expectations that are at the centre of Knobe (2010)'s evaluative account. Questions 6 and 7 directly measure the kind of attitudes relevant to *NoRIA* (agent's belief about normative reasons). Questions 8 to 13 test for a range of alternative internalizing accounts: 8 and 9 internalize Holton's focus on norms, 10 and 11 internalize Knobe's original proposal to understand the asymmetry in terms of the side-effect's valence, and 12 and 13 internalize Nadelhoffer and Alicke's attempt at deriving the asymmetry in judgments about intentional action from an asymmetry in the agent's moral responsibility.

Finally, participants were asked the following question:

Which of the following statements most accurately describes the situation?

- o Saving the black worker's life was the Joe's goal.
- o Saving the black worker's life was a means for Joe to reach his goal
- o Saving the black worker's life was a side-effect of Joe's attempt to reach his goal.
- o None of the above

	Racist Bystander	*Racist Saviour*	*t*	*p*	*d*
(1) Intentionally	4.30 (1.75)	3.07 (1.95)	3.00	<.01	0.66
(2) Desire	2.84 (1.61)	1.96 (1.26)	2.80	<.01	0.62
(3) Reluctance	2.78 (1.57)	4.04 (1.85)	3.30	<.01	0.73
(4) Expected desire	2.16 (1.76)	5.41 (1.44)	9.27	<.001	2.05
(5) Expected reluctance	4.43 (1.85)	2.04 (1.59)	6.32	<.001	1.40
(6) Reason for	3.32 (1.80)	1.83 (1.39)	4.23	<.001	0.95
(7) Reason against	2.16 (1.42)	3.80 (2.09)	4.07	<.001	0.90
(8) Should	3.24 (1.69)	2.04 (1.49)	3.43	<.01	0.76
(9) Should not	2.95 (1.45)	4.17 (1.84)	3.00	<.01	0.73
(10) Good	3.08 (1.74)	2.28 (1.47)	2.27	<.05	0.50
(11) Bad	2.81 (1.76)	4.26 (1.74)	3.75	<.001	0.83
(12) Praise	2.89 (1.76)	1.98 (1.74)	2.58	<.05	0.57
(13) Blame	2.59 (1.54)	3.15 (1.85)	1.47	.15	0.32

Table 1. Mean answer (and standard deviation) for each question in Study 1, in function of case.

Note: Difference between cases were tested using t-tests.

2.2. Results

I began by analysing results to the last question and excluding all participants who did not consider the relevant outcome as a side-effect. I was left with 37 participants in the *Racist Bystander* condition and

46 in the *Racist Saviour* condition. Then, I analysed participants' answers to each remaining question separately. Results (after exclusion) are presented in Table 1.

Internalizing accounts' predictions. As shown by participants' answers to question and 7 (and as expected), participants were significantly more likely to attribute Joe the belief that bringing about the side-effect counted as a reason not to hit the switch in the *Saviour* (M=3.80, SD=2.09) than in the *Bystander* case (M=2.16, SD=1.42): $t(81)=4.07$, $p<.001$, $d=0.90$. This tendency is confirmed by the fact that participants were also significantly more likely to attribute Joe the belief that bringing about the side-effect counted as a reason *to* hit the switch in the *Bystander* (M=3.32, SD=1.80) than in the *Saviour* case (M=1.83, SD=1.39): $t(81)=4.23$, $p<.001$, $d=0.95$. The same pattern can be found for the other measures of Joe's own evaluative attitudes (questions 8 to 13), except the belief that he would be blameworthy for his action (question 13): overall, participants judged Joe more likely to consider that bring about the side-effect was something he should not do / something bad in the *Saviour* than in the *Bystander* case, the only exception being his belief that acting would make him blameworthy. Thus, at first sight, internalizing accounts of the Knobe Effect, including *NoRIA*, should predict higher intentionality ratings in *Saviour* than in *Bystander*.

However, one might point out that, according to participants, Joe is reluctant to save the black worker's life in *Saviour* (M=4.04, SD=1.85), which might undermine intentionality ratings. Moreover, participants' agreement with the claim that 'Joe thought that saving the black worker's life counted as a reason NOT to hit the switch" is still under the midpoint in *Saviour* (M=3.80, SD=2.09). Taking these two factors into account, internalizing accounts might revise their prediction and predict low intentionality ratings in *Saviour*.

In both cases, however, internalizing accounts of the Knobe Effect are bound to predict low intentionality ratings in the *Bystander* case. Indeed, in the *Bystander* case, Joe (i) does not think that causing the black worker's death counts as a reason against hitting the switch / is something he should not do / is bad (all scores are below midpoint), and (ii) Joe is indifferent to the black worker's death. Thus, depending on what factors are highlighted, internalizing accounts can predict two different

outcomes for the study: low intentionality ratings in *Bystander* and high intentionality ratings in *Saviour*, or low intentionality ratings in both cases.

Intentionality ratings. Focusing on participants' answer to question 1, we can see that intentionality ratings are higher in *Bystander* (M=4.30, SD=1.75) than in *Saviour* (M=3.07, SD=1.95): $t(81)=3.00$, $p<.01$, $d=0.66$. Moreover, intentionality ratings are above the midpoint for *Bystander* and below for *Saviour*.[4] This pattern goes against *NoRIA* and other internalizing accounts predictions: (i) there is a difference in intentionality ratings between the two cases, with higher ratings in *Bystander*, and (ii) intentionality ratings are above the midpoint in *Bystander*.

What predicts intentionality ratings? To answer this question, I composed three scores out of my results: an *attitude score* (averaging question 2 and 3, reverse-coded), an *expected attitude score* (averaging question 4 and 5, reverse-coded), and *a reason-not-to score* (averaging question 6, reverse-coded, and 7). I then calculated the Pearson's product-moment correlation between each of these scores and intentionality ratings. As expected by most accounts of the Knobe Effect, there was a positive correlation between intentionality ratings and attitude scores (r=.43, $p<.001$): the greater the agent's desire to bring about the outcome was, the higher intentionality ratings were. As expected by evaluative accounts, there was also a negative correlation between expected attitude scores and intentionality ratings (r= -.25, $p<.05$): the less participants morally expected the agent to desire the outcome, the higher intentionality ratings. Finally, in direct contradiction with *NoRIA*, there was a negative correlation between reason-not-to scores and intentionality ratings (r= -.29, $p<.01$), meaning that the more participants attributed Joe the belief that bringing the side-effect counted as a reason not to act, the less they judge his action as intentional.

Thus, dissociating participants from agents' norms, as preconized by Hindriks (2014), ultimately seems to produce results that go against *NoRIA*'s predictions.

3. Study 2: Knobe's original asymmetry

Study 1 showed that there are counterexamples to *NoRIA*, and internalizing accounts of the Knobe Effect in general: participants' intentionality ratings seem better explained by participants' judgments than by the agent's evaluative attitudes. However, one might argue that all existing accounts have to face counter-examples.[5] So, this is not enough to abandon *NoRIA* or internalizing accounts in general.

Thus, it might be a little too demanding to reject *NoRIA* on the basis of a single counter-example. However it seems reasonable to require from a satisfying account of the Knobe Effect that it explains the original asymmetry between the original *Harm* and *Help* cases. In Study 2, I investigate whether NoRIA, or any internalizing account can explain the original Knobe Effect.

3.1. Materials, participants and methods

To this purpose, 120 participants located in United States (M_{age}=34.0, SD_{age}=10.2; 43.3% women and 55.8% men) were recruited through Amazon Mechanical Turk and paid $0.6 for their participation. Each participant received either Knobe's *Harm* or *Help* case (as presented in introduction). After reading the case, each participants had to rate his or her agreement with the same 13 statements as in Study 1, but adapted to the relevant case ("the chairman" rather than "Joe", "harming the environment/helping the environment" rather "causing the black worker's death/saving the black worker's life", and "starting the new program" rather than "hitting the switch") (presented in a random order).

	Harm case	*Help case*	*t*	*p*	*d*
(1) Intentionally	5.35 (1.27)	1.61 (1.22)	14.27	<.001	3.01

(2) Desire	3.78 (1.40)	1.37 (0.85)	10.10	<.001	2.13
(3) Reluctance	1.65 (1.33)	3.22 (1.77)	4.65	<.001	0.98
(4) Expected desire	1.90 (1.39)	4.98 (1.48)	10.13	<.001	2.14
(5) Expected reluctance	5.30 (1.52)	1.69 (1.22)	12.55	<.001	2.65
(6) Reason for	2.40 (1.48)	1.57 (1.27)	2.88	<.01	0.61
(7) Reason against	1.65 (1.41)	1.88 (1.35)	0.80	.43	0.17
(8) Should	3.98 (1.44)	1.78 (1.32)	7.56	<.001	1.60
(9) Should not	1.98 (1.61)	3.02 (1.78)	2.89	<.01	0.61
(10) Good	2.93 (1.37)	2.10 (1.30)	2.95	<.01	0.62
(11) Bad	2.33 (1.37)	2.47 (1.59)	0.46	.64	0.10
(12) Praise	2.58 (1.65)	1.69 (1.05)	3.13	<.01	0.66
(13) Blame	3.25 (1.46)	1.80 (1.23)	5.11	<.001	1.08

Table 2. Mean answer (and standard deviation) for each question in Study 2, in function of case. Difference between cases were tested using t-tests.

3.2. Results

I began by analysing results to the last question and excluding all participants who did not consider the relevant outcome as a side-effect. I was left with 40 participants in the *Harm* case and 51 in the *Help* case. Then, I analysed participants' answers to each remaining question separately. Results (after exclusion) are presented in Table 2.

Intentionality ratings. As expected, I replicated the original Knobe Effect and found a significant difference in intentionality ratings, such that participants were more likely to judge the chairman's action as intentional in the *Harm* case (M=5.35, SD=1.27) than in the *Help* case (M=1.61, SD=1.22): $t(89)=14.27$, $p<.001$***, $d=3.01$.[6] Now, which of the available accounts can explain this asymmetry?

Internalizing accounts. As shown by participants' answers to question 7, participants did not attribute to the chairman the belief that bringing about the side-effect counted as a reason not to start the new program either in the *Harm* (M=1.65, SD=1.41) or in the *Help* case (M=1.65, SD=1.41). In fact, there was no difference between both cases in participants' answer to this question: $t(89)=0.80$, $p=.43$, $d=0.17$ Thus, *NoRIA* should predict very low intentionality ratings in both cases, as well as no difference between the two. However, this is clearly not what I obtained.

Most of the other internalizing accounts suffer from the same problem. As can be seen in question 11, participants were not more likely to attribute the chairman the belief that the side-effect was "bad" in the *Harm* than in the *Help* case. As for norms, participants were more likely to attribute the chairman the belief that he should not bring about the side-effect in the *Help* case than in the *Harm* case.

The only internalizing account that could account for the results here is the one that focuses on praise and blame (questions 12 and 13). One could imagine an account according to which an agent intentionally brings about a side-effect when this agent believes that bringing about this side-effect would make him blameworthy or praiseworthy. This could explain the difference between the *Harm* and *Help* case. To my knowledge, such an account has never been developed or even argued for, and it might be a possibility to consider in the future.

What predicts intentionality ratings? As in Study 1, I composed three scores out of my results: an *attitude score* (averaging question 2 and 3, reverse-coded), an *expected attitude score* (averaging question 4 and 5, reverse-coded), and *a reason-not-to score* (averaging question 6, reverse-coded, and 7). I then calculated the Pearson's product-moment correlation between each of these scores and intentionality ratings. There was a positive correlation between intentionality ratings and attitude scores (r=.73, p<.001). There was also a negative correlation between expected attitude scores and intentionality ratings (r= -.80, p<.001). Finally, in direct contradiction with *NoRIA*, there was a negative correlation between reason-not-to scores and intentionality ratings (r= -.39, p<.001). Thus, once again, the relationship between intentionality ratings and reason-not-to scores was diametrically opposed to the one predicted by *NoRIA*.

3.3. Reversing internalizing accounts

The results of both studies suggest that internalizing accounts of the Knobe Effect, and in particular *NoRIA*, seem ill-tailored to explain our judgments about intentional action: participants' intentionality judgments are better predicted by participants' evaluations than by the agent's evaluative point of view. However, one could argue that I only reach this conclusion because I followed Hindriks' idea that what matters is whether the agent considers the side-effect as something bad / a reason not to act / something one should not do. But what if one decided to turn that assumption upside-down and to develop

internalizing accounts according to which an agent brings about a given side-effect intentionally if he thinks that this side-effect is something good / a reason to act / something one should do?

Such accounts would be better suited to my data, and would appropriately predict the direction of the difference in intentionality ratings between the cases. However, it is still doubtful that it could explain high intentionality ratings in the *Harm* case, given that all measures of the agent's evaluative attitudes towards bringing about the outcome were below the midpoint. Moreover, it is hard to see how such accounts would handle the *Terrorist* cases: the terrorist probably does not see saving the Americans as a good thing, or as something he should do.

In conclusion, the results of Study 1 and 2 reveal that there are not only counter-examples to *NoRIA* and other internalizing accounts of the Knobe Effect, but they also fail to account for the original Knobe Effect.

4. The dilemma

The results of Study 1 and 2 suggest that *NoRIA* and most of the possible internalizing accounts of the Knobe Effect are empirically inadequate: participants' judgments about intentional action seem to track their own moral evaluations, rather than the agent's point of view. This means that we have a dilemma: either we endorse an internalizing account of the Knobe Effect, and we respect the 'frame-of-mind condition', but sacrifice empirical adequacy, or we endorse an evaluative account of the Knobe Effect, and we have empirical adequacy, to the sacrifice of the 'frame-of-mind' condition.

Is there no way out of this dilemma? In this section, I present three ways in which we might respect Hindriks 'frame-of-mind' *and* endorse an evaluative account of the Knobe Effect.

4.1. The Knobe Effect as a bias

A first way, which has already been widely discussed in the literature, is to treat the Knobe Effect as a bias: the effect of moral (or evaluative) considerations should not be taken as manifesting participants' conceptual competence. Rather, if participants made only judgments following this competence, they

would make judgments that are in line with the 'frame-of-mind condition', and moral considerations would not play a role. Several proposals in this sense have already been made: either participants' judgments about intentional action are biased by negative affective reactions to the agent's moral character (Nadelhoffer 2004a, 2004b, 2006) and their willingness to justify and motivate their attributions of blame (Alicke 2008), or they do not mean their judgments about intentional action literally and use them to convey something else (Adams & Steadman 2004a, 2004b; Guglielmo and Malle 2010). However, most of these accounts face serious empirical challenges. To begin with, the Knobe Effect has been reproduced using pairs of cases involving bad and good outcomes but neither praise nor blame assigned to agents (such as the *Sales* cases; see Knobe and Mendlow 2004; Wright and Bengson 2008), which suggests that the effect has nothing to do with the willingness to attribute praise or blame. Moreover, the effect has been reproduced in participants suffering from emotional impairment (Young *et al.* 2006) and in participants suffering from Asperger Syndrome (Zalla and Leboyer 2011), suggesting that it is due neither to affective bias (which would be absent in people suffering from emotional impairment) nor to pragmatics (the processing of which seems impaired in people with Asperger Syndrome). Finally, participants seem very confident of their judgments in the *Harm* and *Help* cases: if we allow ourselves to doubt this certitude, how much confidence can we still place in the intuitive clarity of the 'frame-of-mind condition'?[7]

4.2. 'Intentionally' as a gradable term

However, there are other, less obvious ways of reconciling empirical adequacy and respect for the 'frame-of-mind condition'. A second way is thus to loosen our understanding of the frame-of-mind condition. We can accept that the truth-value of statements including "intentionally" is sensitive to moral and evaluative considerations without accepting that the word "intentionally" is itself evaluative, or about something else than the agent's frame-of-mind. This might seem like a trick, but let's note that Pettit and Knobe's account of the Knobe Effect begins with an analogy between the semantics of "intentionally" and the semantics of gradable adjectives, such as "cold": sentences about "cold" also seem sensitive to standard and expectations about what temperature a given beverage *should* be.

For example, let's suppose a beer and a cup of coffee are both at the temperature of 20°C. Though they are at the same temperature, we would still say that the coffee is *cold* while the beer is not. Why? Because we evaluate a beverage's coldness not only with respect to the temperature they actually are at, but also taking into consideration the temperature they should be at. A 20°C is cold because it is *colder* than it should be. But does this make "cold" a normative or evaluative notion? And does it mean that "cold" is about something else than an object's temperature? Probably not.

Against this analogy, one might argue that the case of "cold" is different, because the standard involved are not *moral* standards. However, other examples are available. Take the following case (Egré and Cova 2015):

> 10 children were present in a school when a fire broke out. 5 children survived, the other 5 died.
>
> - Would you say that many children died?
> - Would you say that many children survived?

In this case, most people answered 'YES' to the first question and 'NO' to the second. However, exactly the same number of children died and survived. This suggests that the truth-value of sentences containing "many" is impacted by moral and evaluative considerations. Should we then conclude that "many" is a moral and evaluative term? This seems a strange and unwarranted conclusion. What these results suggest, rather, is that the truth-value of sentences containing a given term can be sensitive to moral and evaluative considerations without this term being about something moral or evaluative. In this sense, there is a way of interpreting Knobe's account according to which the Knobe Effect does not make "intentionally" more moral or evaluative than terms like "cold" and "many" (see Egré 2014, for a full proposal along these lines).

4.3. 'Intentionally' as polysemous

For those who might not be convinced by this solution, there is a third way. It has been proposed repeatedly that the word "intentionally" has various meanings (Nichols and Ulatowski 2007; Cushman and Mele 2008; Sousa and Holbrook 2010; Cova, Dupoux and Jacob 2012; Lanteri 2013) and that the

Knobe Effect is mainly due to different contexts leading to different meanings being selected by participants. For example, Cova, Dupoux and Jacob (2012) have suggested that "intentionally" can have the three following meanings:

1) A positive meaning, according to which someone does something intentionally when he actively does it based on his or her desire to do it.
2) A first contrastive meaning, according to which someone does something intentionally when he does it without being forced to do it. In this sense, "intentionally" is opposed to "unwillingly" or "by force".
3) A second contrastive meaning, according to which someone does something intentionally when he does it by having full control upon his action. In this sense, "intentionally" is opposed to "by accident" or "by sheer luck".

According to Cova, Dupoux and Jacob, within the experimental context, participants' understanding of "intentionally" will be influenced by what they take to be the relevant meaning in the present context, and their conclusion will differ according to the case they are presented with. If they are told about something they (morally or statistically) expect the agent to be in favour of (a good action, or something the agent desires), then the first meaning will seem to be the most relevant. If they are told about something they (morally or statistically) expect the agent to be opposed to (a bad action, something the agent does not want), the second meaning will appear to be the most relevant. Finally, if their attention is drawn towards the amount of control and skill a certain action requires, then the third meaning will turn out to be the most salient.

Let's now apply this account to the Knobe Effect. In the *Harm* case, participants expect the agent to be against harming the environment, and thus select the second meaning. But the chairman, because he is indifferent, does not unwillingly harm the environment: thus, his action is judged intentional. In the *Help* case, participants expect the agent to be in favour of helping the environment, and thus select the first meaning. But the chairman does not desire to help the environment; thus, his action is considered unintentional.

This 'linguistic' account explains the Knobe Effect by making moral evaluations a factor determining which meaning of "intentionally" will end up being selected. However, this does not make these meanings "moral" or "evaluative" in any way: they are purely descriptive, and about the agent's capacities and frame of mind. Thus, this account allows moral and evaluative considerations to impact participants' judgments about intentional action while respecting the intuition underlying the 'frame-of-mind condition'. The 'frame-of mind condition' can thus be respected while acknowledging the power of moral and evaluative considerations in shaping our judgments about intentional action.[8]

5. Conclusion

In this paper, my main aim was to put to test a certain family of accounts of the Knobe Effect, namely internalizing accounts, by focusing on its most famous and thoroughly developed representative: Hindriks' *NoRIA*. Arguing that people judge a side-effect as intentional when the agent believes it constitutes a normative reason not to act, but does not care, *NoRIA* was presented as an alternative to evaluative accounts according to which our judgments of intentional action are influenced by our own moral and evaluative judgments. Its main motivation was to respect and conserve a key intuition about intentional action, the 'frame-of-mind condition', according to which judgments about intentional action are about the agent's frame-of-mind, and not about the moral value of his action. Putting *NoRIA* to test through two studies, I have shown that it fails to account for our judgments of intentional action on empirical grounds. However, I have argued that the key intuition behind *NoRIA* could (and should) be conserved, and that there are ways of building accounts of our concept of intentional action that would respect this conceptual insight while being empirically adequate. Thus, even if we should abandon *NoRIA*, and internalizing accounts in general, it is still possible to stay true to their main motivation, and reconcile the empirical data of experimental philosophy with the conceptual results of more traditional methods.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the National Center of Competence in Research (NCCR) in Affective Sciences financed by the Swiss National Science Foundation [n8 51NF40-104897] and hosted by the University of Geneva.

Notes

1. This asymmetry is also called the 'side-effect effect', for Knobe's original experiments only bore on side-effects. Since then, the effect has been shown to hold in case of means towards one's goals. However, since most of the literature on the topic has continued to focus on side-effects, so will this paper. Though it is not easy to give a precise definition of a side-effect, I have argued elsewhere (Cova and Naar 2012a) that one criterion to distinguish means from side-effects could be the following: an event E is a means to my goal G only if E is a necessary component of the causal explanation of my bringing about G. If it isn't a component of the causal explanation, but is still an effect of my reaching G, then it is a side effect.

2. The only known paper that directly tests for *NoRIA* is a paper by Lanteri (2009). However, Lanteri's criticisms to *NoRIA* do not apply to the current version, which takes into account the agent's attitudes (desire or reluctance).

3. The first case is obviously inspired from the famous 'trolley dilemmas'. The inspiration for the second case comes from the *Dog* case in Machery (2008).

4. In *Bystander*, 56.8% of participants gave an answer superior to 4, against 32.6% in *Saviour*.

5. For example, Hindriks (2014) argues that evaluative accounts of the Knobe Effect cannot explain the famous *Nazi Law* pair of cases, while *NoRIA* can. However, in a third study I haven't the space to report in this paper, I have found that *NoRIA* cannot explain the *Nazi Law* cases either: in both cases, very few participants agreed with the claim that the chairman believed that fulfilling/violating the law constituted a reason not to start the new program. Moreover, there was no significant difference between the two cases in participants' agreement ratings.

6. In the *Harm* case, 80.0% of participants gave an answer superior to 4, against 3.9% in the *Help* case.

7. Because these questions have already been debated at length, I admit that I rush through them here. For in-depth discussion of these accounts, see Knobe 2010; Cova 2016.

8. I have argued elsewhere that the second and third solution are compatible (Cova 2016). Indeed, it might be the case both that "intentionally" has different meanings and that each of this meaning has a semantics similar to the semantics of gradable predicates.

Note on contributors

Florian Cova is a postdoctoral researcher at the Swiss Center for Affective Sciences, and principal investigator in the research project "Towards an experimental philosophy of aesthetics", funded by the Cogito Foundation. Outside of aesthetics, his main interests lie in ethics, free will, philosophy of action and philosophy of emotions. He also has a particular interest in the use of empirical methods to address traditional philosophical issues.

References

Adams, F., and A. Steadman. 2004a. “Intentional action in ordinary language: Core concept or pragmatic understanding?” *Analysis* 64: 173-181.

Adams, F., and A. Steadman. 2004b. “Intentional action and moral considerations: still pragmatic.” *Analysis* 64: 268-276.

Alfano, M., J.R. Beebe, and B. Robinson. 2012. “The centrality of belief and reflection in Knobe-Effect cases.” *The Monist* 95: 264-289.

Alicke, M.D. 2008. “Blaming badly.” *Journal of Cognition and Culture*, 8: 179-186.

Cova, F. 2014. “Unconsidered intentional actions: An assessment of Scaife and Webber's “Consideration Hypothesis”.” *Journal of Moral Philosophy* 11: 57-79.

Cova, F. 2016. “The folk concept of intentional action: Empirical approaches.” In *A Companion to Experimental Philosophy*, edited by J. Sytsma and W. Buckwalter, 121-141. Oxford, UK: Oxford University Press.

Cova, F., and H. Naar. 2012a. “Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality.” *Philosophical Psychology* 25: 837-854.

Cova, F., and H. Naar. 2012. “Testing Sripada's deep self model.” *Philosophical Psychology* 25: 647-659.

Cova, F., E. Dupoux, and P. Jacob. 2012. “On doing things intentionally.” *Mind & Language* 27: 378-409.

Cova, F., A. Lantian, and J. Boudesseul. in press. “Can the Knobe Effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality.” *Personality and Social Psychology Bulletin*.

Cushman, F., and A. Mele. 2008. “Intentional action: two and half folk concepts.” In *Experimental Philosophy* edited by J. Knobe and S. Nichols. New York: Oxford University Press.

Egré, P. 2014. “Intentional action and the semantics of gradable expressions (on the Knobe Effect).” In *Causation in grammatical structure*, edited by B. Copley and F. Martin, 176-205. Oxford, UK: Oxford University Press.

Egré, P. & Cova, F. 2015. “Moral asymmetries and the semantics of “many”.” *Semantics & Pragmatics* 8: art. 13.

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: 1635-1647.

Hindriks, F. 2008. “Intentional action and the praise-blame asymmetry.” *The Philosophical Quarterly* 58: 630-641.

Hindriks, F. 2010. “Person as lawyer: How having a guilty mind explains attributions of intentional agency.” *Behavioral and Brain Sciences* 33: 339-340.

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: 190-193.

Knobe, J. 2004. “Folk Psychology and Folk Morality: Response to Critics.” *Analysis* 64: 181-187.

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

Knobe, J. 2010. “Person as scientist, person as moralist.” *Behavioral and Brain Sciences* 33: 315-329.

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: 252-258.

Lanteri, A. 2009. “Judgements of intentionality and moral worth: Experimental challenges to Hindriks.” *The Philosophical Quarterly* 59: 713-720.

Lanteri, A. 2013. “Three-and-a-half folk concepts of intentional action.” *Philosophical Studies, 158,* 17-30.

Machery, E. 2008. “The folk concept of intentional action: Philosophical and experimental issues.” *Mind & Language* 23: 165-189.

Mele, A.R., and F. Cushman. 2007. “Intentional action, folk judgments, and stories: Sorting things out.” *Midwest Studies in Philosophy* 31:184-201.

Nadelhoffer, T. 2004a. “Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality.” *Philosophical Explorations* 9: 203-219.

Nadelhoffer, T. 2004b. “Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow.” *Journal of Theoretical and Philosophical Psychology* 24: 259-269.

Nadelhoffer, T. 2006. “Bad acts, blameworthy agents and intentional actions: some problems for jury impartiality.” *Philosophical Explorations* 9: 203-220.

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

Pettit, D., and J. Knobe. 2009. “The pervasive impact of moral judgment.” *Mind & Language* 24: 586-604.

Phelan, M., and H. Sarkissian. 2009. “Is the ‘Trade-off Hypothesis’ worth trading for?” *Mind & Language*, 24: 164-180.

Sousa, P., and C. Holbrook. 2010. “Folk concepts of intentional action in the contexts of amoral and immoral luck.” *Review of Philosophy and Psychology* 1: 351-370.

Sripada, C. 2010. “The Deep Self Model and asymmetries in folk judgments about intentional action.” *Philosophical Studies* 151: 159-176.

Uttich, K., and T. Lombrozo. 2010. “Norms inform mental state ascriptions: a rational explanation for the side-effect effect.” *Cognition* 116: 87-100.

Wright, J., and J. Bengson. 2009. “Asymmetries in folk judgments of responsibility and intentional action.” *Mind & Language* 24: 237-251.

Young, L., F. Cushman, R. Adolphs, D. Tranel, and M. Hauser. 2006. “Does emotion mediate the effect of an action’s moral status on its intentional status?” *Journal of Cognition and Culture* 1-2: 291-304.

Zalla, T., and M. Leboyer. 2011. “Judgments of intentionality and moral evaluations in individuals with high functioning autism.” *Review of Philosophy and Psychology* 2: 681-698.