Is the 'Trade-off Hypothesis' Worth Trading For?

MARK PHELAN AND HAGOP SARKISSIAN

Abstract: Recently, the experimental philosopher Joshua Knobe has shown that the folk are more inclined to describe side effects as intentional actions when they bring about bad results. Edouard Machery has offered an intriguing new explanation of Knobe's work—the 'trade-off hypothesis'—which denies that moral considerations explain folk applications of the concept of intentional action. We critique Machery's hypothesis and offer empirical evidence against it. We also evaluate the current state of the debate concerning the concept of intentionality, and argue that, given the number of variables at play, any parsimonious account of the relevant data is implausible.

Edouard Machery's paper, 'The Folk Concept of Intentional Action: Philosophical and Experimental Issues' (Machery, 2008) puts forth an intriguing new hypothesis concerning recent, empirically informed work on the concept of intentional action. In this paper, we critique Machery's 'trade-off' hypothesis, offer empirical evidence to reject it, and evaluate the current state of the debate concerning the concept of intentional action.

1. Introduction of the Issue

Before we discuss Machery's hypothesis, we will briefly recap the recent empirical work that is the focus of Machery's analysis. This work, which is most closely associated with Joshua Knobe (Knobe, 2003, 2006; Knobe and Mendlow, 2004), will be familiar to anyone following recent discussions concerning the concept of intentional action. Knobe (2003) reported the results of survey questions asked about paired, minimally divergent vignettes. One pair involved a company chairman who approved a new program which would generate profits and also either a) harm or b) help the environment (side effects which the chairman cared nothing about). In this case, the vast majority of subjects agreed that the chairman intentionally caused the bad side effect of harming the environment and not the good one of helping the environment. Traditional descriptive accounts of the concept of intentional action, focusing on non-evaluative features of the concept, are ill-equipped to explain this result.

Both authors contributed equally to this paper. Our thanks to Edouard Machery and Thomas Nadelhoffer for stimulating discussion of a previous draft of this paper, and to Edouard Machery and Bertram F. Malle for helpful comments in their referee reports for *Mind & Language*. Finally, special thanks to Joshua Knobe for comments and suggestions on numerous previous drafts, and to Fiery Cushman for suggestions on analysis of data.

Address for correspondence: Hagop Sarkissian, Department of Philosophy, The City University of New York, Baruch College, Box B5/295, New York NY 10010, USA.
Email: hagop.sarkissian@baruch.cuny.edu

Mind & Language, Vol. 24 No. 2 April 2009, pp. 164–180.

Faced with a gap in traditional accounts of the concept of intentional action, a number of explanations have been offered for the asymmetry in these cases. Knobe himself, for example, has claimed that the concept of intentional action is sensitive to evaluative considerations, particularly moral considerations. If a side effect is *bad* (morally or otherwise) and foreseen, Knobe suggests that a majority of subjects will judge it to have been brought about intentionally.[1] If, on the other hand, a side effect is *good* and foreseen, subjects will judge that it was not brought about intentionally (Knobe, 2006). Moreover, Knobe thinks that judgments of intentionality that are sensitive to such evaluative considerations are sometimes competent judgments; they are not performance errors, and instead exhibit appropriate use of the concept. Others, such as Nadelhoffer (2004a) and Malle and Nelson (2003; cf. Malle, 2004), contend that the intentionality judgments made in the harming cases are causally subject to evaluative considerations. However, they further believe that, in being so subject, these judgments are biased. These theorists claim that subjects, in a rush to blame the agents for bringing about the bad side effect, inappropriately conclude that these agents acted intentionally.

Machery takes a third line. First, he claims that general considerations about the 'concept' debate in philosophy of mind should lead us to be agnostic as to whether or not the intentionality judgments made in the harming cases are the result of apt processes or biases. Second, he claims that the judgments made in these cases (whether competent or biased) are in fact not sensitive to moral considerations, but rather to non-moral, cost/benefit analyses. In this paper, we aim to critique, and ultimately reject, this latter point. Moreover, we will offer some more general comments about the sort of explanation we might expect for the Knobe effect.

2. Machery's Trade-off Hypothesis

Machery offers a new account of the features to which subjects are sensitive when they judge, for instance, that the chairman harmed the environment intentionally. Machery's hypothesis purportedly explains the asymmetries outlined above without appealing to moral considerations. Machery's central claim is that subjects conceive costs that are knowingly incurred as costs that are intentionally incurred. Since the chairman and other cases depict individuals knowingly incurring certain costs in the pursuit of a goal, subjects take these costs to be incurred intentionally. We follow Machery in calling this the 'trade-off' hypothesis.

Machery's hypothesis represents an intriguingly different sort of explanation than those previously mentioned. For, in spite of their differences, the explanations canvassed above all take moral considerations to be implicated in the asymmetries. According to Machery his explanation denies any relationship between the Knobe effect and moral evaluations, which may be more salutary to existing accounts of

[1] In the face of mounting counterevidence, Knobe (2007) has recently repudiated this explanation.

folk psychology. Insofar as it denies any relationship between the Knobe effect and moral psychology, Machery characterizes his explanation of the Knobe effect as a 'deflationary' one. Furthermore, Machery alleges that his account explains certain asymmetries in intentionality judgments that the moral/evaluative accounts fail to explain. Machery's argument runs as follows: If the trade-off hypothesis explains, and other accounts fail to explain, Knobe and other asymmetries, then it is more generally true than the other accounts, and so it is more parsimonious to embrace it instead of the other accounts.

Our discussion proceeds as follows. First, we discuss an ambiguity in Machery's thesis and argue for its significance. We then challenge the empirical evidence Machery offers in support of his thesis.[2] Next, we offer our own evidence to show that the thesis is, in a strict sense, false. Finally, we motivate some skepticism toward any account that purports to explain the Knobe effect by recourse to one or two factors alone.

2.1 An Ambiguity in the Trade-off Hypothesis

There is an ambiguity in Machery's presentation of the trade-off hypothesis. At times he claims that subjects view the *agents* within the vignettes as incurring costs *to themselves*. For example, consider how he explains the original harming chairman case:

> The chairman desires to obtain something she judges to be beneficial—an increase in profits for her company. She foresees that obtaining this benefit will entail some cost—harming the environment. But because the foreseen cost is offset by the foreseen benefit, the chairman decides to incur the foreseen cost—harming the environment—in order to reap the foreseen benefit—increasing the profits of the company (p. 176).

In other words, subjects read the vignette as involving a chairman who foresees the benefit of profits, conceptualizes the side effect of harming the environment as a cost *to herself*, and decides to proceed based on a cost-benefit analysis. According to this interpretation, the vignette involves a true *trade-off*: an agent agrees to a trade-off because the benefits she gains outweigh the costs she incurs.

At other times, however, Machery claims that it is the *subjects* reading the vignettes (rather than the agents within the vignettes) who consider the side effect a cost, from their own point of view:

> When people read the harm case, they conceptualize the side-effect *harming the environment* as a cost, that is, as something that is negatively valued and that one must incur if one is to reap a greater benefit. They think of this cost as being offset by the benefit *increasing the profits of the company*. That is, they conceptualize the harm case as involving a trade-off between a cost and a benefit (pp. 176–177).

---

[2] Ron Mallon's discussion (2008) independently addresses some of these issues both theoretically and experimentally.

Here the focus is on how the subjects themselves conceive the side effect, not how they conceive the chairman's relationship to the side effect. It seems consistent with this statement that harming the environment need not be conceived by subjects as a cost to the agent within the vignette. It's enough that it constitutes a cost in some more generic sense. Call these two specifications of the trade-off hypothesis the *agent trade-off hypothesis* and the *subject trade-off hypothesis*.

Machery himself is aware of the ambiguity of his thesis, but he remains non-committal about it. In a footnote he writes:

> One might ask whether (i) the subjects in the experiment have to think of the side-effect as being a cost or whether (ii) the subjects have to judge that the agent described in the probe (e.g., the chairman) thinks of the side-effect as a cost. The two cases are not equivalent because the subjects might think of a side-effect as a cost, while the agent might be described as desiring this side-effect. Conversely, the agent might be described as thinking of the side-effect as a cost, while the subjects might think otherwise. I remain noncommittal with respect to (i) and (ii). (p. 177, n.10)

But can he plausibly remain noncommittal? Consider the harming chairman case, where the chairman approves a program that will increase profits but will also harm the environment (Knobe, 2003). Now, according to the agent trade-off hypothesis, subjects read the vignette and interpret the chairman as conceiving 'harming the environment' as a cost she herself is willing to incur for the benefit of making profits. In other words, the chairman views the harm as a cost to herself. However, the chairman's response—'*I don't care at all about the environment. I just want to make as much profit as I can*'—seems to suggest otherwise. Her dismissive attitude suggests that she does not consider it a cost to herself. She neither expresses regret nor pauses to re-evaluate her actions. Hence this case does not involve a clear trade-off on the agent trade-off formulation. But perhaps Machery would maintain that it constitutes a cost to the agent in a technical sense. For while the vignette makes clear that the chairman does not care a great deal about the environment, she is human, after all, so presumably she cares something about it. One goal of our experimental section will be to present a clear counter-example to the agent-centered version of the trade-off hypothesis.

In any case, if harming the environment is not conceived as a cost *to the chairman*, it might still be conceived as a cost in some generic sense. If Machery wants his trade-off hypothesis to explain all of the vignettes he purports it explains, perhaps he should reject the agent trade-off hypothesis in favor of the subject trade-off hypothesis. According to this formulation, subjects recognize something *they* consider a cost incurred for something *they* consider a benefit within the relevant vignettes and therefore apprehend a trade-off. This formulation of the hypothesis does seem to fit the relevant cases thus far discussed.

However, if the trade-off hypothesis is seen as involving things subjects regard as costs, it begins to look very similar to Knobe's hypothesis discussed above. After all, if

people conceive something as a cost, they must conceive it as a bad thing, and Knobe's explanation of the asymmetries was that foreseen bad side effects would be judged intentional. On this reading, it seems that both Knobe's and Machery's accounts focus on subjects' apprehension of the badness of side effects, the chief difference between them being that the relation between side effects and main goal is important on Machery's account but not Knobe's.[3] The distinction between the agent and subject specifications of the trade-off hypothesis will come up again in the following two sections. We now turn to Machery's evidence for the trade-off hypothesis.

2.2 Machery's Evidence for the Trade-off Hypothesis

In this section we critique new findings which Machery offers to support the 'trade-off' hypothesis. Consider the following case.

> *The extra-dollar case*
>
> Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest sized drink available. Before ordering, the cashier told him that the Mega-Sized Smoothies were now one dollar more than they used to be. Joe replied, 'I don't care if I have to pay one dollar more, I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie and paid one dollar more for it.

Machery interprets this case as an instance of the Knobe effect. When asked about this case, nearly all subjects thought Joe paid the extra dollar intentionally. According to Machery, this is because subjects interpret paying an extra dollar as a trade-off: a cost knowingly incurred in the pursuit of a goal. Machery takes this as supporting his claim that the Knobe effect can be explained without recourse to moral considerations.

Let us first make a quick but important point about this case drawing on the discussion of the previous section. If subjects are thinking of paying the extra dollar as a cost, they are thinking of paying the extra dollar as a generally bad thing. So this case seems not to disentangle explanations of the Knobe effect from evaluative considerations generally. Though evaluative considerations are a broader class than moral ones, Knobe claims the concept of intentional action is sensitive to the broader class as well (Knobe, 2006; Knobe and Mendlow, 2004), as Machery himself is aware (p. 168, n.3). So the case does not count

[3] It is important to note that Knobe's views on what explains the relevant effect are not consistent. Knobe (2003) defends the view that negative *moral* evaluations explain the judgments of intentionality in, for instance, the harming chairman case. Knobe (2006) and Knobe and Mendlow (2004), on the other hand, suggest that negative evaluations in general do the work, whether moral or not. Knobe's original view contrasts sharply with Machery's; Knobe's later, less radical account does not.

in favor of the 'trade-off' hypothesis as opposed to (one version of) Knobe's hypothesis.[4]

Be that as it may, the extra-dollar case also faces a more serious objection, which threatens its relevance to the topic of intentional side effects. The objection claims that the case does not involve a side effect at all. Remember that the Knobe effect concerns cases involving two effects, a *main goal* which the agent is explicitly trying to bring about, and a *side effect* which she is not trying to bring about, and which is not a means to the main goal she is trying to bring about. What made Knobe's original findings so puzzling is that *side effects* (effects that were neither deliberately pursued nor were means to an end pursued) were deemed intentional. By contrast, there is nothing puzzling about the intentionality judgments in Machery's extra-dollar case: paying for something is a *means* to getting it, not a side effect incurred in getting it.[5]

Machery replies to this objection, but we find his reply unconvincing. He writes:

> The striking phenomenon is that people make similar judgments when the case involves a negatively valued side-effect such as harming the environment and when the case involves a negatively valued means such as paying an extra-dollar. In both cases, people tend to judge that a foreseen by-product of a goal (paying an extra-dollar and harming the environment) has been intentionally brought about. The most plausible explanation of why people have similar intuitions in the harm case and in the extra-dollar case is that when people read these two cases, they conceptualize both paying an extra-dollar and harming the environment as being a cost that the agent incurs in order to get a desired benefit (p. 183).

However, even if Machery's trade-off hypothesis served to explain both negatively valued means *and* negatively valued side effects, we would still require some other explanation for positively valued means (which are not plausibly costs). Presumably, the intentionality of means admits of the same explanation, whether those means are positive or negative. So it is not more plausible (and certainly not more parsimonious) to assume that the intentionality of negatively valued side effects and negatively valued means admit of a single explanation.

What Machery needs to support his hypothesis are cases in which the relevant effect cannot plausibly be construed as a means to a desired end. In fact, he goes on to provide two such cases. The vignettes are based on standard trolley problems. If successful, these would constitute positive evidence for the trade-off hypothesis, as they involve side effects that are not means. Moreover, Machery takes these as cases Knobe's hypothesis cannot accommodate.

---

[4] But see footnote 2.

[5] Machery contrasts this with a case in which an agent's smoothie will come in a commemorative cup, where few people judged the agent to have received the free cup intentionally. But, on the present objection, this is unsurprising. Receiving a smoothie in a free cup is not a means to receiving a smoothie; it is a side effect of receiving a smoothie.

The worker case

John is standing near the tracks of a trolley. John notices that the brakes of the trolley have failed. Five workmen are working on the tracks with their backs turned. John sees that the runaway trolley is headed for the five workmen who will be killed if it proceeds on its present course. The only way to save these five workmen is to hit a switch that will turn the trolley onto the side tracks. Unfortunately, there is a single workman on the side tracks with his back turned. John knows that workman on the side tracks will be killed if he hits the switch, but the five workmen will be saved. John decides to hit the switch. Sure enough, the trolley turns on the side tracks, the five workmen on the main tracks are saved, and the workman on the side tracks is killed.

The dog case

John is standing near the tracks of a trolley. John notices that the brakes of the trolley have failed. Five workmen are working on the tracks with their backs turned. John sees that the runaway trolley is headed for the five workmen who will be killed if it proceeds on its present course. The only way to save these five workmen is to hit a switch that will turn the trolley onto the side tracks. Moreover, there is a dog on the tracks with its back turned. John knows that the five workmen and the dog will be saved if he hits the switch. John thinks 'I don't care at all about saving the dog. I just want to save the five workmen.' John decides to hit the switch. Sure enough, the trolley turns on the side tracks, the five workmen and the dog on the main tracks are saved.

Subjects were then given one of four probes. The first pair of probes concerned whether causing the death of the worker or saving the dog—the relevant side effects in these cases—was appropriate. In each of these cases, the vast majority of subjects deemed the side effect to be appropriate. On this basis, Machery concludes that, according to Knobe's hypothesis, subjects would judge neither effect intentional.

But it remains unclear why such an inference is warranted. Knobe's hypothesis concerns the *goodness* or *badness* of side effects, not their *appropriateness* or *inappropriateness*. After all, otherwise bad actions (e.g. causing the death of someone) can be deemed appropriate if they are offset by other considerations (saving the lives of five others); similarly, otherwise good actions (e.g. consoling a child) can be inappropriate when offset by other considerations (when done by a pedophile).[6] What Machery should have asked was whether the side effect was *bad* or not, as other studies he cites have done (e.g. Phelan and Sarkissian, 2008; Wright and Bengson, 2008). It seems obvious to us that, though subjects judged both side

[6] The latter example is from Knobe and Mendlow, 2004.

effects appropriate, only causing the death of the worker would be judged bad. (After all, how can saving a dog be a bad thing?) As it stands, Machery cannot use the answers to the appropriateness question to predict anything whatsoever with regards to Knobe's hypothesis.

If we are right in our assertion that people would judge causing someone's death as bad and saving a dog as good, Knobe's hypothesis actually predicts the results to the intentional question correctly. Causing someone's death is bad and the agent knew about it. Saving a dog is good, but the agent didn't care about it. Thus, Knobe's theory would predict a higher level of intentionality judgments in the first case than in the latter. This pattern of judgments was born out, with 56% of subjects judging the former side effect of killing the worker to have been brought about intentionally, and only 23% judging the latter side effect of saving the dog to have been brought about intentionally Knobe's hypothesis accommodates the data as well as Machery's.

To sum up our critique thus far: There is an ambiguity in Machery's presentation of the trade-off hypothesis. On one presentation, the 'subject' formulation, plausible evidence for the hypothesis is likely to constitute evidence for Knobe's hypothesis as well, since both turn on a subject's apprehension of bad effects. On another presentation, the 'agent' formulation, Machery does provide a significantly different theory, but it is questionable whether this theory can accommodate some core instances of the Knobe effect in the existing literature. Regarding Machery's own evidence for his view, the case involving paying an extra-dollar for a smoothie is readily amenable to means-end explanation, and the cases concerning the dog and the worker assimilate to Knobe's hypothesis. As it stands, then, there is no clear corroboration of the trade-off hypothesis that doesn't also corroborate Knobe's hypothesis.

In the following section we will offer direct counter-evidence to the trade-off hypothesis. Note that the Knobe effect has often been tested by using two similar cases, often with only one difference between them—the goodness or badness of the side effect. In Machery's trolley cases, however, there was an additional difference between the two cases. In the second case, information about the agent's *attitude* toward saving the dog is inserted. 'John thinks "I don't care at all about saving the dog. I just want to save the five workmen."' This information about the agent's attitude is lacking in the first case. How does the agent feel about the death of the lone worker? Is he remorseful, or is he indifferent? In the following sections, we will see how changing the agent's attitude and the importance of the goals pursued affect how subjects assess intentionality.

2.3 Evidence Against the Trade-Off Hypothesis

Having discussed Machery's evidence for the trade-off hypothesis, we now turn to our own evidence against that hypothesis. Before we discuss our evidence, let's remind ourselves just how reasoning to intentional side effects purportedly proceeds according to the trade-off hypothesis. We adapt this diagram from Machery (2008):

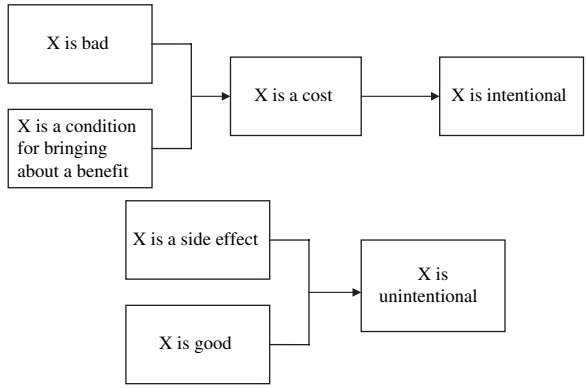

Figure 1 *People's reasoning in relevant cases according to the trade-off hypothesis.*

According to the trade-off hypothesis model, if an effect is both bad and a condition for bringing about a benefit then it is a cost, and because costs are intentionally incurred subjects will judge the effect to have been brought about intentionally. Moreover, we can specify the trade-off hypothesis in each of two ways: the trade-off might involve what the *subject* would consider a cost, or it might involve what the subject believes the *agent* in the vignette would consider a cost.

The clarity of Machery's thesis allows us to specify certain predictions the theory would make. The clearest instance of a trade-off on the agent-centered version would be one in which the agent in the vignette clearly considers the side effect a bad thing, but nonetheless something worth enduring for the benefit being pursued. If the agent trade-off hypothesis is correct, we should expect higher judgments of intentionality in a case in which the agent clearly considers the side effect a bad thing, and lower judgments of intentionality in a nearly identical case in which the agent clearly does not consider the side effect a bad thing. The best instance of a trade-off on the subject-centered version would be one in which the side effect is obviously bad from the subjects' point of view, but nonetheless worth enduring for the main benefit being pursued. If the subject-centered view is correct, we should expect higher judgments of intentionality in a case involving a bad side effect endured for an important goal, and lower judgments of intentionality in a nearly identical case in which the main goal is something few people would consider a benefit.

In the remainder of this section, we show, by means of a single experiment, that neither of these predictions pans out. This suggests that neither the agent-centered nor the subject-centered version of the trade-off hypothesis is generally correct. We asked 294 UNC/Chapel Hill undergraduates about the following four, minimally divergent cases, all of them modifications of an original case involving a lieutenant from Knobe, 2003. The differences between the cases have been highlighted in italics for ease of reading:

Case 1: Caring Lieutenant, Important Goal

A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.' The sergeant said: 'But if I send my squad to the top of Thompson Hill, we'll be moving the men directly into the enemy's line of fire. Some of them will surely be killed!'

The lieutenant answered: 'Look, I know that they'll be in the line of fire, and I know that some of them will be killed. *I care about my soldiers more than anyone else. But it's imperative to the success of this campaign that we take Thompson Hill.*'

The squad was sent to the top of Thompson Hill. As expected, the soldiers were moved into the enemy's line of fire, and some of them were killed.

Case 2: Caring Lieutenant, Unimportant Goal

A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.' The sergeant said: 'But if I send my squad to the top of Thompson Hill, we'll be moving the men directly into the enemy's line of fire. Some of them will surely be killed! *And with the enemy's heavy artillery arriving tomorrow, we won't be able to keep the hill.*'

The lieutenant answered: 'Look, I know that they'll be in the line of fire, and I know that some of them will be killed. *I care about my soldiers more than anyone else. I know we won't be able to keep Thompson Hill, but I just want to take it.*'

The squad was sent to the top of Thompson Hill. As expected, the soldiers were moved into the enemy's line of fire, and some of them were killed. *Thompson Hill was lost shortly thereafter.*

Case 3: Uncaring Lieutenant, Important Goal

A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.' The sergeant said: 'But if I send my squad to the top of Thompson Hill, we'll be moving the men directly into the enemy's line of fire. Some of them will surely be killed!'

The lieutenant answered: 'Look, I know that they'll be in the line of fire, and I know that some of them will be killed. *But I don't care at all about what happens to our soldiers. It's imperative to the success of this campaign that we take Thompson Hill.*'

The squad was sent to the top of Thompson Hill. As expected, the soldiers were moved into the enemy's line of fire, and some of them were killed.

Case 4: Uncaring Lieutenant, Unimportant Goal

A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.' The sergeant said: 'But if I send my squad to the top of Thompson Hill, we'll be moving the men directly into the enemy's line of fire. Some of them will surely be killed! *And with the enemy's heavy artillery arriving tomorrow, we won't be able to keep the hill.*'

The lieutenant answered: 'Look, I know that they'll be in the line of fire, and I know that some of them will be killed. *But I don't care at all about what happens to our soldiers. I know we won't be able to keep Thompson Hill, but I just want to take it.*'

The squad was sent to the top of Thompson Hill. As expected, the soldiers were moved into the enemy's line of fire, and some of them were killed. *Thompson Hill was lost shortly there after.*

Each subject was randomly assigned to one of the four conditions. After reading the vignette, subjects were asked to answer the following question: 'Did the lieutenant intentionally cause the soldiers' deaths?' As in similar studies on the topic, subjects were forced to choose a yes or no answer to the question. The following table summarizes subjects' responses:

We ran an analysis of variance to measure the contributions of caring and importance of the main goal. There was a significant effect of importance, $F(1,290) = 21.428$, $p < .001$, but no significant effect of caring, $F(1,290) = .833$, $p = .362$. There was no significant interaction effect.[8]

	Important	**Unimportant**
Caring	45%	71%
Uncaring	50%	76%

Table 1 *Subjects' judgments that the lieutenant acted intentionally*[7]

[7] Our lieutenant vignettes are based on, but different from, Knobe's (2003). Knobe's cases simply specify that the lieutenant wants to take Thompson Hill, that his men might be saved or killed in the process (manipulating the harm/help dimension), and that he doesn't care what happens to them either way. Knobe did not manipulate the importance of the main goal or the agent's attitude. These differences undoubtedly explain the difference between our results and Knobe's. Interestingly, 77% judged Knobe's uncaring lieutenant who will harm his men for a goal of unspecified importance to have acted intentionally. This result most closely corresponds to our unimportant, uncaring case, both in content and in result.

[8] Although most would agree that ANOVA is robust against violations of certain assumptions and can therefore be used even with dichotomous dependent variables, we re-ran all of the analyses using logistic regression, and all of the original analyses were confirmed.

In what follows, we will examine the relevance of this data to both the agent and subject formulations of the trade-off hypothesis individually. However, we would first like to highlight the results of Case 1. On either the agent or the subject formulations of the trade-off hypothesis, Case 1 should elicit high ascriptions of intentional action, for on either formulation the agent is pursuing a benefit and clearly enduring a heavy cost in doing so. If anything is to count as a trade-off, this should. Yet only 45% of subjects thought that the lieutenant intentionally caused the deaths in this case. This result is much lower than our other cases, which do not as clearly involve trade-offs. It thus represents clear counter-evidence to the trade-off hypothesis in general.

However, there is more to be said about the above data relative to the two formulations. Let's begin with the subject formulation. The asymmetry between subjects' responses in cases involving an important main goal as opposed to an unimportant one (going left to right in the table) marks a counter-example to the subject-centered trade-off view. As discussed above, this version of the trade-off view would predict a high level of intentionality judgments for cases involving a side effect that *subjects* consider a bad thing, which is endured for something *subjects* consider a benefit. It seems clear that the side effect in all of these cases—the deaths of the soldiers—is something most subjects would consider a bad thing. Now, capturing a hill in a decisive battle in a military campaign is something many people would consider a benefit. In any case, it seems clear that more people would consider this a benefit than the main goal in the second case—capturing and quickly losing Thompson Hill. Presumably, subjects would be less likely to misinterpret the relevant features of the case when the main goal is clearly important than when the main goal is of questionable value. Hence, more subjects would be likely to see a trade-off in the case involving the lieutenant who acts for an important goal. In fact, the vignettes received quite the opposite result. Of the subjects who received the vignette involving an important goal, only 45% of subjects claimed that the agent intentionally caused the deaths of his soldiers. This is much lower than the contrast case, where a full 71% of subjects thought the deaths to have been brought about intentionally. This result eludes explanation by, and constitutes counterevidence to, the subject formulation of the trade off view. In fact, the importance of the main goal compared to the side effect is a factor which has been universally overlooked in explanations of folk judgments in side effect cases. So this result, at least, is surprising to all accounts of the Knobe effect that we know of. (In our conclusion, we will discuss the over-simplicity of existing accounts; inattention to the comparative importance of side effect to main goal is but one symptom of such over-simplicity.)

But what of the agent-centered view? Recall that, according to the agent-centered view, if more subjects consider a side effect a cost *to the agent* in one vignette than in another, then, other things being equal, more subjects will judge the agent in the former vignette to have acted intentionally. The lack of a statistically significant asymmetry between subjects' responses in cases involving a caring agent as opposed to an uncaring one (going top to bottom in the table

above) marks a counter-example to the agent-centered trade-off view. Consider, for example, the case involving an *uncaring* lieutenant who sacrifices his soldiers for an *unimportant* goal. In this case, the callous lieutenant says, 'I don't care at all what happens to our soldiers,' signaling that the soldiers' deaths are a matter of little concern to him. With his callous attitude, the uncaring lieutenant may seem to resemble the chairman in Knobe's vignettes, who cares nothing about damaging the environment (discussed above). So one might object that, as with the chairman of the board, the lieutenant is human and presumably cares something about his men. However this defense does not work for the uncaring lieutenant who is pursuing a worthless goal. We can understand why someone would harm the environment in order to make profits (even if we would behave differently) because profits are something we all consider valuable. But the lieutenant sacrifices his men for something no reasonable person could consider valuable: the brief possession of a worthless hill. To make sense of this sacrifice, one has to suppose that the lieutenant cares very little for his men indeed. Presumably, even if the goal is not something they themselves consider worthwhile, more subjects would consider the death of the men a bad thing from the agent's point of view if he expressed concern for his men, as he does in the second case above. Since this is the only divergence between the two cases, if the agent-centered view were correct we would expect a higher level of intentionality judgments for the case involving the *caring* lieutenant who sacrifices his soldiers for an unimportant goal than for the case involving an *uncaring* lieutenant who sacrifices his soldiers for the same goal. In fact, there is a 5% swing in the opposite direction as the agent trade-off hypothesis would predict. This result eludes explanation by, and constitutes counterevidence to, the agent formulation of the trade off view.

Let us take stock for a moment. In this paper we briefly reviewed the history of the debate over intentional side effects, before focusing on a recent purported explanation of these: the trade-off hypothesis, which denies the importance of moral considerations. We discussed an ambiguity in the hypothesis and raised some problems for the evidence marshaled in its favor. Finally, in this section, we presented our own evidence against each of two versions of the trade-off view.[9]

While we believe that the trade-off hypothesis lacks empirical support and does not constitute a satisfactory *full* explanation of the Knobe effect, we wish to emphasize that we do not believe it to be without value. Considerations of cost may figure into people's intentionality judgments in some cases. Certain instances of the Knobe effect might be amenable to a cost-based explanation. However, we

[9] We suggested that Machery's extra-dollar case was beside the point because it involved a means, not a side effect. Similarly, one might object that the soldiers' deaths in our cases are means to capturing Thompson Hill. But in the extra-dollar case, the end of receiving a smoothie *will not* be achieved unless the extra-dollar is paid. In our case, the objective could be achieved even without a single soldier dying, if, for example, the enemy were all terrible shots. Additionally, our results suggest that subjects are not viewing the soldiers' deaths as a means, since judgments as to whether the deaths were brought about intentionally were not universally high.

must emphasize that the trade-off view cannot give a full explanation of what's going on. Cost/benefit analyses do seem to influence assessments of intentionality, but not in a way that the trade-off hypothesis predicts.

So, what is going on in the lieutenant cases? We believe that blameworthiness has a strong role to play in explaining our results. In other studies involving intentional side effects, the agent's attitude seemed to significantly influence people's intentionality judgments. It may seem surprising that it did not appear to do so here. However, such surprise would be based on mere appearance. A closer examination reveals that the agent's attitude does, in fact, influence people's judgments in these cases. To see this, consider the two cases in which the lieutenant was pursuing an unimportant goal. In one of these the lieutenant expressed remorse for his soldiers' deaths, in the other he did not, but the difference between intentionality judgments in these cases was not significant. We believe that the reason for this is that subjects do not take the avowal of remorse to be genuine in the case in which it is voiced. It is no great mystery why people might be skeptical of the remorseful avowal. People sometimes lie about their true feelings, especially when doing so casts them in a better light. We hypothesize that people are inclined to disregard the avowal of concern when the lieutenant is sacrificing his soldiers for the fleeting (and therefore unimportant) goal of capturing Thompson Hill. The lieutenant expresses one attitude, but his actions belie another. If actions speak louder than words, then in both of the cases involving an unimportant goal the lieutenant is announcing that he doesn't care about his men, regardless of what he actually says. So we believe a judgment of a blameworthy attitude explains the high judgment of intentionality in our unimportant cases. If someone is judged to have a blameworthy attitude towards some side effect they stand in a causal relation to, they will be judged to have intentionally brought about the side effect.

We stated that people are likely to lie about their true feelings when doing so casts them in a better light. But what if someone's avowal doesn't put her in a good light? Knobe's original chairman cases are revealing in this regard. In both cases, the chairman volunteers the information, 'I don't care at all about the environment.' This avowal could not cast the chairman in a better light, so that is no reason to doubt its sincerity. Indeed, the chairman who has a blameworthy attitude toward the harm he engenders is judged to have brought it about intentionally. The helping chairman, who simply lacks a praiseworthy attitude, is not so judged.

While there is some tendency to regard an attitude that does not cast an agent in a good light as sincere, we do not think the calculus is so straightforward. The uncaring lieutenant with an important goal leads us to think otherwise. This lieutenant says, 'I don't care at all about what happens to our soldiers,' but we can identify other pressures leading subjects to think that he actually does care. For one thing, the negative side effect is endured for an important goal. But this is true of the negative side effect in the harming chairman case as well (though to a lesser degree, perhaps). The more important factor may well be the proprietary relation

a lieutenant has to his soldiers. The job description of a lieutenant includes caring for the well-being of his soldiers, along with achieving military objectives. A chairman of a company, on the other hand, is not expected to care for the well-being of the environment. These features, and perhaps others (e.g. the prototype of the straight-talking soldier, etc.), presumably serve to diminish the blameworthiness of the uncaring lieutenant with an important goal to some subjects. Therefore, this lieutenant, together with the caring lieutenant with an unimportant goal, is judged not to have acted intentionally.

While we think blameworthiness is an important factor in explaining intentionality judgments in the aforementioned side effect cases, we do not think the calculus to blameworthiness is so straightforward as previous theorists have suggested. Bad effects do not lead directly to blameworthy judgments. It is important to consider the role of the agent in bringing about the effects, his attitude in doing so, and the importance of the goals he was trying to achieve. Furthermore, while blameworthiness has a role to play in explaining these results, it does not appear to do so in other studies, such as Knobe and Mendlow, 2004, and Phelan and Sarkissian, 2008, in which non-valenced side effects were judged to have been intentionally brought about. So, while our results in this study reveal that moral considerations are important in determining the intentionality of side effects, we do not think such considerations explain everything. In the next section, we assess the current status of the debate over intentional side effects.

3. Conclusion: On the Dream of Parsimony

When do the folk consider side effects to have been brought about intentionally? This is the general question at the core of the recent debate concerning intentional action. According to standard philosophical accounts of intentional action, side effects should not be judged intentional, and so many proposals have been put forward to explain why it is that subjects *are* sometimes inclined to judge side effects to have been brought about intentionally. Yet there is little consensus about what features are causing subjects to make the relevant judgments. Knobe (2003, 2006) suggests that the relevant feature is the goodness or badness of the side effect itself; according to Nadelhoffer (2004a, 2004b), and Malle and Nelson (2003), it's not the badness itself but rather the blameworthiness of the agent that is the proximal cause of the judgment;[10] for Wright and Bengson (2008), intentionality judgments reliably follow judgments of badness *and* blameworthiness *together*; Adams and Steadman (2004a, 2004b) claim that denying the chairman did it intentionally would give rise to the conversational implicature that she does not

[10] Malle (2006) no longer takes the position that blameworthiness judgments may interfere with intentionality judgments, embracing, instead, a (preliminary) multi-process explanation, similar to the one we embrace.

bear responsibility, which subjects want to avoid; and for Machery (2008), it is whether the side effect is seen as a cost incurred in pursuit of some desired goal that explains intentional side effects. Each hypothesis comes with its attendant evidence, which is not unconvincing, and with evidence against other views. In addition, there are chiefly critical pieces (cf. Nichols and Ulatowski, 2007; Phelan and Sarkissian, 2008), which jointly disconfirm each of these views as the whole truth. It seems to us that the appropriate response to such a situation is to suppose that there will be no simple explanation of people's judgments of intentional side effects. We believe we have isolated two *further* variables that might be part of the complete explanation of subjects' intentionality judgments in side effect cases: The *importance of the main goal* relative to the side effect is clearly affecting subjects' assessments of intentionality in the above cases. And, although the agent's *avowed attitude* towards the side effect does not result in a *big* shift in intentionality judgments in the above cases, in a case reported elsewhere it did seem to importantly influence such decisions (Phelan and Sarkissian, 2008). The relevance of these features, together with those of others in the literature canvassed above, suggest that one can arrive at the concept of intentional action by means of any number of disparate routes.

We invite further research to help work out the importance of these factors and to come to a more precise understanding of the concept of intentional action. As the debate over intentional side effects stands, though, we must conclude that attempts to account for the Knobe effect by recourse to only one or two variables, though instructive, are incomplete and overreaching in their ambition. It is time to abandon the dream of parsimony.

Department of Philosophy
University of North Carolina, Chapel Hill

Department of Philosophy
Baruch College, CUNY

References

Adams, F. and Steadman, A. 2004a: Intentional action and moral considerations: still pragmatic. *Analysis*, 64, 268–276.

Adams, F. and Steadman, A. 2004b: Intentional action in ordinary language: core concept or pragmatic understanding? *Analysis*, 64, 173–181.

Knobe, J. 2003: Intentional action and side effects in ordinary language. *Analysis*, 63, 190–194.

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

Knobe, J. 2007: Reason explanation in folk psychology. *Midwest Studies In Philosophy*, 31, 90–106.

Knobe, J. and Mendlow, G. 2004: The good, the bad, and the blameworthy: understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24, 252–258.

Machery, E. 2008: The folk concept of intentional action: philosophical and experimental issues. *Mind & Language*, 23, 165–189.

Malle, B.F. 2004: *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA: MIT Press.

Malle, B.F. 2006: Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, 6, 87–112.

Malle, B.F. and Nelson, S.E. 2003: Judging mens rea: the tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences & the Law*, 21, 563–580.

Mallon, R. 2008: Knobe versus Machery: testing the trade-off hypothesis. *Mind & Language*, 23, 247–255.

Nadelhoffer, T. 2004a: Blame, badness, and intentional action: a reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, 24, 259–269.

Nadelhoffer, T. 2004b: On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology*, 24, 196–213.

Nichols, S. and Ulatowski, J. 2007: Intuitions and individual differences: the Knobe Effect revisited. *Mind & Language*, 22, 346–365.

Phelan, M.T. and Sarkissian, H. 2008: The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138, 291–298.

Wright, J. and Bengson, J. 2008: Asymmetries in judgments of responsibility and intentional action. *Mind & Language*, 24, 24–50.