Article

Unconsidered Intentional Actions. An Assessment of Scaife and Webber's 'Consideration Hypothesis'

COVA, Florian

Abstract

The 'Knobe effect' is the name given to the empirical finding that judgments about whether an action is intentional or not seems to depend on the moral valence of this action. To account for this phenomenon, Scaife and Webber have recently advanced the 'Consideration Hypothesis', according to which people's ascriptions of intentionality are driven by whether they think the agent took the outcome in consideration when taking his decision. In this paper, I examine Scaife and Webber's hypothesis and conclude that it is supported neither by the existing literature nor by their own experiments, whose results I did not replicate, and that the 'Consideration Hypothesis' is not the best available account of the 'Knobe Effect'

Unconsidered intentional actions: An assessment of Scaife and Webber's 'Consideration Hypothesis'

(penultimate draft, forthcoming in the *Journal of Moral Philosophy*)

Florian Cova

Swiss Centre for Affective Sciences, University of Geneva

7 rue des Battoirs, 1205 Geneva, Switzerland

florian.cova@gmail.com

Abstract:

The ‘Knobe effect’ is the name given to the empirical finding that judgments about whether an action is intentional or not seem to depend on the moral valence of this action. To account for this phenomenon, Scaife and Webber have recently advanced the ‘Consideration Hypothesis’, according to which people’s ascriptions of intentionality are driven by whether they think the agent took the outcome in consideration when taking his decision. In this paper, I examine Scaife and Webber’s hypothesis and conclude that it is supported neither by the existing literature nor by their own experiments, whose results I did not replicate, and that the ‘Consideration Hypothesis’ is not the best available account of the ‘Knobe Effect’.

Word count:

1. Introduction: the 'Knobe effect' and the 'Consideration Hypothesis'

In a seminal and now famous experiment, Joshua Knobe (2003) gave to a group of participants the following case:

> *Harm Case:*
>
> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also **harm** the environment."
>
> The chairman of the board answered, "I don't care at all about **harming** the environment. I just want to make as much profit as I can. Let's start the new program."
>
> They started the new program. Sure enough, the environment was **harmed**.

In this case, when asked whether the chairman intentionally *harmed* the environment, 82% of participants gave a positive answer. Now, Knobe gave another group a case very similar to this one, except that occurrences of the verb **harm** were replaced by the corresponding occurrences of the verb **help**. In this *Help Case*, only 23% of the participants answered that the chairman intentionally *helped* the environment.

How are we to account for this difference when it seems that the chairman's attitudes about the side effect are identical in both scenarios (i.e. he does not care)? Some have pointed at the most obvious difference between these two cases: differences in participants' *moral evaluations*. While, in the *Harm Case*, most participants judge that harming the environment is *bad* and the chairman *deserves* blame, it is likely that they consider that the chairman in the *Help Case* does something *good* and *does not deserve* praise. Consequently, some have

proposed that differences in ascriptions of intentionality could be explained either by the difference in (perceived) moral valences of the side effects (Knobe 2006; Pettit and Knobe 2009) or by the difference in participants' attributions of moral responsibility[1] (Nadelhoffer 2004a, 2004b; Wright and Bengson 2009), either because moral considerations are central to our ordinary concept of intentional action (Knobe 2010), or because they somehow bias our judgments (Adams and Steadman 2004a, 2004b, 2007; Nadelhoffer, 2006).

Knobe's findings (sometimes called the 'Knobe Effect') have been replicated in different populations (Knobe and Burra, 2006) and extended to new cases. For example, recent findings suggest that this phenomenon extend beyond side effects to also apply to ascriptions of intentionality in the case of means (Cova and Naar, forthcoming-a). However, not everyone accept the conclusion according to which ascriptions of intentionality are driven (even partly) by moral evaluations. Recently, many hypotheses have been advanced according to which the 'Knobe Effect' can be explained without reference to the participants' moral evaluations (Machery 2008; Hindriks 2008, 2010, 2011; Sripada, 2010; Sripada and Konrath 2011; Guglielmo and Malle, 2010; Uttich and Lombrozo 2010) [2]. One of the latest is Scaife and Webber's 'Consideration Hypothesis' (CH), according to which people's ascriptions of intentionality are driven (among other factors) by whether they think the agent took the outcome in consideration when taking his decision (Scaife and Webber forthcoming).

More precisely, Scaife and Webber consider that people ascribe intentionality only when they think that the agent took the side effect into consideration before acting, that is

[1] By "moral responsibility", I mean here the fact that the agent deserves praise or blame for what he has done. Thus, there is a difference in moral responsibility between the chairman in the *Harm Case* who deserves blame for having done something bad and the chairman in the *Help Case* who does not deserve praise for having done something good.

[2] Note that in Hindriks and Sripada's accounts of the 'Knobe Effect', intentionality judgments depend on the agent's moral attitudes (what he thinks to be morally right or wrong). However, both Hindriks and Sripada insist that this claim should not be confused with the claim that participants' moral evaluations influence their intentionality judgments. For this reason (i.e. because they think that ascriptions of intentionality do not depend on participants' moral evaluations), both accounts can be adequately described as rejecting Knobe's thesis that intentionality judgments are influenced by evaluative considerations.

only when 'the agent assigned that side-effect some level of importance relative to the importance they assigned their primary objective'.

How is that hypothesis supposed to account for the asymmetry between the *Harm Case* and the *Help* Case? According to Scaife and Webber, the chairman's 'I don't care' can be interpreted in two very different ways:

1. The sentence 'I don't care about the environment' can mean 'I am not even going to consider what that outcome is worth'
2. The sentence 'I don't care about the environment' can mean 'I have considered this outcome, and it is worth very little'

According to Scaife and Webber, people adopt two different interpretations of the chairman's 'I don't care' in the *Harm Case* and the *Help Case*. In the *Help Case*, the fact that the program is going to help the environment is just another reason to adopt the program – but the chairman has already a sufficient reason to adopt the program (i.e. making money). Thus people adopt the first interpretation: they understand that the chairman didn't take into account the fact that the program would help the environment when deliberating about whether adopting it, and conclude (according to the 'Consideration Hypothesis') that he did not intentionally help the environment. In the *Harm Case*, on the contrary, the fact that the program would harm the environment counts as a reason against adopting the program, while the chairman has an incentive to adopt it. Thus, it is natural to think that the chairman has weighed one option against another and finally given a greater weight to making money. This leads people to adopt the second interpretation: the chairman has taken into account the fact that the program would harm the environment, but judged it of little importance. In this case, the Consideration Hypothesis predicts correctly that people will consider that the chairman

intentionally harmed the environment (because he took it into consideration). To sum up, the asymmetry in attribution of consideration is supposed to drive the asymmetry in ascriptions of intentionality[3].

To support their hypothesis, Scaife and Webber argue (i) that the Consideration Hypothesis is the best explanation for the existing data and (ii) that it is also the best explanation for the new data they collected through experimentation, thus concluding that (iii) the Consideration Hypothesis (henceforth CH) is the best available account of the 'Knobe effect'. In this paper, I will argue that none of these three claims is adequately supported. After presenting two competitive accounts (section 2), I will show that there are clear cases in the literature that CH cannot accommodate while its competitors can (section 3), and that Scaife and Webber's own experiments are not much of a support for their hypothesis, particularly because their results are not easily replicated (section 4). I will conclude that, in light of current data, CH is not the best account available (section 5).

2. Three competing hypotheses

Scaife and Webber's main claim is that their theory is the best available, because it has the greater explanatory power. This is a comparative claim. Thus, to evaluate it, we must choose other theories with which we can compare the Consideration Hypothesis. In their paper, Scaife and Webber contrast their position with two influential kinds of hypotheses: the 'Effect

[3] Scaife and Webber's hypothesis should not be confused with a close kind of account, according to which a side-effect is intentional if the agents considered he had reason not to bring it about (Turner 2004; Hindriks 2008; Machery 2008). For example, Turner (2004) suggests that a side-effect is intentional if (i) the agent was aware that his action was likely to cause this side-effet, (ii) bringing about the side-effect counts against acting from the agent's perspective, and (iii) the agent does not try to prevent the side-effect from occurring. This theory and CH make very different predictions. For example, what if the chairman in the *Help Case* rejoiced about helping the environment rather than saying 'I don't care'? Turner should predict that helping the environment would still be judged unintentional, while Scaife and Webber should predict that this side-effect would be judged rather intentional (because such rejoicing would show that the chairman took the side-effect into consideration. Empirical evidence (Wible 2009; Cova and Naar forthcoming-b) suggests that CH's prediction is the right one.

Evaluation Hypothesis' (EEH) and the 'Action Evaluation Hypothesis' (AEH). Though it might well be that EEH, AEH and CH are not the best theories available, I will not introduce other theories in the competition, for it will be enough for my purposes to show that there is at least one theory that fares better than CH.

2.1. The 'Effect Evaluation Hypothesis' (EEH)

EEH is a family of hypotheses according to which the main factor in the 'Knobe effect' is the difference in the side effect's moral valence (whether it is *good* or *bad*). However, though the various versions of EEH accord themselves to make the normative valence of the side-effect the key factor in explaining the 'Knobe effect', it is worth noting that (i) they do not consider that the side-effect's valence is the only factor people consider in ascribing intentionality and (ii) that they consider that the agent's mental states are also taken into consideration. The different role they give to the agent's mental states might even constitute the key difference between the various versions of EEH.

In this paper, I will focus on the more recent version of this view put forward by Joshua Knobe (2010; Pettit and Knobe 2009). Knobe considers that the valence of the side effect plays a role in setting up a 'default point' to which the agent's pro-attitude towards the side-effect will be compared. Indeed, according to Knobe, we judge that an agent A intentionally brought about an outcome O only if A's pro-attitudes towards O are beyond and above a given 'default point'. Moral evaluations play a crucial role in setting up this default point:

> The central claim will be that people's moral judgments affect their intuitions *by shifting the position of the default*. For morally good action, the default is to have

some sort of pro-attitude, whereas for morally bad actions, the default is to have some sort of con-attitude.

Thus, in the *Harm Case*, we consider that the default is to be opposed to harming the environment because we judge harming the environment to be a bad thing. But the chairman is indifferent to harming the environment, which makes him more prone to harm the environment than if he was opposed: his attitudes towards harming the environment are above the default point, and thus he *intentionally* harms the environment (see Figure 1). Conversely, in the *Help Case*, the default is to be apt to help the environment, and the indifferent chairman is below that threshold; this is why his helping the environment is judged *unintentional* (see Figure 2).

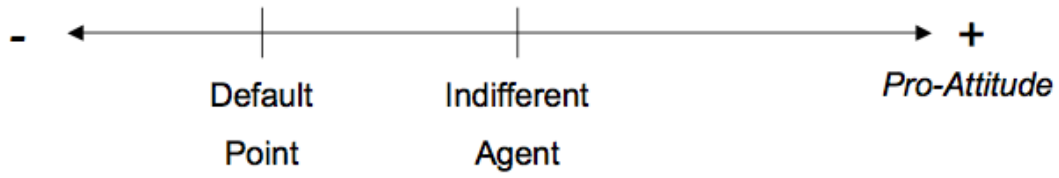

Figure 1. The *Harm Case* according to Knobe (2010)

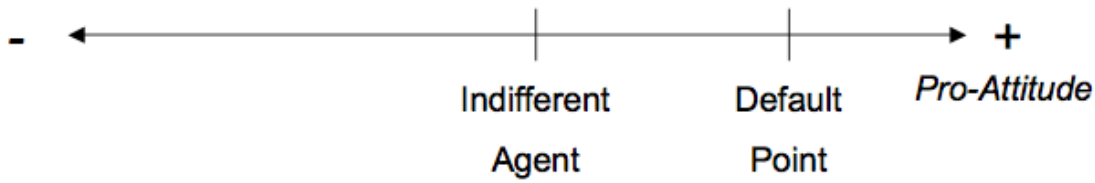

Figure 2. The *Help Case* according to Knobe (2010)

The important point is that this new version of Knobe's theory does not explain intentionality judgments solely by the side effect's moral valence; the agent's attitudes also come into play.

For example, if the side effect is bad but the agent's attitudes are very opposed to it (because the agent was forced to bring it about and brings it regretfully), this theory predicts that the action will be judged unintentional, though it is morally bad.

Because this sort of details is very important in assessing the explanatory power of a hypothesis, I won't compare here the Consideration Hypothesis to EEH in general. Rather, I will focus on this particular version of EEH. From now on, I will use 'EEH' to refer to this particular hypothesis.

2.2. The 'Action Evaluation Hypothesis' (AEH)

While EEH is the hypothesis according to which ascriptions of intentionality are driven by differences in the side effect's moral valence (whether it is *good* or *bad*), AEH is the hypothesis according to which ascriptions of intentionality are driven by differences in the agent's moral responsibility (whether he deserves *praise*/*blame* or not).

Here, I will focus on Nadelhoffer's version of AEH (Nadelhoffer 2004a, 2006), that is the hypothesis according to which a side-effect is intentional to the extent that its moral valence (*good* or *bad*) matches the moral valence of the mental states that motivated it (that is, mental states for which the agent is *praiseworthy* or *blameworthy*).

Having specified the nature of CH's competitors, I will now compare Scaife and Webber's CH to Knobe's EEH and Nadelhoffer's AEH. I will argue that, in most cases, CH fares less well than EEH and AEH.

3. Does the literature support the Consideration Hypothesis? The case of regretful agents

First, Scaife and Webber claim that 'the current experimental debate over the concept of intentional action is best explained by what [they] call the Consideration Hypothesis (CH)'. Of course, they do not review all the experimental literature about the Knobe Effect, a gigantic task that would now require a whole book[4] – but they argue that some famous cases in the literature (more precisely, Machery's cases) can be best accounted for by CH.

Indeed, Machery's *Free Cup* and *Extra Dollar* cases (see Machery 2008) constitute a problem for EEH and AEH. In the *Free* Cup case, participants judge that a person who ordered a smoothie and got it in a commemorative cup did not intentionally obtain a commemorative cup, while, in the *Extra* Dollar case, participants judge that a person who ordered a smoothie and had to pay an extra dollar for it did intentionally pay one dollar more. Now, why would people consider paying an extra dollar as more intentional than receiving a free commemorative cup? There seems to be no moral difference between these two outcomes and the corresponding actions. On the contrary, CH can explain the difference: 'people see Machery's Joe as deciding that the smoothie is worth the extra dollar on this occasion, but as not even considering the value of the free cup'.

Nevertheless, this argument is far from sufficient to establish the superiority of CH, for EEH and AEH can account for these cases once the following problem is noticed: it is possible that most people consider that paying the extra-dollar constitutes a means rather than a side-effect (Phelan and Sarkissian 2009). Since people naturally tend to consider means as more intentional than side-effects (Cova and Naar forthcoming-b), then EEH and AEH do not have to account for the difference between the two cases.

Thus, the fact that CH can provide an explanation of Machery's cases is not sufficient to ground its explanatory superiority. For all other cases surveyed by Scaife and Webber

[4] For (non-exhaustive) reviews of the literature, see (Feltz 2007) and (Cova 2010).

(Knobe and Mallon's cases[5]), EEH and AEH also have an explanation – so that there seems to be nothing in the literature cited by Scaife and Webber in favor of the superiority of CH. Now, I will argue that existing data provide us good reasons to think that CH has a lower explanatory power than the other hypotheses.

3.1. The impact of participants' moral beliefs

A first set of findings that can be opposed to CH are studies suggesting that participants' ascriptions of intentionality vary with participants' explicit (Tannenbaum *et* al. 2009) and implicit (Inbar *et al.* 2009) moral principles while CH only takes into account the agent's attitudes about his action.

For example, Tannenbaum and his colleagues gave the *Harm* and the *Help* cases to participants and asked them about their attitudes towards the environment. They found that participants who considered environment as a protected value (i.e. as having an inviolable status) were more likely to judge that the chairman intentionally harmed the environment in the *Harm Case* and less likely to answer that the chairman intentionally helped the environment in the *Help Case*. This shows that participants' moral evaluations play a role in their ascriptions of intentionality, an effect CH does not predict

However, although CH does not predict this effect, it might still try to accommodate these results. For example, defenders of CH might suggest that participants' moral evaluations have an effect on intentionality judgments only to the extent that they influence participants' perception of the agent's mental states (and considerations). However, Tannenbaum *et al.*'s results are not consistent with this hypothesis: although they found that participants who considered the environment as a protected value were more likely to judge that the chairman

[5] Mallon's cases (2008) are all cases in which participants judge bad side effects brought about by blameworthy agents to be intentional and good side-effects caused by agents who do not deserve praise to be unintentional.

desired to harm the environment in the *Harm Case* (and thus more likely to give this side-effect some consideration), they found no effect of the participants' attitudes towards the environment in desire attributions for the *Help Case*. Moreover, subsequent studies revealed an impact of participants' moral beliefs on ascriptions of intentionality but not on other mental states (such as desire and intention) attributions. Thus, CH seems unable to account for these results, while EEH and AEH clearly can, for they consider participants' moral beliefs to play a central role in ascriptions of intentionality.

3.2. Regretful agents and important goals

Empirical evidence suggests that the more an agent will express regret for having brought about a side-effect, the less this agent will be judged as having intentionally brought about this side-effect. This phenomenon has indeed been observed by Sverdlik (2004), Mele and Cushman (2007), Sripada (2010) and Guglielmo and Malle (2010). Now, it is very likely that the more an agent is seen as bringing about a side effect regretfully, the more participants will consider that he took this side effect into consideration. So, CH should predict that regretful agents' actions will be judged more intentional, whereas it is the opposite trend that has been observed so far.

To test whether CH can account for intentionality judgments in cases of regretful agents, I used the following scenario created by Mele and Cushman (2007):

> *Pond (Regret):*
>
> Al said to Ann: "You know, if you fill in that pond in the empty lot next to your house, you're going to make the kids who look for frogs there sad." Ann

---

Thus, Mallon's results are consistent with both EEH and AEH.

replied: “I know that I’ll make those kids sad. I like those kids, and I’ll definitely regret making them sad. But the pond is a breeding ground for mosquitoes; and because I own the lot, I am responsible for it. It must be filled in.” Ann filled in the pond, and, sure enough, the kids were sad.

I also created a similar scenario, in which the agent does not express regret at all:

Pond (No Regret):

Al said to Ann: “You know, if you fill in that pond in the empty lot next to your house, you’re going to make the kids who look for frogs there very sad.” Ann replied: "So what? Why on earth should I bother about how those kids would feel? The pond is a breeding ground for mosquitoes; and because I own the lot, I am responsible for it. It must be filled in.” Ann filled in the pond, and, sure enough, the kids were very sad.

I gave each scenario to 30 participants (for a total of 60 participants) through the Amazon Mechanical Turk. Age mean was 30.4 and 19 were women. Each participant read only one text and then answered five questions (on a scale ranging from 1 (NO) to 7 (YES), 4 (IN BETWEEN) being the midpoint):

1. *Intentionality Question:* Did Ann intentionally make the kids sad?
2. *Consideration Question:* In taking her decision, did Ann take into consideration the fact that filling the pond would make the kids sad?
3. *Reason Question N°1:* Did Ann consider the fact that it would make the kids sad a reason *not to* fill the pond?

4. *Reason Question N°2:* Did Ann consider the fact that it would make the kids sad a reason *to* fill the pond?
5. *Regret Question:* Was Ann sorry to make the kids sad?

Questions 2 to 5 constitute various ways of measuring how much consideration participants thought Ann gave to the side effect. For analysis, answers to Questions 3 and 4 were summed in a composite score[6].

Question	Regret	No Regret	Comparison (Welsh t-test)
(i) Intentionality	2.77 (2.10)	4.17 (2.28)	t=2.5, df=57.6, p<0.05*
(ii) Consideration	5.59 (1.88)	3.90 (2.76)	t=-2.8, df=51.3, p<0.01**
(iii-iv) Reasons	6.30 (3.16)	5.03 (3.23)	t=-1.5, df=56.8, p=0.13
(v) Regret	5.60 (1.75)	2.41 (1.99)	t=-6.5, df=55.6, p<0.001***

Table 1. Means (and Standard Deviations) for the *Pond* cases.

Results are presented in Table 1, with the results of the Welsh t-test for each comparison. As can been seen, the side effect is judged less intentional in the *Regret* condition that in the *No Regret* condition. However, the measures for consideration go in the opposite direction: participants were more likely to consider that Ann took the outcome into consideration and considered it more as a reason to or not to act in the *Regret* condition. Running Pearson's test

[6] What I want to probe is whether participants consider that Ann took the side-effect into consideration. As an agent is more likely to take the side-effect into consideration if it takes it to count as a reason *for* or a reason

for correlations, I found no correlation at all between intentionality ratings and the Considerations and Reasons answers, but observed a very strong inverse correlation between intentionality and regret attributions (r=-0.51, df=58, p<0.001).

These data can easily be accommodated by EEH which considers the agent's attitude towards the side effect as a crucial factor for intentionality attributions. They also can be accommodated by AEH: although I did not test for it, it is likely that a regretful agent will receive less blame than an agent who shows no regret for his bad action. But I do not think that CH can account for them, as there is no relation here between intentionality judgments and whether the agent was perceived as having taken the side effect into consideration[7].

However, one might object that our measures of consideration are problematic. For example, a reviewer suggested that perhaps the Consideration and Reasons questions are not adequate measures, for they might be interpreted by participants as bearing only on *good* reasons and consideration. Thus, it might be that participants attribute more consideration and reasons in the *No Regret* condition but do not report them, thinking they are only asked about the good reasons.

To test for this alternative hypothesis, I designed a third version of the *Pond* case:

Pond (Hate):

Al said to Ann: "You know, if you fill in that pond in the empty lot next to your house, you're going to make the kids who look for frogs there very sad."Ann replied: "I know. But that's all the more reason to fill the pond! Those little brats always annoyed me! Anyway, the pond is a breeding ground for

---

against acting (or both at the same time), it makes sense to sum questions 4 and 5 to have a measure of whether the side-effect was perceived as a consequence likely to be taken into consideration by Ann.

[7] One objection might be that Ann in the *No Regret* case is in fact perceived as intending to make the kids sad and that is the reason why his action is judged more intentional. However, I have direct measures for this hypothesis: participants' answers to Question 4 ('Did Ann consider the fact that it would make the kids sad a reason *to* fill the pond?') Clearly, if Ann is perceived as intending to make the kids sad, then she will also be

mosquitoes; and because I own the lot, I am responsible for it. It must be filled in." Ann filled in the pond, and, sure enough, the kids were very sad.

The procedure was the same than for the two previous scenarios: 30 participants recruited through Amazon Mechanical Turk read this scenario and answered the same five questions.

Unsurprisingly, participants tended to judge that Ann intentionally made the kids sad (Question 1: *M*=4.43, *SD*=2.34). Also, participants did not hesitate to answer that Ann took the side-effect into consideration (Question 2: *M*=5.4, *SD*=1.85) and that she considered making the kids sad as a reason to fill the pond (Question 4: *M*=4.1, *SD*=2.28)[8]. Thus, it seems that participants did not give these questions a normative interpretation, and that they really judged that Ann did not take the side-effect into consideration in the *No Regret* case, though most of them judge that she intentionally made the kids sad.

4. Scaife and Webber's experiments

However, Scaife and Webber do not rely only on already existing data: they also produce new experiments in support of their hypothesis. Do these experiments really support CH by providing results that only CH is able to explain? If so, maybe that will be enough to counterbalance the difficulties I have previously mentioned.

4.1. Scaife and Webber first experiment

perceived as considering making the kids sad as a reason *to* fill the pond. Overall, I found no significant difference between the two cases for this question (in fact, ratings tended to be higher in the *Regret* case).

[8] For comparison, the mean answer to Question 3 for the *Regret* case was 2.9 (*SD*=2.07). Thus, it doesn't seem that people tend to report less bad reasons than good reasons.

Their first experiment consists in taking the *Harm Case* and introducing a second (good) side effect. Here is their vignette:

> The vice president of the company went to the chairman of the board and said: "We're thinking of changing the way the factory works. There are three factors to consider: it will increase profits, it will improve safety, but it will increase pollution".
>
> The chairman of the board answered: 'All I care about is increasing profits, so let's do it'.
>
> So they altered the factory and, sure enough, this had the effects the vice president had predicted.

In this case, only 45.2% of participants judged that the chairman intentionally increased pollution. In a further replication of the experiment, only 41% of participants gave that same judgment. So, it seems that introducing a second (good) side-effect along the bad one leads participants to judge the bad side-effect less intentional.

These results are indeed surprising, and in favor of CH. Scaife and Webber rightly claim that CH has an explanation for this effect. According to them, introducing a good side-effect leads people to truly believe the chairman when he claims that he does not care about the side effects of his action. Thus, since people consider the chairman less apt to take into consideration the fact that altering the factory will increase pollution, they judge his action less intentional. On the contrary, EEH and AEH do not seem to have a good explanation for this phenomenon.

Nevertheless, there is a problem with this experiment, namely that they compare two scenarios that differ in many respects. To conclude that introducing a second side-effect leads

to a decrease in intentionality ratings, they compare their vignette to Knobe's original *Harm Case*. But there are other differences between their case and Knobe's case; for example, one might think that the phrasing of the side-effect as 'increasing pollution' is less negative than the much stronger 'harming the environment'. Moreover, Scaife and Webber do not observe the fact that introducing the good side effect leads people to consider that the chairman didn't take the bad side-effect into consideration: they just hypothesize it. It would be good to ground this assumption on direct measures of the degree of consideration participants attribute to the agents.

For these reasons, I decided to replicate Scaife and Webber's experiment and introduce an equivalent scenario without the good side effect:

> The vice president of the company went to the chairman of the board and said: 'We're thinking of changing the way the factory works. There are two factors to consider: it will increase profits but it will increase pollution.'
> The chairman of the board answered: 'All I care about is increasing profits, so let's do it'.
> So they altered the factory and, sure enough, this had the effects the vice-president had predicted.

I gave each scenario to 50 participants (for a total of 100 participants) through the Amazon Mechanical Turk. Age mean was 31.1 and 48 were women. Each participant read only one text and then answered five questions:

1. Did the chairman intentionally increase pollution? (on a scale ranging from 1 (NO) to 7 (YES), 4 (IN BETWEEN) being the midpoint)

2. Did the chairman take into account the fact that altering the factory would increase pollution? (on a scale ranging from 1 (NO) to 7 (YES), 4 (IN BETWEEN) being the midpoint)
3. How bad is it to increase pollution? (on a scale ranging from 1 (NOT BAD) to 7 (VERY BAD), 4 (IN BETWEEN) being the midpoint)
4. How much blame does the chairman deserve? (on a scale ranging from 1 (NONE) to 7 (A LOT), 4 (IN BETWEEN) being the midpoint)
5. How much did the chairman regret to increase pollution? (on a scale ranging from 1 (NONE) to 7 (A LOT), 4 (IN BETWEEN) being the midpoint)

Question	One side effect	Two side effects	Comparison (Welsh t-test)	Correlation (Pearson's product-moment correlation)
(i) Intentionality	5.20 (1.85)	4.96 (2.04)	t=0.6, df=97.1, p=0.54	-
(ii) Consideration	3.56 (2.57)	3.24 (2.30)	t=0.7, df=96.8, p=0.54	t=1.96, r=0.19, p=0.05°
(iii) Badness	6.08 (1.48)	6.16 (1.50)	t=-0.27, df=96.0, p=0.79	t=4.49, r=0.42, p<0.001***
(iv) Blame	5.84 (1.52)	5.62 (1.82)	t=0.65, df=94.7, p=0.52	t=4.99, r=0.45, p<0.001***
(v) Regret	2.56 (2.13)	2.12 (1.35)	t=1.23, df=82.9, p=0.22	t=-0.77, r=-0.08 p=0.44

Table 2. Means (with Standard Deviations) and Correlations with Intentionality Judgments for the replication of Scaife and Webber's first experiment

As can be seen in Table 2, there was no question for which I found a significant difference between the two cases. Thus, the two main assumptions of Scaife and Webber's argument are ungrounded: introducing a second side-effect does not reduce intentionality rating and does not make participants less likely to perceive the agent as taking the side-effect into consideration. Correlations do not lend much more support to CH: although there's a marginally significant correlation between intentionality and consideration ratings, I found a much greater and more significant correlation between intentionality and badness and blame ratings, which makes respectively EEH and AEH in better positions.

Nevertheless, there is something very puzzling in these results: although Scaife and Webber found that most participants considered increasing the pollution as not intentional, I found the reverse pattern of answers, with most participants considering increasing the pollution as intentional. For the case with two side effects, 58% of participants gave an answer above 4. Thus, I even failed to replicate Scaife and Webber's own results[9].

4.2. Scaife and Webber's second experiment

Nevertheless, Scaife and Webber have a second experiment in support of their hypothesis. They used the following case:

> *Parent (Original Scenario):*
>
> The doctor said to the parent: 'although your daughter is no longer showing any symptoms, we could run some tests to ensure that it won't recur; but the tests are painful, so it's up to you'.

[9] I tried to replicate Scaife and Webber's results in two other experiments: one similar to the one described here and the other in which participants were asked only the intentionality question and had to answer only by YES or NO. In both cases, I found no difference between the two cases and observed that more than half of the participants judged the side effect to be intentional.

After some consideration, the parent said: 'the tests should be run, to be on the safe side'. And so the tests were run.

Did the parent intentionally inflict pain on the child?

In this case, they found that 66.3% of the participants judged that the parent had intentionally inflicted pain on the child. They use these results as an argument against AEH: clearly, we wouldn't blame the parent for having the tests run, so blame attribution cannot account for these results. From the perspective of EEH, these results are also a bit puzzling; clearly, the parent is perceived as acting reluctantly, as he is *forced* to have the tests run[10]. So, we shouldn't observe high intentionality ratings like these. On the contrary, CH can easily explain these data, as it is clear that the parent took into consideration the fact that having the tests run will harm his daughter, and this is why his action is judged intentional.

To test for this explanation, I designed a similar case, in which the parent does not take time to consider the side-effect of their action:

Parent (Fast Decision):

The doctor said to the parent: "although your daughter is no longer showing any symptoms, we could run some tests to ensure that it won't recur; but the tests are painful, so it's up to you".

Without taking time to think about it, the parent immediately answers: 'I don't care about the pain, the tests should be run'. And so the tests were run.

[10] When I say that the parent is *forced*, I do not deny that he has the possibility to refuse the tests. I just want to say that he has compelling reasons to accept the test even though he would prefer the child not to suffer.

I gave each scenario to 30 participants (for a total of 60 participants) through the Amazon Mechanical Turk. Age mean was 31.5 and 24 were women. Each participant read only one text and then answered the following questions:

1. Did the parent intentionally inflict pain on the child? (on a scale ranging from 1 (NO) to 7 (YES), 4 (IN BETWEEN) being the midpoint)
2. Did the parent take into consideration the fact that running the tests would inflict pain on the child? (on a scale ranging from 1 (NO) to 7 (YES), 4 (IN BETWEEN) being the midpoint)
3. How bad was it to inflict pain on the child? (on a scale ranging from 1 (NOT BAD) to 7 (VERY BAD), 4 (IN BETWEEN) being the midpoint)
4. How much blame does the parent deserve? (on a scale ranging from 1 (NONE) to 7 (A LOT), 4 (IN BETWEEN) being the midpoint)
5. How reluctant was the parent to inflict pain on his child? (on a scale ranging from 1 (NOT AT ALL) to 7 (A LOT), 4 (IN BETWEEN) being the midpoint)

Question	Original Scenario	Fast Decision	Comparison (Welsh t-test)	Correlation (Pearson's product-moment correlation)
(i) Intentionality	1.97 (1.94)	2.20 (1.92)	t=-0.47, df=58, p=0.64	-
(ii) Consideration	4.90 (2.19)	3.13 (2.45)	t=2.95, df=57.3, p<0.01**	t=-0.58, r=-0.07, p=0.57
(iii) Badness	3.40 (1.73)	3.93 (1.96)	t=-1.11, df=57.1, p=0.27	t=1.97, r=0.25, p=0.05°
(iv) Blame	2.2 (1.71)	3.2 (2.02)	t=-2.07, df=56.4, p<0.05*	t=2.20, r=0.28, p<0.05*
(v) Regret	3.83 (1.76)	2.93 (2.00)	t=1.85, df=57.1, p=0.07°	t=1.14, r=0.15, p=0.26

Table 3. Means (with Standard Deviations) and Correlations with Intentionality Judgments for the replication of Scaife and Webber's second experiment

As can be seen in Table 3, manipulation was successful and participants were indeed less likely to consider that the agent took the side-effect into consideration in the *Fast Decision* condition. Nevertheless, intentionality ratings didn't decrease accordingly, in contrast with what CH would have predicted. On the contrary, they (non-significantly) tended to be higher in the case without consideration. This lack of relationship between consideration and intentionality attributions is highlighted by the lack of correlation between answers to the first two questions.

In a more problematic way, I was unable to replicate Scaife and Webber's results: while they found that most participants judged the action intentional, the intentionality ratings I gathered are pretty low. To the intentionality question, only 10% of participants gave an

answer above 4, while 73% gave "1" as answer. This contrasts with the higher ratings to the consideration question, with 53% of participants giving an answer above 4, and only 13% giving "1" as answer.

These results are easily explained by EEH (because it is easy to perceive the parent as being *forced to* inflict pain on his daughter) and AEH (because of the strong correlation between intentionality and blame ratings). But CH seems unable to explain why we find low intentionality ratings for a case with high consideration ratings[11].

5. Assessing the Consideration Hypothesis' explanatory power

How could CH be defended against the results presented in this paper? A first possibility is to emphasize what Scaife and Webber already mentioned in their paper: that it is not possible to directly measure whether people attribute 'consideration' to the agent. Given that 'consideration' can be understood in various ways, it is possible that when participants answer that the agent took the outcome into consideration, they only mean that he merely acknowledged its existence, without integrating it into his deliberation. For this reason, we couldn't rely on participants' answers to the 'consideration question', because it would be impossible to know if they reflect the relevant meaning of 'consideration'. For the same reason, the 'reasons questions' could not help us either, as participants could at the same time consider that the agent considered the side-effect as a reason against acting but did not take it into consideration during his deliberation. In conclusion, the correlation I found could not be used as an argument against CH, because we cannot exactly know what these questions measure.

[11] Following this failure, I once again tried to replicate Scaife and Webber's results in two other experiments: one similar to the one described here and the other in which participants were asked only the intentionality question and had to answer only by YES or NO. In both cases, I observed that more than half of participants judged the side effect not to be intentional.

This line of answers strikes me as quite implausible, for two reasons. The first reason is that it seems to me that the same counter-argument cannot be used against the 'regret question': clearly, if one regrets having to do X to do Y and knows that bringing about Y will cause X, it is clear that one will consider X as a reason not to do Y and will give X some thought when deliberating about whether doing Y or not.

The second reason is that this line of answers cannot account for the fact that, in my experiments, answers to the 'consideration question' varied exactly as we would have expected if these answers reflected attributions of 'consideration' in the sense relevant to Scaife and Webber's hypothesis. For example, in the *Pond* cases, I found higher answers to the 'consideration question' when the agent expressed more regret. And in the *Parent* cases, I found lower ratings when the parent answered quickly and declared that he didn't care. If participants really understood the terms 'taking into consideration' in a sense that is not the one relevant to CH but equivalent to 'acknowledging' or 'thinking about', why did I observe these variations? It seems much more plausible to think that participants' answers track 'consideration' in the sense that is relevant to CH, at least to some extent.

But, let's grant that my measures of 'consideration' are not reliable. Does it mean that my argument against CH collapse? Not at all. For if Scaife and Webber refuse all possibility to ask people about how much consideration they attribute to the agent, they must acknowledge that there is at least another way to guess how much consideration participants will attribute – some kind of *a priori* guess. For, if there is no way of determining how much consideration participants attribute (either by measures or by *a priori* guess), then there will be no way of determining what their intentionality attributions will be. Thus, if there is no way of determining (at least in some cases) how much consideration participants will attribute, then CH cannot make any prediction. And a theory that cannot make any prediction has clearly a very low explanatory power.

So, does CH fare better when we use our own intuitions and 'common sense' to predict how much consideration participants will attribute? I don't think so. It rather seems plausible to think that people will be more likely to attribute consideration to a regretful agent (in the *Pond* cases) and to a parent who takes time to think rather than shrug his shoulders and say 'I don't care' (in the *Parent* cases). But in all these cases, intentionality ratings varied contrary to what CH would predict on this basis.

Maybe Scaife and Webber could say that I'm wrong and that it is not that clear that regretful or caring agents will be more likely to give consideration to the side effects of their actions. So, let's consider the following case, a variation on a case by Knobe and Kelly (Knobe and Kelly 2004; see also Knobe 2004):

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

Did the terrorist intentionally save the Americans?

In this case, I think it is clear that the terrorist gave some importance to the fact that defusing the bomb would save the Americans (after all, he intended to kill them). Also, it seems clear that participants will attribute consideration to the terrorist in the relevant sense. Thus, CH should predict that participants will answer that the terrorist intentionally saved the Americans.

I gave this scenario to 30 participants through the Amazon Mechanical Turk. Age mean was 29.2 and 7 were women. Participants only had to answer the intentionality question in a binary way ('YES' or 'NO')[12]. Only 7 participants (23%) gave the 'YES' answer (as would predict EEH, AEH and IDH). Thus, it seems that CH's predictions are false.

Now, there are two ways of reacting to these results. One is simply to consider that CH is false. The other is to argue for CH by saying that maybe the participants didn't attribute consideration to the terrorist. But if it is impossible to guess whether participants will attribute consideration, even in cases as clear and evident as *Terrorist*, and if it is impossible to measure whether participants attribute consideration, then CH cannot make any prediction. Given that the other hypotheses *can* make predictions, and that some of these predictions *do* work, this line of response inevitably lead to the conclusion that CH should be abandoned.

Thus, for CH to be a legitimate hypothesis, its defendants must grant that there is at least a legitimate way to predict how much consideration participants will attribute. Considering only the cases in which it seemed to me that both answers to the consideration question and guessing on *a priori* grounds clearly led to the same predictions about the consideration participants will attribute, I compared the respective merits of each hypothesis in Table 4 (a '++' indicates that the hypothesis can account for the results and could have

predicted them , a '+' indicates that the hypothesis is merely consistent with the results, a '-' indicates that the hypothesis is incompatible with the results).

Case	EEH	AEH	CH
Machery's *Extra Dollar*	+	+	++
Machery's *Free Cup*	++	++	++
Pond	++	++	-
Parent	++	++	-
Terrorist	++	++	-

Table 4. A comparison of EEH, AEH and CH

Using Table 4, I calculated a score representing the explanatory power of each hypothesis: '++' was equivalent to 1 point, '+' to 0 point, and '-' to -1. The final score are 4 for EEH, 4 for AEH, and -1 for CH. So, contrary to what Scaife and Webber claimed, it seems that CH is far from having the best explanatory power (at least for the cases I examined here).

To sum up, I argued in this paper that Scaife and Webber's claims about CH being the best hypothesis for the Knobe effect is not warranted. The 'Consideration Hypothesis' is unsupported both by the current state of the literature and Scaife and Webber's own experiments. In most of the cases I presented here, the 'Consideration Hypothesis' was the less powerful hypothesis. As a result, I conclude that we have no compelling reason to endorse CH.

[12] Note that, before answering the intentionality question, participants had to answer a comprehension check ('Did the terrorist know that his son was in the nightclub when he planted the bomb?'). Participants who failed the comprehension check were excluded.

Another conclusion is that people willing to propose new accounts of the Knobe Effect should take into consideration past findings in the experimental literature and make sure that their hypothesis is consistent with these results. Here, I pointed at two conditions such an account should fulfill: it should explain (i) why participants' moral opinions seem to impact on their intentionality judgments and (ii) why regretful agents are perceived as acting less intentionally. Some theories cannot account for both phenomena (e.g. Hindriks 2011[13]) while others can account only for the first (e.g. Holton 2010) or only for the second (e.g. Guglielmo and Malle 2010; Sripada 2010). We should then focus on accounts that propose an account of both these phenomena (e.g. Nadelhoffer 2004a; Knobe 2010; Cova *et* al. forthcoming) and find new ways of deciding between them[14][15].

References

Adams, Fred and Steadman, Annie. 2004a. Intentional action in ordinary language: core concept or pragmatic understanding. *Analysis* 64: 173-181.

Adams, Fred and Steadman, Annie. 2004b. Intentional action and moral considerations: still pragmatic. *Analysis* 64: 264-267.

Adams, Fred and Steadman, Annie. 2007. Folk concepts, surveys and intentional action. In *Intentionality, Deliberation and Autonomy: The Action-Theoretic Basis of Practical*

[13] Hindriks (2011, footnote 1) acknowledges that, in its current state, his theory cannot account for the case of Ann. Nevertheless, he suggests that "an expected and undesired effect is not brought about intentionally when the agent has sufficient normative reason to bring it about". However, this qualification is not enough to account for the difference between our two *Pond* cases, for Ann has as much normative reason to fill the pond in the *Regret* and *No Regret* cases. Thus, rather than normative reasons, it seems that it is the agent's attitudes are the real factor responsible for the difference.

[14] For example, Knobe's EEH can be considered superior to Nadelhoffer's AEH because it can account for cases of asymmetry in which there is no blame or praise to be attributed (Phelan and Sarkissian 2008).

[15] I thank Joshua Knobe, Robin Scaife and Jonathan Webber, as well as three anonymous reviewers for their comments on earlier versions of this paper.

Philosophy, edited by Christoph Lumer and Sandro Nannini (Aldershot: Ashgate Publishers): 17-34.

Cova, Florian. 2010. Le statut intentionnel d'une action dépend-il de sa valeur morale ? Une énigme encore à résoudre. *Vox Philosophiae* 2 : 100-128.

Cova, Florian, Dupoux, Emmanuel and Jacob, Pierre. forthcoming. On doing things intentionally. *Mind and Language*.

Cova, Florian and Naar, Hichem. forthcoming-a. Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology.*

Cova, Florian and Naar, Hichem. forthcoming-b. Testing Sripada's deep self model. *Philosophical Psychology.*

Feltz, Adam. 2007. The Knobe effect: a brief overview. *The Journal of Mind and Behavior*, 28, 265-277.

Guglielmo, Steve. and Malle, Bertram. 2010. Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin* 36: 1635-1647.

Hindriks, Frank. 2008. Intentional action and the praise-blame asymmetry. *Philosophical Quarterly* 59: 713-720.

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

Hindriks, Frank. 2011. Control, intentional action and moral responsibility. *Philosophical Psychology* 24: 787-801.

Holton, Richard. 2010. Norms and the Knobe Effect. *Analysis* 70: 1-8.

Inbar, Yoel, Pizarro, David, Knobe, Joshua and Bloom, Paul. 2009. Disgust sensitivity predicts intuitive disapproval of gays. *Emotion* 9: 435-439.

Knobe, Joshua. 2003. Intentional action and side-effects in ordinary language. *Analysis* 63: 190-193.

Knobe, Joshua. 2004. Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology* 24:270-279.

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

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

Knobe, Joshua and Burra, Arudra. 2006. Intention and intentional action: a cross-cultural study. *Journal of Culture and Cognition* 1-2: 113-132.

Knobe, Joshua and Kelly, Sean. 2004. Can one act for a reason without acting intentionally? Unpublished manuscript, Princeton University.

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

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

Mele, Alfred and Cushman, Fiery. 2007. Intentional action, folk judgments and stories: sorting things out. *Midwest Studies in Philosophy* 31: 184-201.

Nadelhoffer, Thomas. 2004a. Praise, side effects and intentional action. *Journal of Theoretical and Philosophical Psychology* 24: 196-213.

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

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

Pettit, Dean and Knobe, Joshua. 2009. The pervasive impact of moral judgment. *Mind and Language* 24: 586-604.

Phelan, Mark and Sarkissian, Hagop. 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.

Phelan, Mark and Sarkissian, Hagop. 2009. Is the trade-off hypothesis worth trading for? *Mind & Language* 24: 164-180.

Scaife, Robin and Webber Jonathan. forthcoming. Intentional side-effects of action. *Journal of Moral Philosophy*.

Sripada, Chandra. 2010. The Deep Self model and asymmetries in folk judgments about intentional action. *Philosophical Studies* 151: 159-176.

Sripada, Chandra and Konrath, Sara. 2011. Telling more that we can know about intentional action. *Mind & Language* 26: 353-380.

Sverdlik, Steven. 2004. Intentionality and moral judgments in commonsense thoughts about action. *Journal of Theoretical and Philosophical Psychology* 24: 224-236.

Tannenbaum, David, Ditto, Peter and Pizarro, David. 2009. Different moral values produce different judgments of intentional action. Unpublished manuscript, University of California.

Turner, Jason. 2004. Folk intuitions, asymmetry, and intentional side effects. *Journal of Theoretical and Philosophical Psychology,* 24: 214-219.

Uttich, Kevin and Lombrozo, Tania. 2010. Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116: 87-100.

Wible, Andy. 2009. Knobe, side effects, and the morally good business. *Journal of Business Ethics*, 85, 173-178.

Wright, Jennifer and Bengson, John. 2009. Asymmetries in folk judgments of responsibility and intentional action. *Mind and Language* 24: 237-251.