

Forthcoming, *Mind and Language*

**Knobe vs. Machery: Testing the Trade-Off Hypothesis**

Ron Mallon

Department of Philosophy

University of Utah

**Abstract:**

Recent work by Joshua Knobe has established that people are far more likely to describe bad but foreseen side effects as intentionally performed than good but foreseen side effects (this is sometimes called the 'Knobe effect' or the 'side-effect effect.' Edouard Machery has proposed a novel explanation for this asymmetry: it results from construing the bad side effect as a cost that must be incurred to receive a benefit. In this paper, I argue that Machery's 'trade-off hypothesis' is wrong. I do this by reproducing the asymmetry between judgments about good and bad side effects in cases that cannot plausibly be construed as trade-offs.

**Word Count:** 3034

# Knobe vs. Machery Deathmatch: Testing the Trade-Off Hypothesis\*

Ron Mallon

Department of Philosophy

University of Utah

**Word Count:** 3034

## 1. The Knobe Effect and Machery's Trade-Off Hypothesis

Recent work in social psychology has revealed a surprising asymmetry in individuals' judgments (sometimes called the 'side-effect effect' or the 'Knobe effect'). Knobe (2003) found that an unintended but foreseen side effect is regarded as intentional if the outcome is bad, but not if the outcome is good. In his best known experiment, Knobe (2003) presented subjects with one of the following vignettes (differences in bold):

### **Harm Condition**

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**.

---

\* I am grateful to Joshua Knobe, Edouard Machery, Shaun Nichols, and Mark Phelan for their generous and helpful assistance with this paper.  
Ron Mallon, Department of Philosophy, 260 S. Central Campus Drive, Rm. 341,  
University of Utah, Salt Lake City, UT, 84112  
rmallon@philosophy.utah.edu

## **Help Condition**

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**.

In each case, subjects were then asked whether or not the chairman harmed/helped the environment intentionally. But the conditions produced sharply divergent results. Most subjects in the Harm condition (82%) said the chairman harmed the environment intentionally, and most in the Help condition said the chairman did not help the environment intentionally. Knobe (2003) concluded that this asymmetry was the result of our concern to blame individuals who have brought about bad side effects, suggesting the hypothesis that it is the foreseen, morally bad side effect of an action that causes the judgment that the action is intentional. Whatever the explanation, the asymmetry Knobe found has now been replicated many times, including in children (Leslie et al. 2006) and cross culturally (Knobe and Burra 2006), and Knobe's analysis of the asymmetry has also attracted considerable critical attention, including a recent critique by Edouard Machery (forthcoming).

Machery's critique has several components, but here I focus upon his deflationary interpretation of the experimental data. Machery suggests that the asymmetry is brought about by experimental subjects viewing the vignettes with a bad side effect as 'trade-offs'

wherein some cost must be assumed to achieve something good (for example, for the chairman to increase profits). Because costs are thought of as intentionally incurred (ms 21), subjects judge the action in the harm condition as intentional. Because there is no such cost in the Help condition, subjects do not judge the action as intentional. Machery calls his view the *trade-off hypothesis*.

For present purposes, we can isolate two important contrasts between Machery's view and that of Knobe (2003). First, on Machery's view, the asymmetry has 'nothing to do with morality or blame' (ms 22). This suggests that one ought to be able to find the asymmetry in cases without a moral dimension. Second, Machery has a particular positive view about the nature of the processing that is occurring - i.e. that it is the result of a construal of the bad side effects as costs traded for a benefit. This suggests that one ought not to be able to find the asymmetry in a case in which such a construal is absent.

Machery himself is focused on producing evidence for the first claim, and he produces evidence of an asymmetry between bad and good side effects even when no moral blame is at issue.<sup>1</sup> Some of Knobe's own recent work, however, reflects this point. Knobe's own considered view now seems to allow that the asymmetry can occur when a side effect is judged to be bad in some way - not just morally bad (Knobe and Mendlow, 2004; Knobe 2006).<sup>2</sup> This is consistent with the trade-off hypothesis. In any case, my aim here is to consider the second contrast: does the asymmetry (in both Knobe's and Machery's cases) result from a cost-benefit construal of the bad side effects, or does it

---

<sup>1</sup> One concern with Machery's experiment is that the "cost" in his scenario is not really a foreseen "side effect" (as that term is normally used), but rather it is a means to acquiring a benefit. Machery addresses this concern, but not to the satisfaction of some critics (e.g. Phelan and Sarkissian, ms). I ignore this problem here.

<sup>2</sup> Recent work by Knobe (Knobe 2007) makes this picture more complex, suggesting that the evaluation may be unconscious.

result from some more direct move from the badness of the side effect to its intentionality?

If the Knobe-effect asymmetries are the result of trade-offs, they ought not to occur in cases in which subjects do not conceive of the bad side effects as costs (Machery 28).

We might wonder, however, whether there are any such cases. Whenever we have a bad side effect, isn't it open to Machery to claim that subjects conceive the bad side effect as a cost?

Not necessarily. A trade-off involves a cost incurred to receive a benefit. Incurring a cost (i.e. a bad side effect) but not getting a benefit is not, therefore, a trade-off. To be more clear, let us continue to call the peripheral costs and benefits incurred in the course of performing an action *side effects*.<sup>3</sup> And let us say that the main reason an actor undertakes the action is the *primary aim* of the action. Trade-offs are cases with the following structure:

**TRADE-OFF STRUCTURE**

Primary aim: good

Side effect: bad

So, if we could produce the asymmetry between intentionality judgments with regard to good and bad side effects in cases that lacked this structure, it would suggest that contrasting trade-off with non-trade-off cases is not what produces the asymmetry in Knobe's (2003) experiments. And that would be strong evidence against the trade-off hypothesis. There are, however, two problems with carrying out this plan.

First, consider the non-trade-off structures available:

---

<sup>3</sup> But see fn. 1 above.

## NON-TRADE-OFF STRUCTURES

	1	2	3
Primary aim:	bad	bad	good
Side effect:	bad	good	good

To compare good and bad side effects, holding the primary aim constant, involves comparing (1) with (2). But cases with structures (1) and (2) do not seem to make sense since they involve an agent performing an action whose primary aim is bad. Why would anyone perform an act whose primary aim was bad?

There is a second problem in challenging the trade-off hypothesis as well, for there is an ambiguity in just what the claim is. Is the claim that the *fictional agent in the story* views the foreseen side effect as a cost necessary to achieving the primary aim? Or is the claim that the *experimental subject* views the foreseen side effect as a cost necessary to achieve the primary aim? Machery himself is explicitly noncommittal on the question (Machery, fn 10),<sup>4</sup> and so he could claim vindication for the trade-off hypothesis in any case in which a trade-off construal can reasonably be attributed to either the agent in the story or the experimental subject. Any decisive experiment will have to consider both versions of the hypothesis.

As it turns out, we can both create intelligible scenarios with non-trade-off structures *and* test both versions of the trade-off hypothesis with a single experimental design. To

---

<sup>4</sup> Phelan and Sarkissian (ms)'s discussion independently addresses this ambiguity both theoretically and experimentally.

do this we must distinguish clearly between the valences the fictional agent attaches to the primary aim and side effect and the valences that the experimental subject attaches. The obvious answer as to why the agent would do something bad is because the agent perceives it as good. In order for the agent's action structure not to be interpretable as a trade-off for the agent, the agent's assignment of valences would (in all vignettes) have to be like (3) above: good primary aim, and good side effect. So in order to contrast good and bad side effects, we must vary these for the experimental subject. The experimental subjects' assignments of valences will have to be like (1) and (2) above: bad primary aims, and good or bad side effects. If we find the Knobe effect in vignette pairs where the subjects judge the pairs to have structures (1) and (2), but the fictional agent judges both vignettes to have structure (3), then that will count as serious evidence against the trade-off hypothesis, for such cases would be trade-offs for neither the experimental

## **2. Testing for Trade-offs**

### **2.1. Experiment 1**

This experiment looked for an asymmetry in intentionality judgments regarding good and bad side effects in cases that lacked trade-off structure for both experimental subjects and fictional agents.

#### **2.1.1. Method**

##### **2.1.1.1. Participants**

Fifty-two undergraduate subjects at the University of Utah participated. Fourteen were female, thirty-eight male.

#### **2.1.1.2. Materials**

The following two scenarios, modeled closely on Knobe (2003)'s scenarios (and thematically based on a case from Knobe and Kelly 2006) were used, varying the valence experimental subjects would attach to the primary aim/side effect (bad/good or bad/bad) but not the valence attached by the fictional agent (good/good in both cases). Scenarios were designed so that typical subjects would think the valence of the primary aims and side effects obvious, but also so that the (sometimes differing) valences attached by the fictional agent would also be obvious.

#### **Harmful Terrorist**

A member of a terrorist cell went to the leader and said, "We are thinking of bombing a nightclub. It will kill many Americans, but it will also harm the Australians since many Australians will be killed too."

The leader answered, "I admit that it would be good to harm the Australians, but I don't really care about that. I just want to kill as many Americans as possible! Let's bomb the nightclub!"

They did bomb the nightclub, and sure enough, the Australians were harmed since many Australians were killed.

And then subjects were asked:

Did the terrorist leader intentionally harm the Australians?

### **Helpful Terrorist**

A member of a terrorist cell went to the leader and said, "We are thinking of bombing a nightclub. It will kill many Americans, but it will also drive down property costs, helping the nearby orphanage acquire the land it needs for the children."

The leader answered, "I admit that it would be good to help the orphanage, but I don't really care about that. I just want to kill as many Americans as possible! Let's bomb the nightclub!"

They did bomb the nightclub, and sure enough, the orphanage was helped by falling property values.

And then subjects were asked:

Did the terrorist leader intentionally help the orphanage?

Both cases probes were presented as open-ended questions (rather than with a fixed choice between, e.g. 'YES' and 'NO').

#### **2.1.1.3. Procedure**

Subjects were given one of the two scenarios in a classroom setting.

#### **2.1.2. Results**

In every case, subjects provided a clearly affirmative or negative answer to the probe.

Subjects in the Harmful Terrorist condition were overwhelmingly more likely to say that

the harmful side effect was intentional (92%) than subjects in the Helpful Terrorist condition were to say that the helpful side effect was intentional (12%). [Insert Table 1 about here.] This difference was highly significant ( $\chi^2(1, N=52) = 33.973, p < .0001$ , two-tailed).

Moreover, some subjects supplemented their answers with explanations that, in the Harmful Terrorist case, drew attention to the fact that the harm was foreseen as evidence of its being intentional ('he knowingly bombed the nightclub knowing Australians would be there'; 'terrorists knew the bombing would harm the Australians'; 'the terrorists knew that there would be Australian casualties before carrying out the order'; 'he had knowledge of the Australians being at the nightclub'), and that, in the Helpful Terrorist case, drew attention to the fact that the terrorist said he didn't care about helping as evidence of its being unintended ('he said he did not care about the orphanage').

### **2.1.3. Discussion**

Finding the asymmetry in cases that cannot plausibly be construed as trade-offs provides evidence against Machery's trade-off hypothesis. Because the effect in this case could have been an artifact of the particular materials used (e.g. perhaps American undergraduates just really hate terrorists) a second experiment with the same structure was performed.

## **2.2. Experiment 2**

### **2.2.1. Method**

### **2.2.1.1. Participants**

Fifty-one undergraduate subjects at the University of Utah took part. Fifteen females and thirty-six males participated.

### **2.2.1.2. Materials**

Two new scenarios were used, again varying the valence experimental subjects would attach to the primary aim/side effect (bad/good or bad/bad) but not the valence attached by the fictional agent (good/good in both cases).

#### **The Harmful Gang Leader**

A member of a local gang went to the leader and said, "We are thinking of trying a new tactic. It will flood the neighborhood with cheaper cocaine, increasing our profits, but it will also harm the cops since more cops will die in drug-related violence."

The leader answered, "I admit that it would be good to harm the cops, but I don't really care about that. I just want to make as much profit as I can. Let's implement the new tactic."

They did implement the new tactic, and sure enough, the cops were harmed since more cops died in drug-related violence.

And then subjects were asked:

Did the leader of the gang intentionally harm the cops?

#### **Helpful Gang Leader**

A member of a local gang went to the leader and said, "We are thinking of trying a

new tactic. It will flood the neighborhood with cheaper cocaine, increasing our profits, but it will also help hard-core addicts to have more money for food and housing."

The leader answered, "I admit that it would be good for people to have more money for food and housing, but I don't really care about that. I just want to make as much profit as I can. Let's implement the new tactic."

They did implement the new tactic, and sure enough, hard-core addicts were helped by having more money.

And then subjects were asked:

Did the leader of the gang intentionally help hard-core addicts?

Again, in both scenarios, probes were presented as open-ended questions.

### **2.2.1.3. Procedure**

Subjects were given one of the two scenarios in a classroom setting.

### **2.2.2. Results**

In every case, subjects provided a clearly affirmative or negative answer to the probe.

Again, most subjects judged a bad side effect to be intentional (62%) while

comparatively few judged a good side effect to be intentional (28%) ( $\chi^2(1, N=51) =$

5.79,  $p < .05$ , two-tailed). [Insert Table 2 about here] Subjects again took the opportunity

to draw attention to fact that the harms were foreseen ('he agreed to implement the new

tactic knowing that it would harm the cops. Having that knowledge makes him

responsible for the outcome'; 'because he knew the consequence of his decision would

hurt the cops'; 'the leader new [sic] the consequences of his actions'; 'he knew about it...') as evidence of the Harmful Gang Leader's intention. And they again drew attention to the fact that the fictional agent (the Helpful Gang Leader) did not care about helping as evidence that the help was not intended ('he only cared about his profits'; 'helping hardcore addicts is just a side effect of this new tactic').

### **2.2.3. Discussion**

Because the cases are not trade-offs, but reproduce the Knobe asymmetry, the second experiment again undercuts Machery's trade-off explanation.

## **3. General Discussion**

Each case provides strong evidence that an asymmetry in the attribution of intentionality in the production of good and bad side effects exists even in scenarios that cannot plausibly be construed as trade-offs. Moreover, the long responses provided by some applicants suggest that the asymmetric answers might be the result of the differential recruitment of available considerations in light of affective arousal, a kind of motivated cognition.<sup>5</sup> On the other hand, the very cases from Machery (forthcoming) and Knobe and Mendlow (2004) that seem to show that the Knobe effect can arise in cases where there is no moral judgment count against the claim that the effect is the result of motivated cognition for these are cases in which a strong affective component seems lacking.

---

<sup>5</sup> See Kunda (1999, Chapter 6) for a review of evidence of motivated cognition.

In any case, our ability to produce the Knobe effect in cases where no trade-off exists suggests that the trade-off hypothesis gives the wrong explanation of this effect, and it is consistent with the view that it is simply the badness of the side effect drives differences in the attribution of intentionality. This alternative view faces its own experimental challenges, including recent work by Thomas Nadelhoffer (2004a, 2004b), Shaun Nichols and Joe Ulatowski (2007), and Walter Sinnott-Armstrong et al. (forthcoming). The best we can say for now is that while the trade-off hypothesis looks to be the wrong explanation for the Knobe effect, consensus on the correct explanation remains elusive.

*Department of Philosophy*

*University of Utah*

## **Bibliography**

- Knobe, J. (2003). 'Intentional action and side-effects in ordinary language.' *Analysis* **63**: 190-193.
- Knobe, J. (2006). 'The concept of intentional action: a case study in the uses of folk psychology.' *Philosophical Studies* **130**: 203-231.
- Knobe, J. (2007). 'Reason explanation in folk psychology.' *Midwest Studies in Philosophy* **31**(1): 90-106.
- Knobe, J. and A. Burra (2006). 'The folk concepts of intention and intentional action: a cross-cultural study.' *Journal of Culture and Cognition* **6**: 113-132.
- 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.
- Knobe, J. and S. Kelly (2006). Can one act for a reason without acting intentionally? Unpublished Manuscript.
- Kunda, Z. (1999). *Social Cognition: Making Sense of People*. Cambridge, MA, MIT Press.
- Leslie, A., J. Knobe, et al. (2006). 'Acting Intentionally and the side-effect effect: 'theory of mind' and moral judgment.' *Psychological Science* **17**: 421-427.
- Machery, E. (forthcoming). 'The folk concept of intentional action: philosophical and experimental Issues.' *Mind and Language*.
- Nadelhoffer, T. (2004a). 'Praise, side effects, and intentional action.' *The Journal of Theoretical and Philosophical Psychology* **24**: 196-213.

- Nadelhoffer, T. (2004b). 'Blame, badness, and intentional action: a reply to Knobe and Mendlow.' *The Journal of Theoretical and Philosophical Psychology* **24**: 259-269.
- Nichols, S. and J. Ulatowski (2007). 'Intuitions and individual differences: The Knobe effect revisited.' *Mind and Language* 22(4): 346-365.
- Phelan, M. and H. Sarkissian (ms). 'Is the "trade-off hypothesis" worth trading for?'
- Sinnott-Armstrong, W., R. Mallon, et al. (forthcoming). 'Intention, temporal order, and moral judgments.' *Mind and Language*.

Figure 1.

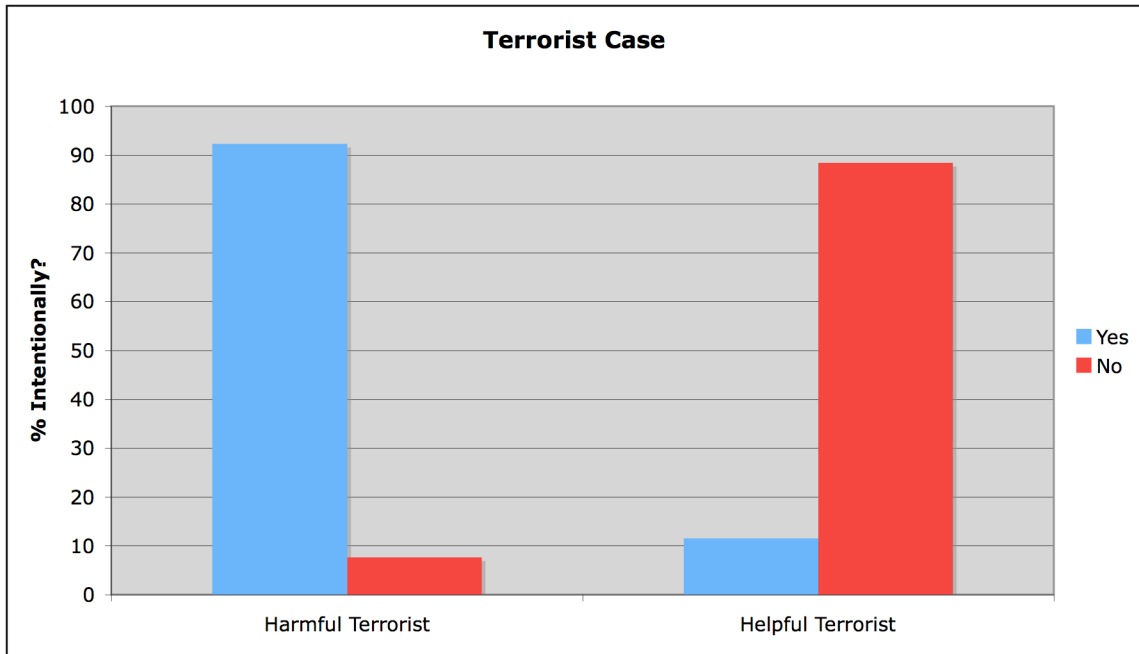


Figure 2.

