The Folk Concept of Intentional Action: Empirical approaches

Florian Cova

Swiss Centre for Affective Sciences, University of Geneva

Penultimate Draft

Final version to be published in W. Buckwalter & J. Sytsma (Eds.), *Blackwell Companion to Experimental Philosophy*. Wiley-Blackwell.

What makes an action intentional and when does someone do something intentionally? This is one of the main and most important questions within the branch of philosophy called "action theory". According to Alfred Mele, "central to the philosophy of action is a concern to understand *intentional* action" (1992, p.199). However, the notion of intentional action does not find its origins in the philosophy of action: rather, it predates philosophy of action, as part of our everyday understanding of human behavior. Thus, an adequate account of intentional action cannot stray too far from our common understanding of which actions count as intentional, and should not lose sight of our "folk" concept of intentional action, or else run the risk of missing its target. For this reason, empirical investigations of the folk concept of intentional action can be expected to be highly relevant for philosophy of action, and it is no wonder that numerous experimental philosophers have tried to dissect the common sense category of "intentional action".

Philosophical accounts have traditionally emphasized three factors: *foreknowledge*, *choice* and *control*. Malle and Knobe (1997) have empirically investigated which factors laypeople deem relevant for intentional action and their results closely matched these models, as people insisted on the five following components: awareness and belief, desire and intention[i], and skill.

However, if these first studies gave results that were mostly in agreement with what should be expected from philosophical accounts of intentional action, later researches revealed puzzling phenomena that suggested that one important feature of our folk concept of intentional action had been overlooked: far from being a purely descriptive component of folk psychology, our concept of intentional action would have a normative (or evaluative) component as well. If this turned out to be true, this would dramatically change our understanding of the folk concept of intentional action.

In this chapter, I provide a critical though comprehensive review of the empirical literature on the folk concept of intentional action. After defending what I think to be the best explanation for the results of these studies, I go back to the implications for action theory.

1. Two puzzles for intentional action: the Knobe Effect and the Skill Effect

Recently, experimental evidence suggested that our judgments about whether an action counts as intentional are sensitive to normative (or evaluative) factors. Evidence for the putative influence of such considerations on ascriptions of intentionality arises from the study of two phenomena, both discovered by Joshua Knobe: the "Knobe Effect" and the "Skill Effect".

1.1. The Knobe Effect

The Knobe Effect (also known in the literature as the "Side-effect Effect") can be described as the observation that whether a side effect is considered intentional highly depends on its valence. Consider the following scenario (Knobe 2003a):

> *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 his original study, Knobe found that 82% of the people surveyed answered 'yes' to the question, "did the chairman of the board intentionally harm the environment?" When given

the same vignette, but this time with the word 'harm' changed into 'help' (the *Help Case*), only 23% responded positively when asked if the chairman of the board intentionally helped the environment. This striking asymmetry has since been replicated in other languages and cultures (Knobe & Burra, 2006; Cova & Naar, 2012a; Dalbauer & Hergovich, 2013), in young children (Leslie, Knobe & Cohen, 2006; Pellizzoni, Siegal & Surian, 2009), in people suffering from Asperger Syndrome (Zalla & Leboyer, 2011) and in patients with cerebral lesions to the prefrontal cortex (Young *et al.* 2006). More recently, it has been shown that the means whereby an agent achieves her goal exhibit asymmetries similar to the Knobe Effect (Cova and Naar, 2012a).

1.2. The Skill Effect

The Skill Effect can be described as the fact that normative (or evaluative) considerations modulate the impact of the control factor on ascriptions of intentionality. Consider the following scenario:

> *Bull's-eye (Skill):* Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bull's-eye. He raises the rifle, gets the bull's-eye in the sights, and presses the trigger.
>
> Jake is an expert marksman. His hands are steady. The gun is aimed perfectly…
>
> The bullet lands directly on the bull's-eye. Jake wins the contest.

In this case, 79% of participants answered that Jake intentionally hit the bull's-eye. Now, consider the *Bull's-eye (No-Skill)* case in which the second paragraph is modified:

> But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild… Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

In this case, only 28% of participants answered that Jake intentionally hit the bull's-eye. These results show that ascriptions of intentionality also depend on the degree of control the agent exerts on his action. Nothing surprising, one would say. But now consider the following pair of scenarios:

> *Aunt (Skill):* Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger.
> Jake is an expert marksman. His hands are steady. The gun is aimed perfectly… The bullet hits her directly in the heart. She dies instantly.
> *Aunt (No-Skill):* […] But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild… Nonetheless, the bullet hits her directly in the heart. She dies instantly.

In the *Skill* condition, 95% of participants answered that Jake intentionally killed his aunt. 76% did so in the *No-Skill* condition. Once again, the outcome is perceived as less intentional when the agent exerts less control. But this difference is much smaller in this case (19%) than in the *Bull's-eye* pair (51%). These results suggest that the contribution of control to ascriptions of intentionality is greatly diminished when the outcome is bad (or good, see Knobe 2003b).[ii]

1.3. Two questions for accounts of our concept of intentional action

These two sets of experiments suggest that normative or evaluative considerations can (i) play a role in our ascriptions of intentionality and (ii) modulate the extent to which the *control* factor has an impact on these ascriptions. Thus, two questions can be asked:

1) Are these effects really due to normative (or evaluative) considerations impacting ascriptions of intentionality?
2) If they are, should we consider the impact of these considerations as part of our normal application of the concept of intentional action, or as a bias leading us to improperly apply this concept?

In the rest of this chapter, I address both questions by surveying the different accounts of the Knobe Effect that have been proposed.[iii] Throughout this survey, I also list the conditions an account of the concept of intentional account must fulfill if it is to count as a proper and satisfying account. Indeed, most accounts available in the literature fail because they do not engage with the whole empirical literature, but only focus on a subset of data. By listing these conditions, I hope to help future accounts not to fall into these shortcomings.

The reason why I chose to focus on the Knobe Effect is that it has attracted much more attention than the Skill Effect. Admittedly, this is a problem: an adequate understanding of our concept of intentional action should account for both phenomena. We thus have our first condition for a proper account of the folk concept of intentional action:

(*Comprehensiveness*) A proper account should explain both the Knobe Effect and the Skill Effect.

2. Normative and evaluative considerations: a constitutive component of intentional action, or just a bias?

Let's admit for the moment that the Knobe Effect is due to normative or evaluative considerations shaping our ascriptions of intentionality, and let's ask whether we should consider this influence as a mere bias, or as revealing something about the deep structure of our concept of intentional action.

2.1. The Knobe Effect as a bias

One possible way of reacting at the Knobe Effect is to accept the existence of the effect but to claim it teaches us nothing about the folk concept of intentional action because it is just an instance of people misapplying this concept or speaking against their mind: normative and evaluative considerations 'bias' participants' answer and distort their judgments about intentional action. This hypothesis comes in two flavors: psychological and linguistic.

The 'psychological' version of this approach argues that people are mistaken, and do not properly apply their concept of intentional action in the *Harm* case (Malle & Nelson, 2003; Pinillos et al., 2011; Sauer & Bates, 2013). A popular version of this account is the 'Blame Bias' account. This account rests on previous psychological evidence that people are ready to distort their attributions of key conditions for blame (such as judgments about causation or attributions of mental states) in order to motivate and justify their negative assessment of a given character (see Alicke, 2008; Alicke & Rose, 2010). According to this account, (Nadelhoffer, 2004a, 2004b, 2006c; Alicke, 2008), when we are not driven by our willingness to praise or blame, we usually do not consider side effects or action performed with lack of control to be intentional, but the perceived blameworthiness or praiseworthiness of the agent can lead us to attribute intentionality to the agent's action to motivate and justify

our attributions of blame and praise. Thus, the asymmetry in the *Help* and *Harm* cases is explained by another asymmetry: that we are prone to blame the chairman in the *Harm Case* but not to praise him in the *Help Case*.

Nevertheless, the scope of this account is limited. First, one version of this account, according to which it is participants' negative affective reactions at the chairman's personality that distorts our attributions of intentionality (Nadelhoffer, 2006), fails to explain why the Knobe Effect can be found in populations that have impoverished affective reactions due to brain damage (Young *et al.* 2006). Second, no available version of this account can explain the existence of asymmetries similar to the Knobe Effect in non-moral, non-emotionally biased cases, such as the following (drawn from Wright & Bengson, 2009 and inspired from Knobe & Mendlow, 2004):

> *Sales (Decrease):* The VP of a company went to the chairperson of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also decrease sales in New Jersey." The chairperson of the board answered, "I don't care at all about decreasing sales in New Jersey. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, profits increased and sales in New Jersey decreased.

In this case, most participants answered that the chairperson intentionally decreased the sales in New Jersey. Now, in the *Sales (Increase)* case, the word "decrease" was replaced by "increase", and most participants answered that the chairperson did not intentionally increase the sales in New Jersey. However, this asymmetry was not correlated with an asymmetry in judgments of praise and blame: participants did not tend to attribute more blame to the VP in the *Decrease* case than praise in the *Increase* case.

In response to these challenges, Wright and Bengson (2009) have developed an affect-free version of this account, according to which people use the following heuristic: an agent must have done something intentionally if he is responsible for it. Switching from 'blame' and 'praise' to the more neutral concept of 'responsibility' allows them to accommodate non-moral cases such as the *Sales* cases. However, two criticisms can be raised. First, in participants with Asperger Syndrome, the Knobe Effect is present even though these patients attribute a lot of praise (and thus responsibility) to the chairman in the *Help* case (Zalla & Leboyer, 2011). Second, even in normal participants, responsibility judgments do not accurately track ascriptions of intentionality in all cases (Cova & Naar, 2012b).

The 'linguistic' version of the hypothesis is the one that grants the less depth to the Knobe Effect: according to it, participants do not even really believe that the chairman intentionally harmed the environment in the *Harm* case. A first version, developed by Adams and Steadman (2004a, 2004b, 2007), consider the Knobe Effect as the product of conversational implicatures: people use the word "intentionally" in morally bad cases to conversationally imply that the agent deserves blame. However, we already saw that blame judgments were not the best predictor of intentionality judgments.

Another version, put forward by Guglielmo and Malle (2010a), argues that the Knobe Effect is only due to the forced-choice setting of the original experiment. Indeed, people might choose to answer that the chairman intentionally harmed the environment to convey something else: that he did it 'willingly' or 'recklessly'. To test for this hypothesis, Guglielmo and Malle had participants in the *Harm* choose between several descriptions of the chairman's action which ones were correct and most accurate. The four descriptions were that the chairman 'willingly, 'knowingly', 'intentionally' and 'purposefully' harmed the environment. Guglielmo and Malle observed that most participants (86%) chose the description involving 'knowingly' and that very few (1%) chose the description containing intentionally.

Such results might suggest that the original Knobe Effect was just an artifact imputable to the constraints of the original task and that people only speak truly when they are given the choice between several options (Woolfolk, 2013). However, this presupposes that Guglielmo and Malle's multiple choice is a better measure than forced-choice settings. To determine whether a new method of measurement is reliable, one has to calibrate it on uncontroversial cases. To this purpose, I created a modified version of the *Harm* case in which harming the environment is a means rather than a side-effect (Cova, 2014b). Most philosophical accounts of intentional action would predict that, in this case, harming the environment is intentional, since means are intentional (but see: Cova & Naar, 2012a). However, when I used Guglielmo and Malle's method, it turned out that the claim according to which the chairman intentionally harmed the environment was rarely chosen. So, we face a choice: conclude that their measure is unreliable, or get ready for a very revisionary account of intentional action.

To summarize, most 'biasing' accounts fail because the asymmetry observed by Knobe can be reproduced in cases involving neither upsetting events, nor blame attributions, such as the *Sales* case. This leads us to formulate our second condition:

> (*Morally neutral cases*) A proper account should explain why asymmetries similar to the original Knobe Effect can be observed in cases involving no moral violation.

2.2. Evaluative considerations as part of our concept of intentional action

Thus, it seems that the Knobe Effect cannot be simply explained away as the product of bias. But if we accept that the effect is driven by normative or evaluative considerations, what conclusions should we draw from its existence? According to Knobe himself (2006), his results show that folk psychology is not purely descriptive, as is often assumed, but also

designed to fulfill evaluative functions. Starting from the rather mundane observation that ascriptions of intentionality are important inputs for our judgments of praise and blame, Knobe advances the hypothesis that our concept of intentional action has in fact specifically 'evolved' to play this role: that ascriptions of intentionality play a fundamental role in attributions of praise or blame is not an accident, but reveals their true function.

Knobe distinguishes between two kinds of evaluations: the judgment that an action has led to a bad (or good) outcome[iv], and the judgment that one deserves blame (or praise) for a given action. With this distinction in mind, the question is: how do we go from the first kind of judgment (that someone did something *bad*) to the second kind (that this person is *blameworthy*)? Knobe's answer is that we use certain tools, among which the concept of intentional action: the function of our concept of intentional action is "to track the psychological features that are most relevant to praise and blame judgments." (p.225)

However, it is not clear that the way we go from evaluation of the outcome (*good*/*bad*) to evaluation of the agent (*praiseworthy*/*blameworthy*) is the same for both bad and good actions. As Knobe points out, "different psychological features will be relevant depending on whether the behavior itself is good or bad" (p.225). Indeed, as I already pointed out, there is an asymmetry in the way we attribute blame for bad actions and praise for good actions: knowing that one's action will have bad consequences seems to be enough to deserve blame for them, while only knowing that one's action will have good consequences is not enough to deserve praise – one also had to intend to bring about these specific consequences. Thus, if ascriptions of blame or praise are asymmetric, and if our concept of intentional action is designed to drive our ascriptions of blame and praise, then it is only natural that it is sensitive to different features according to whether we apply it to good or bad actions.

In Knobe's first account of the Knobe Effect, if an outcome counts as bad, then it should be considered intentional if the agent either *tried* to bring it about or *foresaw* that

acting in the way he did would bring it about. Thus, a bad behavior can be intentional even if it is only foreseen (as harming the environment in the *Harm Case*). On the contrary, when the outcome is good, it should be considered intentional only if the agent was specifically trying to bring it about (which is not the case for helping the environment in the *Help Case*) (see Figure 1).

However, although this account can explain the asymmetry between the *Harm* and the *Help* cases, it fails to explain other cases such as Mele and Cushman (2007)'s *Pond* case in which the protagonist must fill the empty pond next to her lot to prevent an infestation of mosquitoes, thereby making the children who used to play next to the pond sad. The key idea is that, in this case, the protagonist expresses deep regrets at bringing about the bad outcome of making the children sad. Most of the participants judged that the protagonist did not intentionally made them sad, though she clearly foresaw this bad outcome, which speaks directly against the account I presented in this section.[v]

Thus, Knobe's first account fails because it does not accommodate the fact that the attitude an agent takes towards a given side-effect also impacts our ascriptions of intentionality, independently of what he intends or believes. Therefore, an adequate theory must also satisfy the following condition:

> (*Attitudes*) A proper account should accommodate the fact that the agent's attitude towards a side-effect (whether he brings it about reluctantly, indifferently, or joyfully) has an impact on our ascriptions of intentionality.

2.3. The pervasive impact of moral judgment

However, Knobe has substantially modified his account since 2006, and the current version actually satisfies (*Attitudes*). In a series of papers (Pettit & Knobe, 2009; Knobe, 2010a),

Knobe has extended his thesis about the concept of intentional action (i.e. that the concept is not only descriptive but partly evaluative) to a wide array of psychological concepts: "desiring", "deciding", "advocating", "being in favor of", "believing" or "knowing". Indeed, asymmetries similar to the Knobe Effect for intentional action have been observed for these concepts: for example, people are more likely to answer that the chairman desired to *harm* the environment (in the *Harm* case) than to say that the chairman desired to *help* the environment (in the *Help* case) (Pettit & Knobe, 2009). Similar patterns have also been found to affect the way people understand causation (Hitchcock & Knobe, 2009) and the way they build action trees and see one action as counting as a means or a side-effect (Knobe, 2010b). This led Knobe to conclude that the impact of evaluative considerations on our application of psychological concepts is not specific to our concept of intentional action, but is rather a deep and fundamental feature of folk psychology.

According to Knobe (2010), the fact that asymmetries similar to the Knobe Effect can be found in many other cases put pressure on what counts as an adequate account of the Knobe effect: it is not enough to have an explanation that works for judgments about intentional action, if this account does not also apply to these other asymmetries. Surely, one cannot just postulate that *all* asymmetries have the same source and explanation. However, psychological concepts such as desire and choice seem tightly closed to our concept of intentional action, and it seems that a proper account of the Knobe Effect should also apply to these asymmetries. This argument has since been dubbed the 'argument from unification' (Hindriks, 2014) and has become a topic of debate (Sauer, 2014). It also suggests that if, as I proposed, a proper account of the folk concept of intentional action should account for the Knobe Effect, then a proper account should also account for asymmetries in related concepts. This condition might seem too demanding for an account of our folk concept of intentional action, and I will get back to this question in section 4, but it seems likely that an account of

our folk concept of intentional action will be *ceteris paribus* better if it explains the relationship between the Knobe Effect and these other asymmetries.

Pettit and Knobe (2009) propose an account that satisfies this demand by explaining all asymmetries in our application of motivational psychological concepts. Let's start with the following example: suppose a beer and a cup of coffee are both at the temperature of 20°C. Application of 'cold' to these beverages might plausibly yield a true statement in the coffee case and a false statement in the beer case because people rate each liquid relative to a *default value* that specifies what it is supposed to be like for it to be cold (Figure 2). Similarly, in the *Harm* and *Help* cases, judgments about intentional action are evaluated by comparing the chairman's actual attitudes towards the outcome to a *default value*, and this default value differ depending on the outcome. More precisely, Pettit and Knobe argue that the default value is partly determined by what we *normatively expect* the agent to desire: we think people *should* desire to help the environment, so the default point in the *Help* case is an attitude above indifference. But we think people *should* be reluctant to harm the environment, so the default point is set below indifference in the *Harm* case (Figure 3). This is why the same attitude on the chairman's behalf (his indifference) leads people to judge his action intentional in the *Harm* case (in which indifference is above the default point) but not in the *Help* case (in which indifference is below the default point).

This account has two main advantages. First, it can be extended to a great number psychological attitudes, and thus not only explain the Knobe Effect, but also a lot of similar asymmetries, thus fulfilling the demand of unity. Second, because it relies on the comparison of the agents' attitudes to a standard, it takes into account both the effect of normative considerations and the fluctuation of the agent's attitudes, thus fulfilling (*Attitudes*). For example, it is true that, in the *Pond* case, the side-effect is bad, and that the default point for

intentionality is set below indifference (the agent *should* be reluctant to cause the relevant side-effect). However, in this case, the agent is not indifferent: she is in fact very reluctant to make the children sad. Thus, her attitudes towards the outcome are still below the default point, and this is why she is judged not to have brought about a bad side-effect intentionally.

2.4. Norms and the Knobe Effect

Thus, Knobe's current account is an improvement on his earlier account, since it fulfills both (*Attitudes*) and (*Morally neutral cases*). However, one should note that it does not fulfill (*Comprehensiveness*), since he does not account for the Skill Effect. Moreover, Knobe himself found a case that poses a threat to his later account. Consider the following case (drawn from Knobe, 2007):

> *Nazi Law (Violation):* In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps. Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes. The Vice-President of the corporation said: 'By making those changes, you'll definitely be increasing our profits. But you'll also be violating the requirements of the racial identification law.' The CEO said: 'Look, I know that I'll be violating the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!' As soon as the CEO gave this order, the corporation began making the organizational changes.

In the *Fulfillment* case, all occurrences of 'violating' were replaced by 'fulfilling.' In the *Violation* case, 81% of participants said the CEO intentionally violated the requirements of the law, while only 30% of participants said he intentionally fulfilled the requirements of the law in the Fulfillment case. This asymmetry cannot be explained by Knobe's account: surely, it is better if a Nazi law is violated than fulfilled, and we consider that the CEO *should* desire to violate it.

To account for such cases, Holton (2010) has proposed a competing account according to which the asymmetry in ascriptions of intentionality is directly caused by another asymmetry in norms. Thus, Holton claims that there is a fundamental asymmetry concerning norms: to intentionally violate a norm, all one needs to do is to knowingly violate it, whereas to intentionally conform to a norm one needs to be counterfactually guided by it. And since whether a norm was intentionally violated or conformed are supposed to influence our ascriptions of intentionality, we have an explanation for the Knobe Effect.

One advantage of Holton's account is that it allows the asymmetry in ascriptions of intentionality to be driven by whatever norm is salient to participants, and not only by the norms participants actually endorse. Thus, participants' answers to the *Nazi Law* cases nicely fits Holton's account, provided that we suppose that the Nazi law is a more salient norm in this context that moral norms. And indeed, recent evidence suggest that judgments about intentional action can be manipulated by making certain norms more or less salient (Robinson, Stey & Alfano, in press). Thus, we can add a new condition to our collection:

> (*Norms*) A proper account should account for the fact that norms seem able to drive asymmetries similar to the Knobe Effect independently of side-effects' valence.

However, accounts in terms of norms also fail on neutral cases, in which it seems possible to have asymmetries without norms, such as the *Sales* case: surely, there is no norm against the VP decreasing sales. Moreover, in its current form, Holton's account does not fit the *(Attitudes)* condition either. However, it is an intriguing alternative to Knobe's account, which allows for a wider impact of normative considerations – for it allows participants' ascriptions of intentionality to be impacted by norms these participants do not share.[vi]

3. A Knobe Effect without evaluative considerations?

So far, I have only surveyed accounts that acknowledged some influence of normative and evaluative considerations upon ascriptions of intentionality. However, a substantial number of accounts do not, and aim at giving an account of the Knobe Effect that does not appeal to such considerations. Some of them even claim that there is no effect to begin with.

3.1. Can the Knobe Effect be explained away?

Indeed, it has been suggested by some that the Knobe Effect can be reduced to a more fundamental, and hopefully less puzzling phenomenon. The more popular version of this strategy appeals to an asymmetry in the agent's desires and claims that the Knobe Effect is just an artifact that can be easily explained by the fact (i) that we consider an outcome more intentional when the agent actually intended or desired this outcome and (ii) that participants tend to consider that the chairman in the *Harm* case desires to harm the environment while the chairman in the *Help* case does not desire to help the environment (Guglielmo & Malle, 2010a).[vii] Of course, in Knobe's original vignettes, both chairmen explicitly state that they just do not care, but participants need not take these statements at face value. Some have

argued that participants in fact tend to attribute more desire towards the outcome in the *Harm* case, because behaviors that are contrary to norms are more informative than behaviors that conform to norms: it takes nothing to conform a norm, but since transgressing a norm is a deviation from 'normality', it seems to tell us something about the agent's motivations (Uttich & Lombrozo, 2010).

From this point of view, the Knobe Effect seems to tell us nothing new or surprising about our concept of intentional action: people just attribute more intentional action in the *Harm* case because they are more likely to see the agent in this case as desiring harming the environment (Guglielmo & Malle, 2010a). As evidence in favor of this account of the Knobe Effect, it is often pointed at the fact that intentionality ratings tend to be lower in the *Harm* case when the agent is regretful and tend to be higher in the *Help* case when the agent expresses his joy at helping the environment. However, as we have seen, this is not incompatible with Knobe's latest account of the asymmetry: rather, Knobe explicitly acknowledges that variations in agents' attitudes will be reflected in ascriptions of intentionality. What then is the difference between Knobe's latest account and the kind of account we just described? Simply that Knobe would claim that the level of attitude required for the action to be judged intentional will be lower than in the *Harm* than in the *Help* case, because the default point is set lower in the *Harm* case. On the contrary, those who wish to 'explain away' the Knobe Effect by making attitudes the only relevant factor should predict that, attitudes towards the side-effect being equal, ascriptions of intentionality should be identical in the *Harm* and *Help* cases. However, this is not what is happening when one takes a careful look at Guglielmo and Malle (2010)'s results: rather, their results seem to vindicate Knobe's predictions (see Figure 4). Thus, it is far from clear that the influence of moral considerations on judgments of intentionality can be dismissed by the mere observation that intentionality ratings vary along the agent's attitudes towards the outcome.

Other accounts have followed Guglielmo and Malle in trying to reduce the Knobe Effect to a more basic and less puzzling asymmetry. For example, Sripada (2010, 2012) proposed that the asymmetry could be explained by the concordance between the outcome and the agents' 'Deep Self' (i.e. his deeply held values) and argued that the influence of normative considerations on judgments of intentionality disappeared once controlled for these attitudes (Sripada & Konrath, 2011).[viii] Similarly, Shepard and Wolff (2013) have proposed to reduce the asymmetry to another asymmetry in causal judgments: indeed, participants are more willing to say that the chairman caused the side-effect in the *Harm* case than in the *Help* case. However, it seems that, even once controlled for all these factors, normative and evaluative considerations still play a role in shaping our judgments about intentional action (Cova, 2014b).

2.2. A normativity-free Knobe Effect

Another class of account acknowledges the existence of a genuine and irreducible asymmetry in ascriptions of intentionality, but claims that normative and evaluative considerations have nothing to do with it: to explain the Knobe Effect, one only needs to broaden the range of agent's attitudes one deems relevant to judgments about intentional action. Two families of such accounts can be distinguished:

(i) According to the first kind of accounts, a side-effect counts as intentional when the agent has *a reason not to* bring it about. A famous example is Machery's 'Trade-Off Hypothesis' (Machery, 2008) according to which the difference between intentional side effects and non-intentional side effects amounts to a difference in the level of cost that is foreseen by the agent in order to achieve his goal. When we conceptualize something as a cost incurred by an agent, we come to think that the cost has been incurred intentionally. For

example, in the *Harm Case*, harming the environment is a cost. Thus, it is intentional. But, in the *Help Case*, helping the environment is not a cost.[ix]

As has been pointed out (e.g. Mallon, 2008), Machery's claim is ambiguous. Are intentional side-effects those conceptualized as a cost by the participant reading the vignette or by the agent described in the vignette? If one chooses the first reading, then the account can no longer explain the asymmetry in the *Nazi Law* case. Thus, the account can only escape the difficulty faced by other accounts by choosing either the second reading or a blend of both reading (according to which a side-effect becomes intentional if it is conceived as a cost by the participant or by the agent).

This is probably why most versions of such accounts have endorsed a version of the thesis according to which it is the agent's attitude (and not the participant's) that matters (e.g. Turner, 2004). For example, in his 'Normative Reasons account of Intentional Action' (in short: NoRIA), Hindriks (2008, 2010, 2011, 2014) claims that a side-effect can only be intentional if the agent acted (i) in spite of the fact that he did not want to bring this side-effect about, or (ii) in spite of the fact that he believed that bringing this side-effect about constituted a normative reason against acting the way he did.

Accounts that adopt the agent's perspective fail on two points. First, these accounts cannot fulfill (*Attitude*). According to this account, the more an agent sees the side-effect as a cost, an obstacle, or a reason not to act, the more their bringing about this side-effect should be considered as intentional. However, take again the *Pond* case: in this case, the agent (Ann) is very reluctant to bring about the side-effect (making the children sad). Thus, she certainly conceives of the side-effect as a cost, and as a reason not to act – but the side-effect is judged smostly unintentional. What happens now when we contrast this case to a similar case in which the agent makes the children sad, but does not worry about it, and is wholly indifferent? The account in terms of 'reasons not to' should predict that ascriptions of

intentionality will be even lower. However, we get exactly the opposite: in such cases, the side-effect is judged much more intentional (Cova, 2014a). In summary, these accounts make predictions that run directly against (*Attitudes*).

Second, these accounts cannot explain cases such as the *Terrorist* case (Knobe, 2004b; Cova, 2014a), in which a terrorist has planted a bomb in a nightclub to kill Americans but reluctantly defuses the bomb when he discovers that his own son is in it. When asked whether the terrorist intentionally save the Americans, most people answer that the terrorist *did not* intentionally save the Americans. However, he clearly had *a reason not to* save them, and this clearly was a *cost* to him. Thus, the accounts we just surveyed fail to account for this case.[x]

(ii) A second kind of accounts focuses on the agent's deliberation and the extent to which he takes the potential side-effect of his future action in consideration. According to the 'Deliberation Model' (Alfano, Beebe, & Robinson, 2012), the asymmetry in ascriptions of intentionality should be explained by earlier asymmetries in other kinds of mental states such as beliefs, desires, and intentions. An agent who learns that one course of action leads to violating a norm is perceived by participants as more likely to stop and deliberate carefully about whether he should violate this norm or not. Since deliberation leads to the formation of other mental states – such as beliefs, desires, and intentions – this difference underlies all asymmetries we have mentioned so far. Similarly, Scaife and Webber (2013) advanced the 'Consideration Hypothesis', according to which people ascribe intentionality only when they think that the agent took the side-effect into consideration before acting, that is only when the agent assigned that side-effect some level of importance relative to the importance they assigned their primary objective. However, these accounts fail on the same points as accounts in terms of *reason not to*: they should predict that regretful agents are seen as acting more intentionally than indifferent agents, and should predict that most people consider the side-effect intentional in cases such as *Terrorist*. Moreover, recent data suggest that the

consideration the agent gave to a side-effect is not positively but negatively correlated with intentionality ratings (Cova, 2014a) and that attributions of beliefs do not predict judgments about intentional action once controlled for other factors (Cova, 2014b).

4. The multiple meanings of "intentionally"

So far, we haven't met an account of the folk concept of intentional action that satisfies all the condition I have listed, that is an account that

> (*Comprehensiveness*) explains both the Knobe Effect and the Skill Effect,
>
> (*Norms*) explains the apparent norm-sensitivity of intentionality ascriptions,
>
> (*Morally neutral cases*) explains why we have asymmetries in cases in which there seems to have no moral violations,
>
> (*Attitudes*) and explains why regretful agents are considered as acting less intentionally.

In this last section, I will sketch what I believe constitutes such an account, and I will do so by starting from a hypothesis we haven't considered yet: that there is not only one folk concept of intentional action, but several.

4.1. The 'Interpretive Diversity' hypothesis

The 'Interpretive Diversity' hypothesis has first been advanced by Nichols and Ulatowski (2007). According to them, people actually ascribe two different meanings to the noun phrase "intentional action": (i) 'having a motive' and (ii) 'foreknowing'. Furthermore, one and the same person can adopt one or the other according to the context. Thus, in the *Harm Case*,

when they use "intentionally", most people mean "done with foreknowledge" and they judge the chairman as having harmed the environment "intentionally" (because he knew his action would harm the environment). But, in the *Help Case*, when they use "intentionally", most people mean "done with a motive" and they consider the chairman as not having helped the environment "intentionally" (because he lacked a motive for helping the environment).

Following Nichols and Ulatowski, others have tried to distinguish the different meanings in which the world "intentionally" is used. For example, Cushman and Mele (2008) have defended that people have "two and half folk concepts" of intentional action, a proposal that has since been outbid by Lanteri (2013), who proposed to cut the folk concept of intentional action in "three and a half". Meanwhile, Sousa and Holbrook (2010) have tried to explain the Skill Effect by distinguishing between two interpretations of "intentionally". Though none of these accounts can explain all the asymmetries we have seen so far[xi], I think they are on the right track: "intentionally" is polysemous and part of what is happening in the Knobe Effect is due to this polysemy.

4.2. The Knobe Effect as a linguistic phenomenon

But why think that the Knobe Effect is a linguistic phenomenon? Did we not rule out this possibility in section 2.1? Not really: we only ruled out the possibility that the Knobe Effect might be a purely pragmatic phenomenon, but not the possibility that it might be a semantic phenomenon.

There are several reasons to think that the Knobe Effect take place at the semantic level, rather than at a deeper level (the one of folk psychology, for example). Let's go back to Pettit and Knobe's analogy with the beer and coffee temperature: it is true that norms seem to impact our judgments about whether a beverage should count as hot, and that these norms vary along beverage. However, this clearly does not mean that such norms have an impact on

our capacity to estimate temperatures: though we judge the coffee cold and the beer hot, we can simultaneously judge that they both are at the same temperature. Thus, the fact that the truth of linguistic judgments is impacted by norms and normative considerations does not show that the underlying psychological competencies are. Thus, it could be that the truth-value of sentences containing "intentionally" is sensitive to norms and normative considerations without our folk psychology being suffused with such considerations.

One might argue that the case of normative considerations is different from the one of standards about a beverage's temperature. But take the following case (Egré & Cova, in press):

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

In this case, most people answered 'YES' to the first question and 'NO' to the second. However, exactly the same number of children died and survived. This suggests that the truth-value of sentences containing "many" is impacted by moral and evaluative considerations. Should we then conclude that our ability to count and estimate quantities is fundamentally moral and driven by moral considerations? This seems preposterous.

Another clue that asymmetries such as the Knobe Effect only are present at a linguistic level and not at the deeper level of folk psychology is that they do not influence the way participants predict agents' behavior. Indeed, although people are more likely to say that the chairman *desired* to bring about the side-effect in the *Harm* case, they do not perceive him as

more likely to deliberately harm the environment (for no other reason) than the chairman of the *Help* case (see Cova, Dupoux & Jacob, 2010 for details).

As I said, Knobe's account, according to which attributions of intentionality depend on default points that are influenced by evaluative and normative considerations, is directly inspired from the semantics of gradable expressions. It is thus possible to keep this intuition while discarding Knobe's claim about our folk psychology fulfilling irreducible normative and evaluative functions. We thus reach a purely linguistic account of the Knobe Effect in which we treat statements including the word "intentionally" as having truth-conditions similar to the ones of sentences including gradable terms, so that statements about "intentionally" are true only if the agent's attitudes towards the side-effect go beyond a certain default point (Egré, 2010, 2013).[xii]

Of course, we have seen that Knobe's original account could not explain all cases we surveyed. Thus, some extra adjustments are necessary. The main one is introducing the notion of 'expectation' (Mandelbaum & Ripley, 2010; Cova, Dupoux & Jacob, 2012). Expectations can be both descriptive (we expect someone to do something because it is something he often does) and normative (we expect someone to do something because he ought to). We can thus modify Knobe's account by postulating that the default point is set by both kinds of expectations. When several expectations are in conflict, the most salient (most of the time, the moral ones) win and determine where the default point should be set. This explain why the asymmetry seems to be driven primarily by what we morally expect the agent to desire (in the *Harm*, *Help* and *Terrorist* cases) but why, in absence of salient moral expectations, the default point is set by conventional norms (the *Nazi Law* case) or by what, based on his situation, the agent seems more likely to desire (the *Sales* case).

4.3. Three concepts of intentional action

The account we just sketched works for most of the asymmetries, at least those about the agent's attitudes: it works not only for "intentionally" but also for "desire", "intent" or "being opposed to". Thus, we have an account that does not apply only to "intentionally". However, it should be noted that there seems to be something peculiar to "intentionally". First, asymmetries tend to be more extreme in the case of intentionally than in the case of other psychological predicates (Pettit & Knobe, 2009). Second, it is the only psychological predicate for which the Skill Effect has been observed so far.[xiii] Therefore, the account based on analogy with the semantics of – which applies to different psychological predicates – is not the whole story: there must be something more, something specific to "intentionally".

Based on informal observation – most of my participants, when tested in groups of friends, disagreed about what "intentionally" meant, I have argued that the word "intentionally" can have three different meanings (Cova, Dupoux & Jacob, 2012):

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

Note that, though such an account might account for the Knobe Effect by itself, it is also compatible with the account sketched in the previous section: "intentionally" can have different meanings that are all such that the truth-conditions of statements including them

mimic the truth-conditions of statements including gradable terms. One advantage of combining the two accounts is that only the account based on an analogy with the semantics of gradable predicates can be extended to other predicates.

The main idea of this account is that participants' understanding of "intentionally" will be influenced by what they take to be the meaning of interest, and this will differ according to the context. If they are told about something they expect the agent to be in favor of (a good action, or something the agent desires), then the first meaning will seem to be the most relevant. If they are told about something they expect the agent to be opposed to (a bad action, something the agent does not want), the second meaning will appear to be the most relevant. Finally, if their attention is drawn towards the amount of control and skill a certain action requires, then the third meaning will turn out to be the most salient.

Let's apply this account to the Knobe and Skill Effects. In the *Harm* case, participants expect the agent to be against harming the environment, and thus select the second meaning. But the chairman, because he is indifferent, does not unwillingly harm the environment: thus, his action is judged intentional. In the *Help* case, participants expect the agent to be in favor of helping the environment, and thus select the first meaning. But the chairman does not desire to help the environment; thus, his action is considered unintentional. In the *Bull's-eye* case, the setting makes salient questions of skill, and thus the third meaning, which is sensitive to the amount of control the agent exerts upon his action is selected. But, when the bull's-eye is replaced by the agent's *Aunt*, someone we expect him not to kill, then the second meaning becomes the most salient, and this meaning is not sensitive to the agent's control.

Aside from explaining both the Knobe and Skill Effects, this account has other non-negligible advantages. First, contrary to pragmatic accounts, it does not have to claim that participants do not really speak their mind and use "intentionally" to convey something else. Second, contrary to bias accounts, it does not tax participants with general irrationality.

However, contrary to Knobe's account, it does not commit us to the paradoxical thesis that folk psychology is suffused with moral considerations, and it fits the general intuition that our concept of intentional action is primarily a descriptive psychological concept. Surely, moral considerations do influence the interpretation of "intentionally" the participants adopt, and have an impact on where the default point is set. But this does not make "intentionally" a more 'moralized' term than "cold", "hot" or "many".[xiv] Thus, this account preserve the main intuitions between what seemed conflicting accounts by acknowledging at the same time (i) that judgments about intentionality are impacted by moral considerations without participants being mistaken and (ii) that the concept of intentional action has a descriptive, psychological function.

5. What consequences for action theory?

The Skill Effect and more notably the Knobe Effect are two puzzling phenomena that have drawn a lot of attention, sometimes at the expense of the bigger picture. As can be seen from the present survey, numerous theories or the folk concept of intentional action have been proposed, and many do not live up to the task because they are too focused on a certain phenomenon, or a rather limited set of cases. My aim in this chapter was thus to shed light on the multiple puzzling phenomena a satisfying account of the folk concept of intentional action should account for.

Now, one might wonder what (philosophical) use there is for an appropriate understanding of our folk concept of intentional action. I began the survey by stressing that a philosophical account of intentional action should not stray too far from the folk concept. But what lessons should action theory draw from the empirical investigations I surveyed? In fact, the answer to this question depends on what turns out to be the right account of phenomena

such as the Knobe Effect. At one end of the spectrum, we have accounts for which the Knobe Effect doesn't teach us anything new or interesting about the folk concept of intentional action or intentional action itself: the Knobe Effect is just a bias, and lead people to use "intentionally" in ways that do not reflect their core understanding of what constitutes an intentional action (Guglielmo & Malle, 2010a). Thus, empirical investigations of the folk concept of intentional action have very little theoretical import: at best, they have practical value, by allowing us to detect biases that we should strive to correct (Nadelhoffer, 2006c).

At the other end of the spectrum, we have Knobe's view, according to which the results of these researches should lead us to completely revise our understanding of folk psychology and of the nature and function of the concept of intentional action (Knobe, 2010). In this case, philosophers working in action theory would have to decide whether they choose to follow the folk understanding of intentional action, or whether they consider better to take a more revisionist stance. In both cases, these results would reveal a tension between the goal of following commonsense notions of intentional action and the goal of reaching a concept of intentional action one would be able to apply to actions independently from one's moral values and commitments (an important goal, given the role that the concept of "intentional action" is supposed to play in moral and legal debates).

Finally, between these two extremes, there are several intermediate positions. If we follow accounts that consider that the Knobe Effect cannot be explained away but can nonetheless be explained without appeal to normative and evaluative considerations, then these empirical investigations do not have the revolutionary consequences Knobe expect them to have, but can still teach us interesting facts about the factors people take into account when considering an action as intentional (such as the consideration one gives to a particular side-effect; see Scaife & Webber, 2013). In this case, philosophical accounts of intentional action might consider integrating these factors into their account. And if we follow accounts

according to which "intentionally' comes into different senses, then we might end up criticizing as doomed the philosophical project aiming at finding a definition of "intentional" that would encompass all our (non-biased) intuitions about the truth-value of sentences including "intentionally" (Cova, Dupoux & Jacob, 2012). In this case, philosophers might renounce the project of finding a unitary and proper definition of intentional action, and instead start discussing which meaning of "intentionally" is relevant for the moral and normative questions the concept is supposed to address.

Anyway, it is impossible to determine on *a priori* grounds what will be the philosophical consequences of current empirical investigations of our folk concept of intentional action. To put it otherwise: this is also an empirical question.

Acknowledgments

This research was supported by the National Centre of Competence in Research (NCCR) Affective sciences financed by the Swiss National Science Foundation (n° 51NF40-104897) and hosted by the University of Geneva. I also thank Wesley Buckwalter, Hichem Naar, Justin Sytsma and one anonymous reader for their comments on a previous version of this chapter.

References

Adams, Fred, and Annie Steadman. 2004a. "Intentional action in ordinary language: core concept or pragmatic understanding." *Analysis*, 64: 173-181. DOI: 10.1111/j.1467-8284.2004.00480.x

Adams, Fred, and Annie Steadman. 2004b. "Intentional action and moral considerations: still pragmatic." *Analysis*, 64: 264-267. DOI: 10.1111/j.0003-2638.2004.00496.x

Adams, Fred, and Annie Steadman. 2007. "Folk concepts, surveys and intentional action." In *Intentionality, Deliberation and Autonomy: The Action-Theoretic Basis of Practical Philosophy*, edited by Christoph Lumer and Sandro Nannini, 17-34.Aldershot: Ashgate Publishers.

Alfano, Mark, Beebe, James, and Brian Robinson. 2012. "The centrality of belief and reflection in Knobe-effect cases." *The Monist*, 95: 264-289. DOI: 10.5840/monist201295215

Alicke, Mark D. 2008. "Blaming badly." *Journal of Cognition and Culture*, 8: 179-186. DOI: 10.1163/156770908X289279

Alicke, Mark D., and David Rose. 2010. "Culpable control or moral concepts?" *Behavioral and Brain Sciences*, 33: 330-331. DOI10.1017/S0140525X10001664

Cokely, Edward T., and Adam Feltz. 2009. "Individual differences, judgment biases, and Theory-of-Mind: Deconstructing the intentional action side effect asymmetry." *Journal of Research in Personality*, 43: 18-24. DOI:10.1016/j.jrp.2008.10.007

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

Cova, Florian. 2014a. “Unconsidered intentional actions: An assessment of Scaife and Webber’s ‘Consideration Hypothesis’”. *Journal of Moral Philosophy*, 11: 57-79. DOI 10.1163/17455243-4681013

Cova, Florian. 2014b. “Can the Knobe Effect be explained away?” Unpublished manuscript, University of Geneva.

Cova, Florian, and Hichem Naar. 2012a. “Side-effect effect without side effects: the pervasive impact of moral considerations on judgments of intentionality.” *Philosophical Psychology*, 25: 837-854. DOI:10.1080/09515089.2011.622363

Cova, Florian, and Hichem Naar. 2012b. “Testing Sripada’s deep self model.” *Philosophical Psychology*, 25: 647-659. DOI:10.1080/09515089.2011.631996

Cova, Florian, Dupoux, Emmanuel, and Pierre Jacob. 2010. “Moral evaluation shapes linguistic report of others’ psychological states, not theory-of-mind judgments.” *Behavioral and Brain Sciences*, 33: 334. DOI: 10.1017/S0140525X10001718

Cova, Florian, Dupoux, Emmanuel, and Pierre Jacob. 2012. “On doing things intentionally.” *Mind & Language*, 27: 378-409. DOI: 10.1111/j.1468-0017.2012.01449.x

Cova, Florian, Dutant, Julien, Machery, Edouard, Knobe, Joshua, Nichols, Shaun, and Eddy Nahmias, eds. 2012. *La Philosophie Expérimentale*. Paris: Vuibert.

Cushman, Fiery, and Alfred Mele. 2008. “Intentional action: two and half folk concepts.” In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols, 170-184. New York: Oxford University Press, NY.

Dalbauer, Nikolaus, and Andreas Hergovich. 2013. “Is what is worse more likely? – The probabilistic explanation of the side-effect effect.” *Review of Philosophy and Psychology*, 4: 639-657. DOI: 10.1007/s13164-013-0156-1

Egré, Paul. 2010. “Qualitative judgments, quantitative judgments, and norm-sensitivity.” *Behavioral and Brain Sciences*, 33: 335-336. DOI: 10.1017/S0140525X1000172X

Egré, Paul. 2013. “Intentional action and the semantics of gradable expressions (On the Knobe Effect).” In *Causation in Grammatical Structure*, edited by Bridget Copley and Fabienne Martin. Oxford: oxford University Press.

Egré, Paul, and Florian Cova. in press. “Moral asymmetries and the semantics of “many”.” *Semantics & Pragmatics*.

Falkenstien, Kate. 2013. “Explaining the effect of morality on intentionality of lucky actions: The role of underlying questions.” *Review of Philosophy and Psychology*, 4: 293-308. DOI: 10.1007/s13164-013-0135-6

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

Feltz, Adam, and Edward T. Cokely. 2007. “An anomaly in intentional action ascriptions: More evidence of folk diversity.” In *Proceedings of the 29th Annual Cognitive Science Society*, edited by D.S. McNamara and J.G. Trafton, 1748. Austin, TX: Cognitive Science Society.

Feltz, Adam, and Edward T. Cokely. 2011. “Individual Differences in Theory-of-Mind Judgments: Order Effects and Side Effects.” *Philosophical Psychology*, 24: 343-355. DOI:10.1080/09515089.2011.556611

Guglielmo, Steve, and Bertram F. Malle. 2010a. “Can unintended side effects be intentional? Resolving a controversy over intentionality and morality.” *Personality and Social Psychology Bulletin*, 36: 1635-1647. DOI: 10.1177/0146167210386733

Guglielmo, Steve, and Bertram F. Malle. 2010b. “Enough skill to kill: Intentionality judgments and the moral valence of action.” *Cognition*, 117: 139-150. DOI: 10.1016/j.cognition.2010.08.002

Hindriks, Frank. 2008. “Intentional action and the praise-blame asymmetry.” *The Philosophical Quarterly*, 58: 630-641. DOI: 10.1111/j.1467-9213.2007.551.x

Hindriks, Frank. 2010. “Person as lawyer: How having a guilty mind explains attributions of intentional agency.” *Behavioral and Brain Sciences*, 33: 339-340. DOI: 10.1017/S0140525X10001767

Hindriks, Frank. 2011. “Control, intentional action, and moral responsibility.” *Philosophical Psychology*, 24: 787-801. DOI:10.1080/09515089.2011.562647

Hindriks, Frank. 2014. “Normativity in action: How to explain the Knobe Effect and its relatives.” *Mind & Language*, 29: 51-72. DOI: 10.1111/mila.12041

Hitchcock, Christopher, and Joshua Knobe. 2009. “Cause and norm”. *Journal of Philosophy*, 11: 587-612.

Holton, Richard. 2010. “Norms and the Knobe effect.” *Analysis*, 70: 417-424. DOI: 10.1093/analys/anq037

Knobe, Joshua. 2003a. “Intentional action and side-effects in ordinary language.” *Analysis*, 63: 190-193. DOI: 10.1111/1467-8284.00419

Knobe, Joshua. 2003b. “Intentional action in folk psychology: an experimental investigation.” *Philosophical Psychology*, 16: 309-324. DOI: 10.1080/09515080307771

Knobe, Joshua. 2004a. “Intention, intentional action and moral considerations.” *Analysis*, 64: 181-187. DOI: 10.1111/j.1467-8284.2004.00481.x

Knobe, Joshua. 2004b. “Folk psychology and folk morality: response to critics.” *Journal of Theoretical and Philosophical Psychology*, 24, 270-279. DOI: 10.1037/h0091248

Knobe, Joshua. 2006. “The concept of intentional action: a case study in the uses of folk psychology”. *Philosophical Studies*, 130: 203-231. DOI: 10.1007/s11098-004-4510-0

Knobe, Joshua. 2007. “Reason explanation in folk psychology.” *Midwest Studies in Philosophy*, 31: 90-107. DOI: 10.1111/j.1475-4975.2007.00146.x

Knobe, Joshua. 2010a. “Person as scientist, person as moralist.” *Behavioral and Brain Sciences*, 33: 315-329. DOI: 10.1017/S0140525X10000907

Knobe, Joshua. 2010b. “Action trees and moral judgment.” *Topic in Cognitive Science*, 2: 555-578. DOI: 10.1111/j.1756-8765.2010.01093.x

Knobe, Joshua, and Arudra Burra. 2006. “Intention and intentional action: a cross-cultural study.” *Journal of Culture and Cognition*, 1-2: 113-132. DOI: 10.1163/156853706776931222

Knobe, Joshua, and Gabriel S. Mendlow. 2004. “The good, the bad and the blameworthy: understanding the role of evaluative reasoning in folk psychology.” *Journal of Theoretical and Philosophical Psychology*, 24: 252-258. DOI: 10.1037/h0091246

Lanteri, Alessandro. 2009. “Judgments of intentionality and moral worth: experimental challenges to Hindriks.” *The Philosophical Quarterly*, 59: 73-720. DOI: 10.1111/j.1467-9213.2009.626.x

Lanteri, Alessandro. 2013. “Three-and-a-half folk concepts of intentional action.” *Philosophical Studies*, 158: 17-30. DOI: 10.1007/s11098-010-9664-3

Leslie, Alan M., Knobe, Joshua, and Adam Cohen. 2006. “Acting intentionally and the side-effect effect: ‘Theory of mind’ and moral judgment.” *Psychological Science*, 17: 421-427. DOI: 10.1111/j.1467-9280.2006.01722.x

Machery, Edouard. 2008. “The folk concept of intentional action: philosophical and experimental issues.” *Mind & Language*, 23: 165-189. DOI: 10.1111/j.1468-0017.2007.00336.x

Malle, Bertram F. 2006. “Intentionality, morality, and their relationship in human judgment.” *Journal of Cognition and Culture*, 6: 87-112. DOI: 10.1163/156853706776931358

Malle, Bertram F., and Joshua Knobe, J. 1997. “The folk concept of intentionality.” *Journal of Experimental Social Psychology*, 33: 101-121. DOI: “The folk concept of intentionality.”

Malle, Bertram F., and Joshua Knobe. 2001. “The distinction between desire and intention: A folk-conceptual analysis.” In *Intentions and Intentionality: Foundations of Social Cognition*, edited by Betram F. Malle, Louis J. Moses, and Dare A. Baldwin. Cambridge, MA: MIT Press.

Malle, Bertram F., and Sarah E. Nelson. 2003. “Judging *mens rea*: the tension between folk concepts and legal concepts of intentionality.” *Behavioral Sciences and the Law*, 21: 563-580. DOI: 10.1002/bsl.554

Mallon, Ron. 2008. “Knobe versus Machery: testing the trade-off hypothesis.” *Mind & Language*, 23: 247-255. DOI: 10.1111/j.1468-0017.2007.00339.x

Mandelbaum, Eric, and David Ripley. 2010. “Expectations and morality: A dilemma”. *Behavioral and Brain Sciences*, 33: 346. DOI: 10.1017/S0140525X10001822

Mandelbaum, Eric, and David Ripley. 2012. “Explaining the abstract/concrete paradoxes in moral psychology: the NBAR hypothesis.” *Review of Philosophy and Psychology*, 3: 351-368. DOI: 10.1007/s13164-012-0106-3

Mele, Alfred, and Fiery Cushman. 2007. “Intentional action, folk judgments and stories: sorting things out.” *Midwest Studies in Philosophy*, 31, 1, 184-201. DOI: 10.1111/j.1475-4975.2007.00147.x

Nadelhoffer, Thomas. 2004a. “The Butler problem revisited.” *Analysis*, 64: 277-284. DOI: 10.1111/j.0003-2638.2004.00497.x

Nadelhoffer, Thomas. 2004b. “Praise, side effects and intentional action.” *Journal of Theoretical and Philosophical Psychology*, 24: 196-213. DOI: 10.1163/156853706776931222

Nadelhoffer, Thomas. 2004c. “Blame, badness and intentional action: a reply to Knobe and Mendlow.” *Journal of Theoretical and Philosophical Psychology*, 24: 259-269. DOI: 10.1037/h0091247

Nadelhoffer, Thomas. 2005. “Skill, luck, control and intentional action.” *Philosophical Psychology*, 18: 343-354. DOI: 10.1080/09515080500177309

Nadelhoffer, Thomas. 2006a. “On trying to save the simple view.” *Mind & Language*, 21: 565-586. DOI: 10.1111/j.1468-0017.2006.00292.x

Nadelhoffer, Thomas. 2006b. “Foresight, moral considerations and intentional actions.” *Journal of Cognition and Culture*, 6: 133-158. DOI: 10.1111/j.1468-0017.2006.00292.x

Nadelhoffer, Thomas. 2006c. “Bad acts, blameworthy agents and intentional actions: some problems for juror impartiality.” *Philosophical Explorations*, 9: 203-220. DOI: 10.1080/13869790600641905

Nanay, Bence. 2010. “Morality of modality? What does the attribution of intentionality depend on?” *Canadian Journal of Philosophy*, 40: 25-39. DOI:10.1353/cjp.0.0087

Nichols, Shaun, and Joseph Ulatowski. 2007. “Intuitions and individual differences: the Knobe effect revisited.” *Mind & Language*, 22: 346-365. DOI: 10.1111/j.1468-0017.2007.00312.x

Pellizzoni, Sandra, Siegal, Michael, and Luca Surian. 2009. “Foreknowledge, caring and the side-effect effect in young children.” *Developmental Psychology*, 45: 289-295. DOI: 10.1037/a0014165

Pettit, Dean, and Joshua Knobe. 2009. “The pervasive impact of moral judgment.” *Mind & Language*, 24: 586-604. DOI: 10.1111/j.1468-0017.2009.01375.x

Phelan, Mark, and Hagop Sarkissian. 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. DOI: 10.1007/s11098-006-9047-y

Phelan, Mark, and Hagop, Sarkissian. 2009. “Is the trade-off hypothesis worth trading for?” *Mind & Language*, 24: 164-180. DOI: 10.1111/j.1468-0017.2008.01358.x

Pinillos, N. Angel, Smith, Nick, Nair, Shyam, Marchetto, Peter, and Cecilea Mun. 2011. “Philosophy’s new challenge: experiments and intentional action.” *Mind & Language* 26: 115-139. DOI: 10.1111/j.1468-0017.2010.01412.x

Robinson, Brian, Stey, Paul, and Mark Alfano. in press. “Reversing the side-effect effect: the power of salient norms.” *Philosophical Studies*. DOI: 10.1007/s11098-014-0283-2

Rose, David, Livengood, Jonathan, Sytsma, Justin, and Edouard Machery. 2012. “Deep trouble for the deep self.” *Philosophical Psychology*, 25: 629-646. DOI:10.1080/09515089.2011.622438

Sauer, Hanno. 2014. “It’s the Knobe Effect, Stupid!” *Review of Philosophy and Psychology*, 5: 485-503. DOI: 10.1007/s13164-014-0189-0

Sauer, Hanno, and Tom Bates. 2013. “Chairmen, cocaine, and car crashes: the Knobe Effect as an attribution error.” *Journal of Ethics*, 17: 305-330. DOI: 10.1007/s10892-013-9150-1

Scaife, Robin, and Johnathan Webber. 2013. Intentional side-effects of action. *Journal of Moral Philosophy*, 10: 179-203. DOI: 10.1163/17455243-4681004

Shepard, Jason, and Phillip Wolff. 2013. “Intentionality, evaluative judgments, and causal structure.” In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, edited by M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, 3390-3395. Berlin: Cognitive Science Society.

Shepherd, Joshua. 2012. “Action, attitude, and the Knobe Effect: another asymmetry.” *Review of Philosophy and Psychology*, 3: 171-185. DOI: 10.1007/s13164-011-0079-7

Sousa, Paulo, and Colin Holbrook. 2010. “Folk concepts of intentional action in the contexts of amoral and immoral luck.” *Review of Philosophy and Psychology*, 1: 351-370. DOI: 10.1007/s13164-010-0028-x

Sripada, Chandra S. 2010. “The Deep Self Model and asymmetries in folk judgments about intentional action.” *Philosophical Studies*, 151: 159-176. DOI: 10.1007/s11098-009-9423-5

Sripada, Chandra S. 2012. “Mental state attribution and the side-effect effect.” *Journal of Experimental Social Psychology*, 48: 232-238. DOI: “Mental states attribution and the side-effect effect.”

Sripada, Chandra S., and Sara Konrath. 2011. “Telling more than we can know about intentional action.” *Mind and Language*, 26: 353-380. DOI: 10.1111/j.1468-0017.2011.01421.x

Sverdlik, Steven. 2004. “Intentionality and moral judgments in commonsense thoughts about action.” *Journal of Theoretical and Philosophical Psychology*, 24: 224-236. DOI: 10.1037/h0091244

Turner, Jason. 2004. “Folk intuitions, asymmetry and intentional side effects.” *Journal of Theoretical and Philosophical Psychology*, 24: 214-219. DOI: 10.1037/h0091242

Uttich, Kevin, and Tania Lombrozo. 2010. “Norms inform mental state ascriptions: a rational explanation for the side-effect effect.” *Cognition*, 116: 87-100. DOI: 10.1016/j.cognition.2010.04.003

Wible, Andy. 2009. “Knobe, side effects, and the morally good business.” *Journal of Business Ethics*, 85: 173-178. DOI: 10.1007/s10551-008-9936-4

Woolfolk, Robert L. 2013. “Experimental philosophy: A methodological critique.” *Metaphilosophy*, 44: 79-87. DOI: 10.1111/meta.12016

Wright, Jennifer C., and John Bengson. 2009. "Asymmetries in folk judgments of responsibility and intentional action." *Mind & Language*, 24: 237-251. DOI: 10.1111/j.1468-0017.2008.01352.x

Young, Liane, Cushman, Fiery, Adolphs, Ralph, Tranel, Daniel, and Marc Hauser, M. 2006. "Does emotion mediate the effect of an action's moral status on its intentional status?" *Journal of Cognition and Culture*, 1-2: 291-304. DOI: 10.1163/156853706776931312

Zalla, Tiziana, and Marion Leboyer. 2011. "Judgments of intentionality and moral evaluations in individuals with high functioning autism." *Review of Philosophy and Psychology*, 2: 681-698. DOI: 10.1007/s13164-011-0048-1

Figures

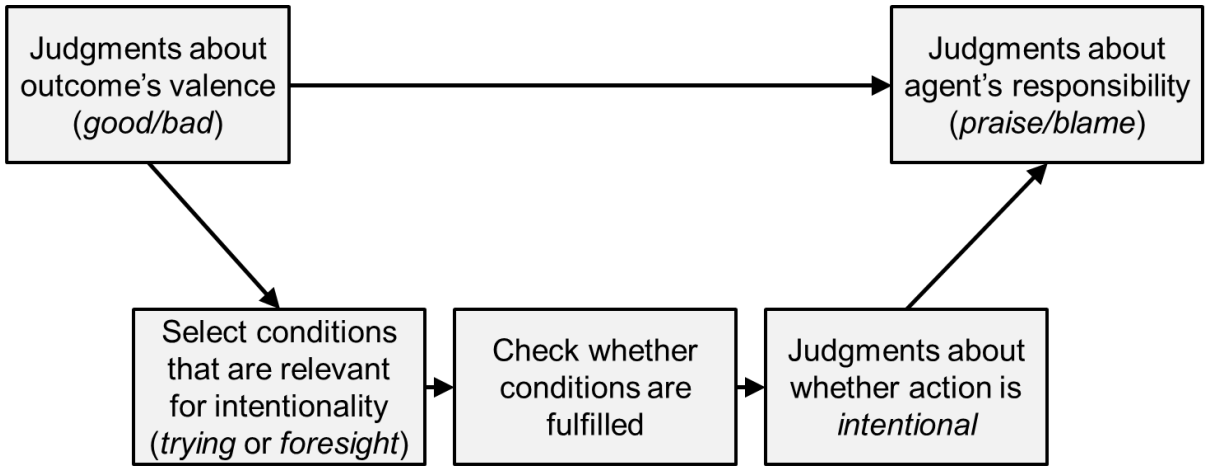

Figure 1. Knobe's original account (after Knobe, 2006). Judgments of praise and blame are, as is widely assumed, the results of both the outcome's evaluation (is it good or bad?) and ascriptions of intentionality (did the agent bring about the outcome intentionally?) However, what conditions need to be fulfilled for the action to count as intentional is itself determined by judgments about the outcome's valence.

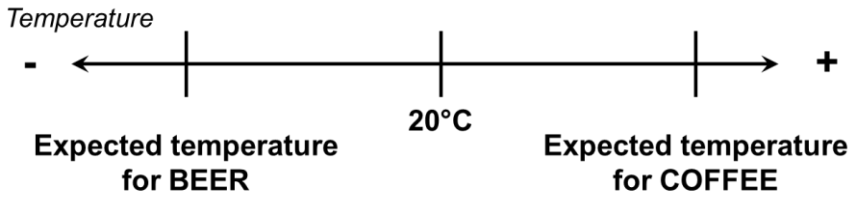

Figure 2. Default points for the 'coldness' of beverages according to Pettit and Knobe (2009). As one can see, whether a beverage counts as cold is not function only of its actual temperature (here 20°C), but also of whether this temperature is situated beyond or above the expected temperature for each beverage. Thus, a 20°C coffee counts as cold, while a 20°C beer does not, while both have the same temperature.

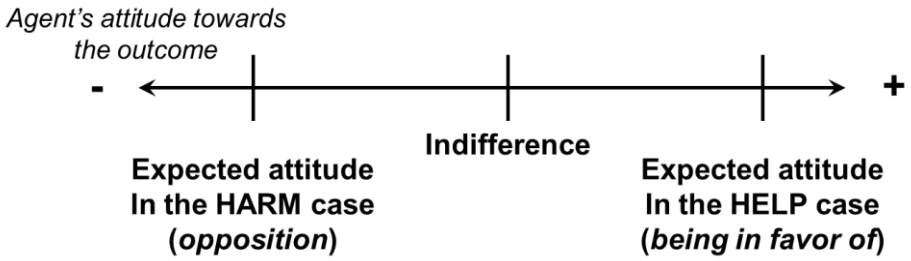

Figure 3. Default points for the agent's attitudes towards the outcome in the *Harm* and *Help* cases according to Pettit and Knobe (2009). The expected attitude is not the same in both cases: in the *Harm* cases, the expected attitude is below indifference ("being opposed to") while it is beyond in the *Help* case ("being in favor of"). Thus, a chairman holding the very same attitude ("indifference") does not reach the relevant default point for intentionality in the *Help* case, but more than fulfill it in the *Harm* case.

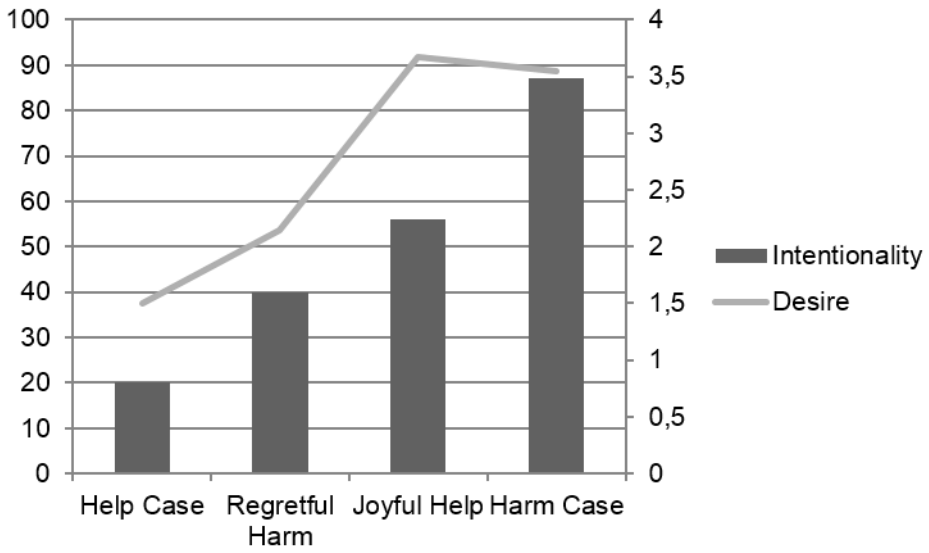

Figure 4. Percentages of participants judging the side-effect intentional (dark bars) and desire ratings on a scale from 0 to 6 (grey curve) in Guglielmo and Malle (2010) for the original *Harm* and *Help* cases and for the modified cases ("Regretful Harm" and "Joyful Help"). Intentionality ratings seem to covary with desire ratings when one look only for the three cases on the left. However, one can also observe that desire ratings are equal for the "Harm" and "Joyful Help" case, while intentionality ratings are much higher in the second than the first. This suggests that the same attitude towards the outcome generates higher intentionality ratings when the outcome is a bad one, and that ascriptions of intentionality are not fully explained by the agent's attitude towards the outcome.

[i] For empirical evidence on the distinction between *intentions* and *desires*, see Malle and Knobe (2001).
[ii] For similar phenomena, see Nadelhoffer (2004a, 2005, 2006a, 2006b).
[iii] For previous surveys, see Feltz (2007), Cova (2010) and Cova, Dupoux & Jacob (2012).
[iv] This 'goodness' and 'badness' do not have to be specifically moral, as shown by the fact that there can be instances of the Knobe Effect involving aesthetic values (Knobe, 2004b).
[v] For similar cases featuring regretful agents, see Sverdlik (2004), Phelan and Sarkissian (2008) (the *City Planner* case), Lanteri (2009) (the *Lever* case) and Shepherd (2012).
[vi] For another example of an asymmetry that does not seem to involve norms, see the pair of *Apple Tree* cases in Nanay (2010) and Cova and Naar (2012b).
[vii] For a similar attempt at 'explaining away' the Skill Effect, see Guglielmo and Malle (2010b).

[viii] For critical discussion, see Rose et al. (2012) and Cova and Naar (2012b).

[ix] To defend his hypothesis, Machery relies on two often used and much discussed cases: the *Free Cup* and *Extra Dollar* cases. For criticism of Machery's hypothesis, see Mallon (2008) and Phelan and Sarkissian (2009). For an account very similar to Machery's but using the notion of 'obstacle' in place of the notion of 'trade-off', see Sauer (2014).

[x] A related but slightly different account is proposed by Nanay (2010). According to Nanay, we judge a foreseen side effect to be intentional if the following modal claim is true: if the agent had not ignored considerations about the foreseen side effect, her action might have been different (other things being equal). However, it is not clear how this account would explain participants' answers in both *Pond* cases and in the *Terrorist* case.

[xi] For a detailed defense of this claim, see Cova, Dupoux & Jacob (2012).

[xii] Does this make "intentionally" itself a gradable term? Not necessarily. To my knowledge, there is no empirical evidence allowing us to conclude that people do or do treat "intentionally" as a gradable term.

[xiii] There is no published study investigating the Skill Effect for other psychological predicates. However, I have collected preliminary evidence that psychological predicates such as "desire", "intend" and "believe" are not subject to the Skill Effect. "Know", though, seems to display a pattern similar to the Skill Effect, but (i) the effect is much smaller and (ii) only holds for bad outcome, and does not exist in the case of good outcomes, which suggests that we are dealing with a different phenomenon.

[xiv] Additional advantages of this account include its ability to explain how ascriptions of intentionality can be manipulated by phrasing questions differently (Malle, 2006; Falkenstien, 2013) and why there seems to be so many individual differences on judgments of intentionality (Feltz & Cokely, 2007, 2011; Cokely & Feltz, 2009).