Judgments of Intentionality and Moral Worth: Experimental Challenges to Hindriks'

FORTHCOMING: *Philosophical Quarterly*
(doi: 10.1111/j.1467-9213.2009.626.x)

ABSTRACT: Joshua Knobe uncovered two asymmetries in laypersons' judgements about intentional action and moral blameworthiness: people are more likely to describe an action as intentional if it had a bad outcome than a good outcome, and are more likely to blame the bad outcome than to praise the good one. These asymmetries raised numerous questions about lay moral judgment. Frank Hindriks recently proposed a theory of intentional action, according to which one acts intentionally if one fails to comply with a normative reason against performing the action, and he also argued that moral praise requires appropriate motivation, whereas moral blame does not. According to Hindriks, the above asymmetries are normal features of a theory of intentional action and not anomalies. I present two empirical studies that reveal asymmetries in laypersons' judgements of intentionality and moral blameworthiness, which cannot be explained by Hindriks' theory of intentional action.

WORDS COUNT: 3,587 (3,957 including footnotes)

In an investigation of folk conceptions of intentional action, Joshua Knobe conducted an empirical study in which he gave half of the participants the following vignette:[1]

> *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.*

Knobe asked the participants how blameworthy they deemed the chairman and whether the harm had been brought about intentionally. 82% of the respondents said that the chairman caused the harm intentionally and deserved to be harshly blamed (the mean value was 4.8, on a 0 to 6 scale). The other half of the participants received a different vignette in which "help" replaced "harm." The respondents in the help scenario found that the chairman did not intentionally help the environment (only 23% say he did) and that he deserved very little praise (the mean response was 1.4, on a 0 to 6 scale).

These asymmetries in the ascription of intentionality and in the attribution of praise and blame are now known as the 'Knobe effect.' According to Knobe, laypersons attribute intentionality to a decision depending on their evaluation of its outcome. This is an anomaly, in that it defies the commonsense notion that people do (and should) first assess intentions and only afterwards express moral judgments. Frank Hindriks proposed an alternative explanation for the observed asymmetries that does not treat them as anomalous, but as normal features of a (normative) theory of intentional action. According to Hindriks, one can bring about an effect intentionally, although one does not intend to bring it about, if he performs an action that he knows will bring about the undesired effect, while having a normative reason against bringing about such effect.[2] Indeed, Hindriks' theory of intentional action dispels the problems of the Knobe effect. However, I shall present evidence from two experimental studies that reveal

[1] Joshua Knobe, 'Intentional Action and Side Effects in Ordinary Language', *Analysis*, 63 (2003), pp. 190-4.

[2] Frank Hindriks, 'Intentional Action and the Praise-Blame Asymmetry', *The Philosophical Quarterly*, 58 (2008), pp. 630-41.

several asymmetries similar to the Knobe effect, which cannot be entirely explained by Hindriks' theory of intentional action.

I. HINDRIKS' INTENTIONAL ACTION

In seeking a folk account of intentionality that explains the Knobe effect, Hindriks rejects the 'simple view,' according to which an agent brings about an effect intentionally only if he intends to bring it about. For example, I am intentionally aboard a plane if I boarded it and intended to board it. Laypersons also hold more sophisticated views of intentionality, since they say that some effect may have been brought about intentionally even though its actor did not intend to bring it about.

According to Hindriks, therefore, an action is also intentional if one of three theses associated with the 'single phenomenon view' is verified.[3] The first is the 'positive significance thesis' (PST), according to which an agent intentionally brings about some effect, if he intends to bring about another effect that he expects to result in the former effect. Imagine that the main goal in my life is to earn as many air miles as possible and that, at some moment in time, the airline with which I habitually travel launches a special promotion awarding a substantial air miles bonus for flying to New York. It could be said that I arrived in New York intentionally, if I boarded a plane and intended to board it in anticipation of the fact that the plane would have brought me to New York, because I wanted to take advantage of the special offer. Hindriks' second thesis is the 'negative significance thesis' (NST), according to which an agent intentionally brings about some effect if he brings about another effect *in spite* of the fact that he expects it to result in the former effect. Therefore, I arrived in New York intentionally, if I boarded a plane and intended to board it in anticipation of the fact that the plane would have brought me to New York, in spite of the fact that I did not want to be in New York, perhaps

[3] Michael Bratman, *Intentions, Plans, and Practical Reason* (Cambridge: Harvard University Press, 1987). For qualification, see F. Hindriks, 'Intentional Action and the Praise-Blame Asymmetry', p. 638, fn. 11.

because there the cost of living is too high, but simply wanted to earn the bonus air miles. The third thesis is the 'deontic significance thesis' (DST), according to which an agent intentionally brings about some effect if he brings about another effect in spite of the fact that he expects it to result in the former effect and that this would be a normative reason against doing so. For example, I arrived in New York intentionally, if I boarded a plane and intended to board it in anticipation of the fact that the plane would have brought me to New York, in spite of the fact that I believed that my being in New York would be wrong, perhaps because I am the target of a terrorist group and my presence in a large city increases the chances that the city is attacked by the terrorists thereby threatening its inhabitants, which is something I have moral reasons to avoid. A distinction surfaces here between motivating reasons for (not) bringing about an effect, such as a desire to earn air miles, and normative reasons (not) to bring about an effect, such as increasing the chances of a terrorist attack. What Hindriks defines the 'side effect deliberation norm' (SDN) is the condition according to which, when deliberating about what to do, one should take into account morally significant side effects, which can therefore constitute reasons to or not to perform an action. More formally, Hindriks defines an 'intentional action' (IA) as follows:

> An agent *S* φs intentionally if *S* intends to ψ, φs by ψing, expects to φ by ψing, and
>
> (PST) *S* intends to ψ, because he expects thereby to φ, or
>
> (NST) *S* ψs in spite of the fact that he does not want to φ, or
>
> (DST cum SDN) *S* ψs in spite of the fact that he believes his expected φing constitutes a normative reason against ψing.

Hindriks suggests that (IA) constitutes a plausible account of the way laypersons attribute intentionality. In Knobe's study, for example, the failure to satisfy (PST) explains why people say that the chairman does not bring about the desirable effect intentionally. Conversely, (DST) can be satisfied even if the chairman does not care about the side effect in the harm condition,

because of (SDN). I should add that laypersons are likely to resort unconsciously to this or any other theory, and so to express judgements based on its conditions while remaining unaware of doing so.

I shall now present the evidence from two experiments, in which the participants answered questions concerning the intentionality of certain acts, with either positive or negative moral worth. The findings of both experiments are at odds with (IA) and pose a challenge to the capacity of Hindriks' theory of making sense of the ways in which laypersons conceive of intentionality.

II. THE EXPERIMENTS

II.1. *A trolley problem experiment*

The first study was a trolley problem experiment conducted with 62 undergraduates at the Università del Piemonte Orientale, who were randomly assigned to either the standard or the reversed treatment, both of which featured two scenarios – the lever and the stranger – in alternate order.[4] The vignettes of the first scenario read as follows:

> [LEVER] *A trolley without passengers and without conductor is travelling at full speed down a track. On the track there are five people, who will surely be killed if the trolley keeps riding on the actual path. There is also a sidetrack, on which there is one person. A passer-by could pull a lever next to the track, and this way deviate the trolley onto the sidetrack. The passer-by realises that, if he does not pull the lever, the five people will be killed. If he pulls the lever instead, the five people will be saved. The passer-by is aware, however, that by pulling the lever, the person on the sidetrack will be killed.*

The participants then answered several yes/no questions, including the following:

> * *If the passer-by does not pull the lever, is he intentionally killing five people?*
>
> * *If the passer-by pulls the lever, is he intentionally killing one person?*

[4] These results were first published in: Alessandro Lanteri, Chiara Chelini, and Salvatore Rizzello, 'An Experimental Investigation of Emotions and Reasoning in the Trolley Problem', *Journal of Business Ethics*, 83 (2008), pp. 789-804.

The second scenario, and the related questions, were instead the following:

> *[STRANGER] A trolley without passengers nor conductor is travelling at full speed down a track. On the track there are five people, who will surely be killed if the trolley keeps riding on the actual path. A passer-by stands next to the track, and he could push a very fat stranger on the trolley's path, halting its ride. The passer-by realises that, if he does not push the stranger, the five people will be killed. If he pushes the stranger instead, the five people will be saved. The passer-by is aware, however, that by pushing him, the stranger will be killed.*
>
> ** If the passer-by does not push the stranger, is he intentionally killing five people?*
>
> ** If the passer-by pushes the stranger, is he intentionally killing one person?*

Since both scenarios rules out misunderstandings, if the passer-by pulls the lever or pushes the stranger, he intends to do so, and so performs either action intentionally. What about the side effects of these acts?

I do not discuss Hindriks' (IA) with regard to the question about intentionally killing five people, since (IA) does not formally encompass the negative option – i.e., *S* φs intentionally if *S* intends to *not* ψ, φs by *not* ψing, expects to φ by *not* ψing…. Suffice it to point that 86% of the respondents say that not pulling the lever and not pushing the stranger do not amount to intentionally killing of five people, regardless of the treatment.

Let us instead consider how should respondents answer questions about the intentionality of killing one person, according to Hindriks' (IA). In the lever scenario, the passer-by intends to pull the lever, kills one person by pulling the lever, expects to kill one person by pulling the lever and he nevertheless pulls the lever. Hence, though he does not intend to, he intentionally kills one person. Also in the stranger scenario the passer-by intentionally kills one person, because he kills one person by pushing the stranger, expects to kill one person by pushing the stranger and he both pushes the stranger and intends to do so.

In both scenarios, (IA) returns the verdict that the killing of one person was intentional. Yet, the respondents of this experiment did not abide by such judgment. Instead, their answers reveal a distinct asymmetry between the two scenarios. Whereas 90% of the respondents say that

pushing the stranger is an intentional killing, only 42% consider pulling the lever an intentional killing. This asymmetry probably reflects the fact that in the first scenario the passer-by intends to pull the lever and thereby kills one person as a side effect, whereas in the second scenario the passer-by intends to push the stranger and thereby brings about the side effect of saving five lives.[5] According to Hindriks' (IA), however, this should not influence the intentionality of killing either the stranger or the person on the sidetrack.

One way to salvage Hindriks' theory might be to limit its ambition of generality. Perhaps not everybody employs the very same theory and perhaps not everybody regards both situations as identical. As for the stranger, that its killing is intentional seems uncontested, since it is intended and since there are moral reasons not to (intend to) kill a stranger. The lever scenario is less straightforward. The passer-by intends to save five lives and does not intend to kill one person, yet he does, by means of pulling the lever. Depending on whether one sees a normative reason against pulling the lever, this could be either a case of (NST) or (DST cum SDN). The experimental observations of the lever scenario can perhaps be explained by distinguishing the participants who believe there is a moral reason against pulling the lever – and according to whom, this is a case of (DST cum SDN) – from those who do not – and thus regard this as (NST). I'd be inclined to suggest that those who find a moral reason against pulling the lever almost inescapably also see intentionality, whereas those who consider this a case of (NST) may or may not see intentionality. They may reason roughly as follows: the passer-by did not want to kill the person on the sidetrack and, though he in fact killed him, he did not do so intentionally. Others could nevertheless retort that he should have known better and thus find the killing intentional. Such explanation can be tested and, if confirmed, would suggest a possible amendment to Hindriks' theory: to qualify (NST) and consider it as a peculiar case which is sometimes considered intentional and sometimes not. I am not suggesting that (NST) is not a

[5] For a discussion of this difference in connection with the doctrine of double effect, see A. Lanteri et al., 'An Experimental Investigation of Emotions…'.

genuine case of intentionality, but that perhaps not every layperson recognizes it as such in their intuitive folk theories.

The data of this trolley experiment also point to a second asymmetry in the lever scenario. In the standard treatment (i.e. lever scenario first and then stranger), 29% of the respondents say that pulling the lever constitutes an intentional killing of one person, but in the reversed treatment (i.e. stranger scenario first and then lever), the figure scales up to 59%.[6] It is likely that being exposed to the more emotionally intense stranger scenario makes moral violations more salient to the respondents, so that they are now more likely to spot a normative reason against pulling the lever. It is unclear how else this second asymmetry could be accommodated within the (IA), if not by invoking another psychological explanation or admitting to the existence of some bias.

The results of this trolley problem experiment cast the shadow of several anomalies that conflict with Hindriks' attempt to explain away judgment biases. Unless some amendment is successfully introduced – e.g. to qualify (NST) – Hindriks' (IA) proves inadequate to capture the responses elicited in this experiment. Hindriks' (IA) would still be a valuable contribution without being valid for everyone and so he would be on safer grounds by suggesting that in certain situations – e.g. the lever scenario – laypersons regard matters of intentionality and moral worth in a plurality of ways, some of which depart from his theory and some of which are plainly biased. Indeed, even an adjusted (IA) would not easily capture the asymmetries revealed in the next experiment.

II.2. *A modified chairman experiment*

In Knobe's original experiment, the chairman intends to make money. He does not intend to help/harm the environment and he does not even care about it. In the harm scenario, however, he

[6] Both asymmetries are highly statistically significant: χ^2 (1, N=124) = 32.4042, $p < .005$ and χ^2 (1, N=62) = 5.8949, $p < .025$, for the overall responses in the lever and the stranger scenarios and for the responses in the lever scenarios in the two treatments, respectively.

should care because he has a normative reason against harming the environment. Hence, according to Hindriks' (IA), the side effect he brings about to the environment is intentional. On the other hand, the chairman does not have a normative reason against helping the environment, so he does not help it intentionally. Hindriks' theory thus draws a distinction between the two cases along the expectations of the side effects. It is essential that the chairman know in advance whether his decision will have harmful or beneficial consequences on the environment, because only in the former case does he have a normative reason against launching the new product, which makes his bringing about those consequences intentional. Had the chairman not known in advance the new product's consequences on the environment, perhaps he would no longer have had a normative reason not to launch it. Or wouldn't he?

The second study was administered to 52 undergraduates at the Università Cattolica del Sacro Cuore, who were randomly assigned to either the harm or the help treatment of a variant of Knobe's chairman experiment. The vignettes read as follows.

> *An executive of a company goes to the Chief Executive Officer and tells him: 'We are thinking of launching a new product, which will increase profits. This product requires a new technique and we do not know what consequences this will have on the environment.'*
>
> *The CEO replies: 'I do not care at all about the consequences. I just want to make profits. We shall launch the new product immediately.'*
>
> [HARM] *The new product is launched and, as predicted, profits increase. However, the new technique turns out to be polluting and it harms the environment.*
>
> [HELP] *The new product is launched and, as predicted, profits increase. Moreover, the new technique turns out to be ecological and it helps the environment.*

The participants were then asked the following two questions:

> ** According to you, the CEO harmed/helped the environment…*
>
> *a) INTENTIONALLY* *b) NOT INTENTIONALLY* *c) NEITHER*
>
> ** According to you, for his decision, the CEO should be…*
>
> *a) PRAISED* *b) BLAMED* *c) NEITHER*

If bringing about unknown consequences for the environment constitutes a normative reason against launching the product, both treatments of this novel study would be cases of (DST cum SDN) and, accordingly, Hindriks' theory would suggest that the chairman brought about *both* the helpful and the harmful consequences intentionally. If, instead, the unknown consequences on the environment do not count as normative reasons against launching the new product, (IA) would stipulate that the CEO did *not* intentionally help/harm the environment.[7] Neither of these predictions was observed.

In the harm treatment, 58% of the respondents found the CEO's decision intentional, compared to a mere 7% in the help treatment.[8] In the original study the percentage of respondents indicating that the decision was made intentionally was higher than here (remind: 82% in the harm condition and 23% in the help condition). This is because in Knobe's vignettes the chairman knew the side effect in advance. Knowing the side effects in advance explains the difference between Knobe's study and mine above, but it cannot account for the intentionality asymmetry between the two scenarios in Knobe's study, as Hindriks suggests, because the asymmetry remains in this novel study though such knowledge is identical in the two treatments. Therefore, (SDN) alone is insufficient to explain the asymmetries in the original experiment. Even if, as we did in the trolley study above, we allowed for a differential interpretation of the vignettes – i.e. with some participants regarding the case as (NST) and hence alternatively intentional or not – we would still have to explain why the final state of affairs matters when attributing intentionality and blame for two identical decisions.

It remains unclear, moreover, whether the unknown consequences on the environment truly constitute a normative reason against launching the new product. One could take the

[7] This is perhaps best seen as a case concerning luck, a case Hindriks admits to not having yet covered and intends to investigate in the future. F. Hindriks, 'Intentional Action and the Praise-Blame Asymmetry', p. 638, fn. 11.

[8] The responses to the Intentionality question are statistically independent: χ^2 (1, N=52) = 7.4363, $p < .01$. Note that, given the low number of observations in certain responses, the 'Intentional' and 'Neither' answers were grouped together for the Chi-squared test. Hence, there is 1 degree of freedom.

judgments of blameworthiness as proxies that the respondents believe the chairman had moral reasons against disregarding the environment. The data support such view: 79% of the respondents said that the chairman deserved blame in the harm treatment, and also in the help treatment 43% found him blameworthy (while 0% said that he deserved praise).[9,10] Since the decision in either scenario is identical, the praise-blame asymmetry should not occur. Instead, we observe a novel puzzle – what we might deem a 'blame-blame' asymmetry. Explaining away the seeming bias in this asymmetry requires once again finding the differences between the decisions in the two treatments, although they were *not* different.

These figures might also pose another challenge to Hindriks' theory. Here, the respondents who find the CEO blameworthy are more numerous than those who judge the side effect intentional. If Hindriks' (IA) were true, however, we should probably observe more judgments of intentionality than of blame, because it is possible for the CEO to have harmed/helped the environment intentionally without having a moral reason against doing so and so (presumably) without being blameworthy, but not vice versa.

This new chairman experiment questions the capacity of Hindriks' theory to fully capture laypersons' responses to the chairman vignettes. I also doubt that (IA) could be easily adjusted to match this new evidence. Awareness or advance knowledge is indeed a precondition for Hindriks' – and probably anyone's – account of intentionality.[11] An account of intentionality that dispenses with the knowledge of the object about which the intention is held would probably be untenable. The responses elicited in this study should therefore be regarded as anomalies.

---

[9] Also the responses to the Judgement question are statistically independent: χ^2 (1, N=52) = 7.0767, $p < .01$. For the Chi-squared test, the 'Praised' and 'Neither' answers were grouped together. Again, this results in 1 degree of freedom.

[10] Knobe elicited moral judgements that did not allow the respondents in the help treatment to blame the chairman, but only to (not) praise him. A comparison between the two studies along this dimension should thus be taken with caution.

[11] These findings are also problematic for other accounts of folk intentionality that require awareness. Bertram Malle and Joshua Knobe, 'The Folk Concept of Intentionality', *Journal of Experimental Social Psychology, 33* (1997), pp. 101-121.

III. CONCLUDING REMARKS

Hindriks' account deserves ample credit for warning us against the temptation to instate a bias every time an observed behaviour or judgement appears to deviate from the received philosophical view. Laypersons surely hold non-trivial theories of intentionality, which sometimes escape philosophers' capacity to imagine them without gathering relevant empirical evidence, and for which there may exist plausible accounts that explain away the alleged biases. Hindriks' theory of intentionality goes a long way establishing this with respect to the Knobe effect in the context of the original study. When applied to a different experimental setting or to a variant of the original problem, however, (IA) has not proved very robust.

Hindriks' suggestion that folk intuitions concerning intentionality are the (aware or unaware) application of (IA) calls for some amendments in order to address the challenges presented by the evidence I introduced above. Yet, amendments may not suffice because, even when a coherent normative theory is shown capable of accounting for some otherwise anomalous observations, there remain certain aspects of folk theories which cannot be easily, or at all, confined within the borders of a rigorous account of intentionality.[12]

[12] I am grateful for helpful comments from Frank Hindriks and an anonymous referee.