Word count: 10200 (including footnotes and references)

# The Folk Concept of Intentional Action: Philosophical and Experimental Issues*

EDOUARD MACHERY

---

**Abstract:** Recent experimental findings by Knobe and others (Knobe, 2003a; Nadelhoffer, 2006b; Nichols and Ulatowski, forthcoming) have been at the center of a controversy about the nature of the folk concept of intentional action. I argue that the significance of these findings has been overstated. My discussion is two-pronged. First, I contend that barring a consensual theory of conceptual competence, the significance of these experimental findings for the nature of the concept of intentional action cannot be

---

* Thanks for helpful comments to Gregory Currie, Josh Knobe, Ron Mallon, Thomas Nadelhoffer, Shaun Nichols, Steve Stich, Liane Young, the readers of the blog Experimental Philosophy (http://experimentalphilosophy.typepad.com/) as well as two anonymous reviewers. Thanks also to my research assistant on this project, Julie Sokolow, for her help and her comments.

**Address for correspondence:** Edouard Machery, Department of History and Philosophy of Science, University of Pittsburgh, 1017CL, Pittsburgh, PA, 15260, USA.

**Email:** machery@pitt.edu

determined. Unfortunately, the lack of progress in the philosophy of concepts casts doubt on whether such a consensual theory will be found. Second, I propose a new, deflationary interpretation of these experimental findings, 'the trade-off hypothesis,' and I present several new experimental findings that support this interpretation.

Consider the following scenario. A CEO decides to start a new program, because this program will increase the profits of her company. She foresees that this program will harm the environment as a side-effect. People tend to judge that the CEO *intentionally* harmed the environment. Now, consider a second scenario. A CEO decides to start a new program, because this program will increase the profits of her company. She foresees that this program will help the environment as a side-effect. People tend to judge that the CEO did *not intentionally* help the environment. Using these and other stories, Joshua Knobe found that from an early age on (Leslie et al., 2006) and in several cultures (Knobe and Burra, 2006), people tend to judge that agents intentionally bring about foreseen, blameworthy side-effects, but that agents do not intentionally bring about foreseen, praiseworthy side-effects.[1] I call this effect '*the Knobe effect*.'

The Knobe effect has been at the center of a controversy about the nature of the folk concept of intentional action. Some philosophers, such as Knobe and, more recently, Nichols, argue that these findings bring to light some important properties of this folk

[1] Knobe, 2003a, 2003b, 2006; Knobe and Mendlow, 2004; Mele, 2003; Nichols and Ulatowski, forthcoming.

concept. Other philosophers demur.[2] Typically, these skeptics argue that the Knobe effect might say something about how the folk concept of intentional action is used in specific circumstances. However, they insist that the Knobe effect says little about what is constitutive of people's grasp of this concept. To use a terminology explained later on in this article, according to these skeptics, the Knobe effect says precious little about our *conceptual competence* with the concept of intentional action.

In this article, I argue that the significance of the Knobe effect has been probably overstated. My discussion has two parts. First, I argue that a stumbling block stands in the way of settling the philosophical debate about the implications of the Knobe effect for understanding the nature of the folk concept of intentional action. The philosophical debates about the Knobe effect suppose a distinction between performance and competence with a concept, but the lack of progress in the philosophy of concepts casts doubt on whether such a distinction will be made out. Second, I argue that contrary to the consensus among philosophers and psychologists, the Knobe effect probably says little about our moral psychology. Rather, it results from the fact that people take the costs that are incurred in order to reap some benefits to be intentionally incurred. I call this new interpretation of the Knobe effect 'the trade-off hypothesis.' I present some new experimental evidence in support of the trade-off hypothesis.

Here is how I will proceed. In the first section, I review Knobe's empirical findings as well as his interpretation of these findings. In the second section, I develop the first part of my critique. I argue that because there is no consensual theory of conceptual competence, the philosophical debate about the significance of the Knobe

[2] Adams and Steadman, 2004a, 2004b; Malle, 2006; Nadelhoffer, 2006 a, b.

effect for the folk concept of intentional action cannot be resolved. In the third section, I turn to the second part of my critique. I present some experimental evidence that supports the trade-off hypothesis and I argue that if the trade-off hypothesis is correct, then the Knobe effect fails to say anything about our moral psychology.

## 1. The Knobe Effect

### 1.1 The Experimental Findings

Knobe presented people with pairs of stories (or 'probes'). Within each pair, the probes are assumed to be identical, save for one element. The probes describe a decision made by an agent. The agent is aware that her decision will have a side-effect. The nature of the side-effect distinguishes the probes within each pair. For instance, the side-effect in the first probe in a given pair might be morally wrong, while the side-effect in the second probe might be morally desirable. Consider, particularly, 'the harm case' and 'the help case' (Knobe, 2003a).

> **The 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. Did the chairman intentionally harm the environment? YES / NO

**The help 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, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped. Did the chairman intentionally help the environment? YES / NO.

Both probes are identical, save for one element—the nature of the side effect. Particularly, harming the environment is morally wrong and blameworthy, while helping the environment is morally right and praiseworthy. In the harm case, people tend to judge that the vice-president of the company intentionally harmed the environment. On the contrary, in the help case, people tend to judge that she did not intentionally help the environment (Table 1).

Put table 1 about here.

The asymmetry between the two cases is extremely robust, showing up with different probes, in different cultures and at different ages.

### 1.2 Knobe's Interpretation

This result is surprising. In both probes within a pair, the side-effect is foreseen by the agent. The only difference between the two probes is assumed to lie in the nature of the side-effect, for instance, whether the side-effect is morally right or morally wrong or whether one side-effect is blameworthy while the other is praiseworthy. The puzzle is to understand why the nature of a side-effect, for example its moral value or its blameworthiness, matters for people when they are asked to decide whether this side-effect has been intentionally brought about.

Knobe has an explanation. He writes (2006, 225-226):

> 'We are now in a position to offer a new hypothesis about the role of moral considerations in people's concept of intentional action. The key claim will be that people's intentional action intuitions tend to track the psychological features that are most relevant to praise and blame judgments. But — and this is where moral considerations come in — different psychological features will be relevant depending on whether the behavior itself is good or bad. That is to say, we use different psychological features when we are (a) trying to determine whether or not an agent deserves blame for her bad behaviors from the ones we use when we are (b) trying to determine whether or not an agent deserves praise for her good behaviors.'

According to Knobe, the folk concept of intentional action plays an important role in blame and praise, including moral blame and moral praise (Knobe, 2006).[3] We blame and praise people, depending on their intentional actions. Because the actions that are

[3] While Knobe initially emphasized *moral* praise and blame, he now emphasizes blame and praise *in general* (Knobe, personal communication, May 2006).

blameworthy differ from the actions that are praiseworthy, the properties that matter for classifying blameworthy actions as intentional differ from the properties that matter for classifying praiseworthy actions as intentional. One way to capture Knobe's hypothesis is to propose that people follow the following categorization procedure when they decide whether a side-effect is intentional (Figure 1).

Put Figure 1 about here.

The hypothesis that people follow this categorization procedure explains the asymmetry found between the harm case and the help case. As I'll say for the sake of simplicity, blame and praise have shaped the folk concept of intentional action. Knobe (2006) takes the asymmetry between the judgments elicited by the harm and help cases to be tentative evidence that the function of the theory of mind is not merely to predict and explain behavior, but also to enable us to make moral judgments.

Although Knobe explains the asymmetry between the two probes within a pair by reference to the role of the concept of intentional action for blaming and praising, he is not committed to the claim that this asymmetry will show up *only* in cases where the two side-effects differ with respect to their moral value. Neither is he committed to the claim that this asymmetry will show up *only* in cases where the two side-effects differ with respect to their blameworthiness or praiseworthiness (Knobe and Mendlow, 2004). Rather, Knobe contends that the asymmetry is to be found in cases where one outcome is bad (including, but not exclusively, morally bad), while the other outcome is good (including, but not exclusively, morally good). Indeed, Knobe and Mendlow (2004) found that a probe that does not involve a morally wrong side-effect elicited the same judgments as the harm case. Subjects were presented with the following probe.

> Susan is the president of a major computer corporation. One day, her assistant comes to her and says, 'We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts but decreasing sales in New Jersey.' Susan thinks, 'According to my calculations, the losses we sustain in New Jersey should be a little bit smaller than the gains we make in Massachusetts. I guess the best course of action would be to approve the program.' 'All right,' she says. 'Let's implement the program. So we'll be increasing sales in Massachusetts and decreasing sales in New Jersey.'

75% of the subjects answered that the side-effect was intentional. Knobe and Mendlow explain that subjects make this judgment because the side effect decreasing sales in New Jersey 'is in some sense a bad one.' Since Knobe and Mendlow (2004) first found that a pair of non-moral cases could elicit the asymmetry between the judgments about the intentionality of side-effects, evidence has accumulated that this asymmetry is not found exclusively in cases involving morally relevant actions. Particularly, Phelan and Sarkissian (forthcoming) replicated the Knobe effect with other pairs of non-moral cases.[4] This fact ought to be kept in mind while evaluating Knobe's views. If the asymmetry between the judgments about the intentionality of side-effects can be found in pairs of cases that have nothing to do either with morality or with blame and praise, why would we believe that this asymmetry depends on the role of our concept of intentional action in folk morality or in ascribing blame and praise? I revisit this issue in section 3.

## 2. What is our Conceptual Competence with the Concept of Intentional Action?

[4] See also Wright and Bengson (ms) and the free-cup and extra-dollar cases in section 3.

### 2.1 Alternative Interpretations of the Knobe Effect

Knobe argues that his and others' experimental findings cast some light on the nature of the folk concept of intentional action. These findings show that this concept has somehow been shaped by its role in blaming and praising. Nichols and Ulatowski (forthcoming) and Mele (2003) concur.

The claim that the experimental findings under consideration cast some light upon the folk concept of intentionality has recently been under intense attack. The critics of this claim contend that these experimental findings might cast some light on how the concept of intentional action is used, but not on what is *constitutive* of people's grasp of the concept of intentional action. Two main types of critique can be distinguished.

Adams and Steadman (2004a, 2004b) have argued that the asymmetry found between the two probes in a given pair, for instance between the harm case and the help case, is a pragmatic phenomenon.[5] In substance, they contend that people's intuitions are derived from the conversational implicatures implied by the two possible answers for each probe. Consider the harm case. People might feel that if they were to answer that the chairman did not intentionally harm the environment, they would conversationally imply that the chairman is not to be blamed for her choice. Since they want to blame the chairman, they answer that she did intentionally harm the environment. Consider now the help case. People might feel that if they were to answer that the chairman did intentionally help the environment, they would conversationally imply that the chairman is to be praised for her choice. Since they do not want to praise the chairman, they answer that she did not intentionally help the environment. If this pragmatic interpretation were

[5] For discussion, see Knobe, 2004, 2006; Nichols and Ulatowski, forthcoming.

correct, then, according to Adams and Steadman, the Knobe effect would say nothing about what is constitutive of grasping the concept of intentional action. Rather, what would explain subjects' answers are their beliefs about how other people would interpret their assertions if they were to assert that an action, for instance, harming or helping the environment, was done intentionally.

Nadelhoffer (2006a, b) has proposed a second kind of explanation of the experimental findings under consideration.[6] He contends that the harm case (and similar cases) triggers some emotion that prevents the correct application of the concept of intentional action. He writes that 'affective or emotional responses … *inappropriately* bias our otherwise rational judgments' (2006a, 214). That is, if the concept of intentional action were appropriately applied, people would have the same intuitions for the two cases within a pair. For both cases, people would judge that the side-effect has not been intentionally brought about. For instance, they would say that the chairman has neither intentionally harmed the environment nor intentionally helped the environment. If this interpretation of the empirical findings were correct, then, according to Nadelhoffer, the Knobe effect would say nothing about what is constitutive of grasping the concept of intentional action, because people misapply this concept in one of the two probes within a pair.

### 2.2 Conceptual Competence vs. Conceptual Performance

[6] For a related idea, see Malle, 2006; for discussion, see Knobe and Mendlow, 2006; Young et al., 2006; Nichols and Ulatowski, forthcoming.

There are two dimensions in this philosophical controversy. First, Knobe, Nichols, Adams, Nadelhoffer and others disagree about why people make asymmetric judgments about the intentional status of side-effects in the harm case and the help case (and similar cases). That is, they disagree about the psychological events that underlie people's judgments in the harm case and in the help case (and similar cases). For instance, Nadelhoffer proposes, while Knobe or Nichols deny, that people judge that harming the environment is intentional because they experience a given emotion. Or Adams proposes, while Knobe or Nichols deny, that people judge that harming the environment is intentional because they want to avoid a conversational implicature.

Second, Knobe, Nichols, Adams, Nadelhoffer and others disagree about whether the asymmetry between people's judgments about the intentional status of the side-effect in the harm case and in the help case casts any light on people's *conceptual competence* with the concept of intentional action. Adams and Steadman, Malle, and Nadelhoffer doubt that it is the case. They believe that the asymmetric use of this concept in the probes within a pair, for example in the harm and help cases, results from factors beyond what is constitutive of possessing this concept, such as negative emotions or our desire to avoid unwanted conversational implicatures. The asymmetric use of this concept in these probes is merely an aspect of *our conceptual performance*. Knobe and Nichols and Ulatowski, on the contrary, argue that these findings cast some light on people's conceptual competence with this concept.

The distinction between conceptual competence and conceptual performance derives from Chomsky's distinction between linguistic competence and linguistic performance (Chomsky, 1965). As Chomsky puts it (1965, p. 3):

> ‘Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its (the speech community's) language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance.’

The generative syntactician uses people’s intuitions about the grammaticality of sentences to determine people’s implicit knowledge of the grammar of the natural language they speak. Grammatical intuitions, however, are not supposed to be a direct reflection of this implicit knowledge. Rather, grammatical intuitions are supposed to result from this implicit knowledge together with non-linguistic factors, such as attention, memory, fatigue, and so on (see Figure 2).

Put Figure 2 about here.

An individual’s linguistic competence is her implicit knowledge of her language, on the basis of which she is able to utter and understand an infinite number of sentences. Her linguistic performance is a specific event, that is, the production of sentences at a given time on a given occasion.

The distinction between competence and performance has been used in a few domains besides language. Following Cohen (1981), numerous epistemologists have used this distinction to distinguish between our rational capacity to judge and reason and our actual judgments and reasonings, including irrational judgments and invalid reasonings in experimental tasks. Dwyer (1999), Mikhail (2000), and Hauser et al. (forthcoming) have

also used this distinction to draw a distinction between our knowledge of moral principles and our moral judgments and decisions at a given time.

The debate between Knobe, Nichols, Adams, Nadelhoffer and others supposes a distinction between competence and performance applied to the possession of concepts. The conceptual competence with a given concept might be the knowledge one might have about the referent of this concept by virtue of having this concept or the inferences one is willing to draw by virtue of having this concept. By contrast, the conceptual performance with a given concept is a specific event, that is, the use of this concept at a given time on a given occasion. Conceptual performance with a given concept is typically affected by many factors, besides what is constitutive of our competence with this concept.

It is important to distinguish the two dimensions in the controversy at hand distinguished above. For, even if it turned out that one cannot decide whether or not the Knobe effect bears on our conceptual competence with the concept of intentional action (as will be argued below), one might still be able to decide whether any account of the psychological events that underlie the asymmetry between our judgments in the harm and help cases is correct. For instance, one might be able to decide whether or not a negative emotion causes people to make these asymmetric judgments, as proposed by Nadelhoffer, even though one might not be able to decide whether if Nadelhoffer's account were correct, one would be entitled to conclude that the Knobe effect does not cast any light on our conceptual competence with the concept of intentional action.

### 2.3 The Problem of Conceptual Competence

Does the Knobe effect bear merely on people's performance with the concept of intentional action or does it cast some light on people's conceptual competence with this concept? As we saw above, the debate is raging. I do not intend to take a stance in this debate. For, I argue in the remainder of this section that a stumbling block stands in the way of resolving this issue. To put it simply, the controversy can be resolved only if there is a principled distinction between what constitutes our competence with a given concept and what results merely from the multitude of factors that affect our use of this concept at a given time on a given occasion. Unfortunately, the literature on concepts has not converged and does not seem to be converging on such a principled distinction. Barring such a distinction, however, the controversy about whether the Knobe effect bears on our conceptual competence with the concept of intentional action cannot be properly resolved.[7]

Philosophers have proposed many accounts of conceptual competence. For our purposes, it is useful to briefly distinguish three accounts, which we might call 'the holistic account,' 'the molecularist account' and 'the atomist account.' Block's (1986) theory of concepts illustrates the holistic account. According to Block, *any* inference or judgment that involves a given concept is constitutive of the identity of this concept.

[7] Importantly, the argument proposed here is *not* that the asymmetry found by Knobe could be a mere performance error, rather than an aspect of the competence with the concept of intentional action. Rather, I argue that the philosophical debates about this asymmetry suppose a distinction between performance and competence with a concept, but that the lack of progress in the philosophy of concepts casts doubt on whether such a distinction will be found.

Thus, any inference or judgment is constitutive of what it is to possess this concept rather than another or, to put it differently, of the conceptual competence with this concept rather than with another. There is no distinction to be drawn between the inferences or judgments involving this concept that are constitutive of the conceptual competence with this concept and those that are not. Peacocke's (1992) theory of concepts illustrates the molecularist account. According to Peacocke, as a first approximation, *some* judgments or inferences involving the concept, but not other inferences or judgments, are constitutive of the identity of a concept. Thus, some specific judgments or inferences are constitutive of our conceptual competence with this concept. For instance, an individual possesses the concept SQUARE only if she is disposed to assent to a judgment that a seen square object is square when this object is presented visually with the right orientation in the right conditions and when she takes her experience at face-value (Peacocke, 1992, p. 74). Finally, Fodor's (1998) theory of concepts illustrates the atomist account of conceptual competence. Contrary to Block and Peacocke, Fodor proposes that *no* inference or judgment involving a concept is constitutive of the identity of this concept. To possess a concept is not a matter of how one uses it. Rather, one possesses a concept if one stands in a nomological relation with the referent of this concept.

This diversity of accounts of conceptual competence bears on whether the Knobe effect casts some light on our conceptual competence with the folk concept of intentional action. Suppose for a moment that Adams and Steadman (2004a, 2004b) are right, when they contend that the Knobe effect is to be explained as a pragmatic phenomenon. What follows with respect to our conceptual competence with the concept of intentional action? Well, it depends on which account of conceptual competence is correct. If the holistic

account of conceptual competence is correct, then the Knobe effect still bears on our conceptual competence with the concept of intentional action, in spite of merely resulting from pragmatic factors. For, according to the holistic account of conceptual competence, *any* inference that involves a concept is constitutive of the conceptual competence with this concept. Suppose, on the contrary, that the molecularist account of conceptual competence is correct. Then, if the Knobe effect is really to be explained in pragmatic terms, as Adams and Steadman would have it, one could argue that the Knobe effect is irrelevant for spelling out our conceptual competence with the concept of intentional action. For, according to the molecularist account of conceptual competence, only *some* inferences or judgments involving a concept are constitutive of the conceptual competence with this concept.

Now, suppose that as Knobe (2004b, 2006) and Nichols and Ulatowski (forthcoming) have argued, the Knobe effect is not to be explained in pragmatic terms. What follows with respect to our conceptual competence with the concept of intentional action? Well, again, it depends on which account of conceptual competence is correct. If the atomist account of conceptual competence is correct, the Knobe effect fails to cast any light on the conceptual competence with the concept of intentional action. For, according to the atomist account, conceptual competence does not depend on how a concept is used. How we use a concept is not part of what it means to have this concept. Suppose, on the contrary, that the molecularist account of conceptual competence is correct. Then, if the Knobe effect is not to be explained in pragmatic terms, it might plausibly cast some light on the nature of our conceptual competence with the concept of intentional action.

Thus, the debate about the significance of the Knobe effect for our conceptual competence with the folk concept of intentional action hangs on which account of conceptual competence is correct. Depending on which account is correct, the implications of specific interpretations of the Knobe effect, such as Adams and Steadman's or Nadelhoffer's account, differ. Unfortunately, it is entirely unclear which account of conceptual competence is correct. The literature on concepts has failed to decide between the three accounts discussed above. Moreover, it does not seem to be heading toward a consensus. Barring such a consensus, however, the philosophical debate about the implications of the Knobe effect for our conceptual competence with the concept of intentional action cannot be satisfyingly resolved.

### 2.4 An Objection

It might be thought that the problem raised in this section is easily circumvented. There might be an internal agreement between Knobe, Nichols, Adams, Nadelhoffer and others about what is constitutive of conceptual competence and what is merely a property of our conceptual performance. For instance, these philosophers might agree that when a judgment involving a concept, such as the concept of intentional action, is caused by some emotion, as has been proposed by Nadelhoffer, our performance is not the result of our conceptual competence.

This objection is problematic for three reasons. First, it is unclear whether Knobe, Nichols, Adams, Nadelhoffer and others agree on an account of conceptual competence. They have been silent on this issue. Second, it would be curious to propose that a mere internal agreement about conceptual competence is sufficient for understanding the

implications of the Knobe effect for the nature of the concept of intentional action. What philosophers really need is the *correct* account of conceptual competence (if there is such a thing). What if these philosophers agreed on the wrong account? Finally, suppose that Knobe, Nichols, Adams, Nadelhoffer and others agree on what seems to be the most congenial account for the debate at hand, namely the molecularist account of conceptual competence.[8] As we saw, according to this account, some inferences or judgments, but not others, are constitutive of our conceptual competence. For the sake of the argument, suppose also that this is the right account of conceptual competence. Unfortunately, it would still be unclear what the consequences of specific interpretations of the Knobe effect are for the nature of the folk concept of intentional action. For, we still would need an account of *which* inferences or judgments are constitutive of conceptual competence and which are not. From different accounts of which inferences or judgments are constitutive of conceptual competence, different implications for our conceptual competence with the concept of intentional action can be drawn from Adams' interpretation of the Knobe effect. The same is true of Nadelhoffer's or of Knobe's interpretation of this effect. However, as is well-known in the philosophical literature on concepts, an account of *which* judgments or inferences are constitutive of conceptual competence is lacking, except, maybe, for a few logical concepts.

Philosophers interested in the Knobe effect have debated at length about whether this effect casts any light on our conceptual competence with the concept of intentional

[8] The molecularist account seems to be the most congenial, because it naturally allows for a distinction between those inferences and judgments involving a concept that are constitutive of possessing this concept and those judgments and inferences that are not.

action. If the argument developed in this section is sound, this debate is misguided. This suggests that philosophers should shift focus toward understanding the psychological events that underlie the asymmetry between our judgments in the harm and help cases (and similar cases). This is the object of the next section.

## 3. A New Explanation of the Knobe Effect

### 3.1 Costs, Benefits and Intentional Action

I turn to the second part of my critique of the debate about the Knobe effect. Whether or not they believe that the Knobe effect casts light on our conceptual competence with the concept of intentional action, philosophers and psychologists agree that the Knobe effect has something to do with our moral psychology. Knobe argues that this effect is evidence that the function of our folk theory of mind is not merely to predict and explain behavior, but also to enable moral judgments. Critics, such as Nadelhoffer and Adams, believe that moral emotions or a desire to avoid conversational implicatures about blame explain the Knobe effect. I disagree. In this section, I argue that the Knobe effect probably does not tell us anything about *moral* psychology. Specifically, it does not provide evidence that the folk concept of intentional action and, a fortiori, the theory of mind have been shaped by their role in folk morality or in ascribing blame and praise.

An important cue that this may well be the case is that, as recognized by Knobe himself, the asymmetry between the judgments elicited by the two probes within a pair, for instance, between the judgments elicited by the harm case and the help case, is found in pairs of stories that involve neither a moral evaluation of the side-effect resulting from

the agent's decision nor an evaluation of the blameworthiness or praiseworthiness of this side-effect. This suggests that the explanation of the Knobe effect might not be essentially related to morality or to blame. A convincing explanation should account for the fact that an asymmetry between people's intuitions about the intentionality of bringing about a side-effect is to be found in moral *and* in non-moral cases. None of the explanations of the Knobe effect meets this constraint.

A plausible explanation is not hard to come by. Consider the harm case. The chairman desires to obtain something she judges to be beneficial—an increase in profits for her company. She foresees that obtaining this benefit will entail some cost—harming the environment. But, because the foreseen cost is offset by the foreseen benefit, the chairman decides to incur the foreseen cost—harming the environment—in order to reap the foreseen benefit—increasing the profits of the company. To put the same idea differently, the harm case describes a trade-off: An agent, i.e., the chairman, is willing to incur a cost in order to get a benefit. Consider now the help case. The chairman desires obtaining something she judges to be beneficial—an increase in profits for her company. She foresees that obtaining this benefit will bring about some other benefit—helping the environment. Because helping the environment is not a cost, the help case does *not* describe a situation where the chairman decides to incur a cost in order to reap a benefit. Or, to put it differently, the help case does not describe a trade-off between a cost and a benefit.

Now, suppose that people conceptualize the harm case (or similar cases) in the way just described. When people read the harm case, they conceptualize the side-effect *harming the environment* as a cost, that is, as something that is negatively valued and that

one must incur if one is to reap a greater benefit.[9] They think of this cost as being offset by the benefit *increasing the profits of the company*. That is, they conceptualize the harm case as involving a trade-off between a cost and a benefit. The help case cannot be conceptualized in this way. For, *helping the environment* cannot be plausibly thought of as a cost, since it is not negatively valued.[10] People are then asked whether the chairman intentionally harmed or intentionally helped the environment. Since we think of costs as being intentionally incurred in order to reap some foreseen benefits, people tend to give a positive answer to this question. Since people do not conceptualize helping the environment as a cost, they answer that the chairman did not intentionally help the environment. I call this explanation '*the trade-off hypothesis*.' Importantly, in keeping with the fact that the Knobe effect is found in moral and in non-moral cases and contrary to previous explanations, the trade-off hypothesis does not hang on the side-effect being morally relevant.

---

[9] Costs include means, that is, those actions or events that bring about desired goals. They also include side-effects of goals. Side-effects are not means, because they do not bring about the goals, but merely result from these goals.

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

Compare the trade-off hypothesis with Knobe's explanation of the asymmetry between our judgments elicited by the two cases within a pair. According to Knobe, this asymmetry results from the role of the concept of intentional action in ascribing blame and praise. Because of this role, when a foreseen side-effect is judged to be bad, it is judged to be intentional; when it is judged to be good, it is judged to be unintentional (Figure 1). According to this view, when people read the harm case, they categorize harming the environment as a foreseen side-effect and as being bad. On the basis of these two categorizations, they judge that harming the environment is intentional. When people read the help case, they categorize helping the environment as a foreseen side-effect and as being good. On the basis of these two categorizations, they judge that helping the environment is unintentional. Figure 3 summarizes the reasoning that leads to the judgments in the harm and in the help cases according to Knobe.

Put Figure 3 about here.

By contrast, according to the trade-off hypothesis, the asymmetry between people's judgments elicited by the harm case and the help case (as well as similar cases) has nothing to do with morality or with blame. It merely results from the fact that people conceptualize one of the two probes, namely the harm case, as involving a trade-off between a foreseen cost and a foreseen benefit. Thus, when people read the harm case, they categorize harming the environment as a condition for getting a benefit and as being bad. On the basis of these two categorizations, they categorize harming the environment as being a cost. Because they believe that costs are intentionally incurred, they judge that harming the environment is intentional. When people read the help case, they categorize helping the environment as a side-effect and as being good. They cannot categorize it as

being a cost, because it is not negatively valued. On the basis of these two categorizations, they judge that helping the environment is unintentional. Figure 4 summarizes the reasoning that leads to the judgments in the harm and in the help cases according to the trade-off hypothesis.

Put Figure 4 about here.

Which of the two hypotheses under consideration is correct? The trade-off hypothesis, but not Knobe's hypothesis, accounts naturally for the fact that the Knobe effect is to be found in cases that have nothing to do with morality or with blame. Moreover, some new experimental evidence further supports the trade-off hypothesis. I developed a pair of probes, closely modeled on Knobe's harm and help cases. Most important, one of these two probes—the extra-dollar case—involves a clear trade-off between a cost and a benefit. In the extra-dollar case, the agent is confronted with a decision concerning whether to incur an extra cost (paying an extra-dollar) in order to reap a desired benefit (getting a smoothie). In the other case—the free-cup case—the agent is given a benefit (a free cup) in addition to the foreseen benefit that results from her decision (a smoothie). The prediction was that an asymmetry similar to the asymmetry found with the harm and help cases would be found with this pair of probes. Moreover, since I contend that the asymmetry has nothing to do with blame, the two probes were designed in such a way that the actions were neither blameworthy nor praiseworthy. Subjects were asked to evaluate the blameworthiness of the action chosen by the agent. The prediction is that subjects would find the two probes equally neutral.

The two probes are the following:

**The free-cup case**

Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest sized drink available. Before ordering, the cashier told him that if he bought a Mega-Sized Smoothie he would get it in a special commemorative cup. Joe replied, 'I don't care about a commemorative cup, I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie in a commemorative cup.

Did Joe intentionally obtain the commemorative cup?

YES NO

Was obtaining the commemorative cup blameworthy, praiseworthy, or neutral?

BLAMEWORTHY PRAISEWORTHY NEUTRAL

**The extra-dollar case**

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

Did Joe intentionally pay one dollar more?

YES NO

Was paying one dollar more blameworthy, praiseworthy, or neutral?

BLAMEWORTHY PRAISEWORTHY NEUTRAL

### 3.2 Experimental Evidence

126 undergraduate students from the University of Pittsburgh took part in the experiment. Subjects were asked to fill a short demographic questionnaire. This questionnaire was used to determine whether English was their native language. Eight subjects answered that they were not native speakers of English. Excluding these subjects did not affect the data analysis. For this reason, these subjects were not excluded from the sample.

In a classroom setting, subjects were randomly given one of the two probes, the extra-dollar probe or the free cup probe (see above). 62 subjects read the extra-dollar probe, 64 the free-cup probe. The scoring procedure was straightforward. The question about the intentional nature of the side-effect ('the intentionality question') was scored binomially. A negative answer was scored 0 and a positive answer was scored 1. The question about the blameworthiness of the action ('the value question') was scored as follows. The answer 'blameworthy' was scored 0, the answer 'neutral' was scored 1, and the answer 'praiseworthy' was scored 2. Percentages are presented in tables 2 and 3.

Put tables 2 and 3 about here.

A chi-square test yielded a highly significant difference between the two cases for the intentionality question ($\chi^2$ (1, $N$ = 126) = 37.2, $p < 0.001$) (Figure 5). As predicted by the trade-off hypothesis, subjects were significantly more likely to judge that the agent intentionally paid an extra dollar than to judge that the agent intentionally obtained a free cup. Importantly, this asymmetrical pattern of answers is analogous to the Knobe effect.

Put Figure 5 about here.

A chi-square test failed to yield any significant different between the two conditions for the value question ($\chi^2$ (2, $N$ = 126) = 3.2, $p$ > 0.1, n.s.) (Figure 6). As predicted, across conditions, subjects did not give significantly different answers to the question about the value of the action of the agent. In both conditions, subjects tended to judge that the side-effect was neutral (by contrast to blameworthy or praiseworthy).

Put Figure 6 about here.

### 3.3 Discussion

The hypothesis under consideration is that people make asymmetric judgments in the stories used by Knobe and others because they conceptualize the negative side-effect (e.g., harming the environment) as a cost that the agent incurs in order to reap a benefit (e.g., making profits). Because costs are intentionally incurred in order to reap a benefit, people judge that the foreseen, negative side-effect was intentionally brought about. To provide evidence for this hypothesis, I predicted that the asymmetry found by Knobe and others would occur in pairs of stories that contrast a clear cost-benefit relation (incurring a foreseen cost to reap a foreseen benefit) with a gain as a foreseen side-effect of a foreseen benefit. The findings summarized above confirm the trade-off hypothesis.

What is the significance of these findings and of the trade-off hypothesis? The findings reported here and the trade-off hypothesis suggest that *pace* Knobe, Mele and Nichols, the asymmetry found with the harm and help cases as well as with similar probes might have nothing to do with our folk morality and with ascribing blame and praise. This asymmetry is found in pairs of cases that have nothing to do with blame and praise, because trade-offs between foreseen costs and foreseen benefits are not found

only in stories involving blameworthy or praiseworthy choices. Moreover, *pace* Nadelhoffer, Malle, and Adams, this asymmetry might have nothing to do with avoiding conversational implicatures about blame or with negative emotions. The asymmetry is found in pairs of cases that describe actions judged by people to be neither blameworthy nor praiseworthy, but neutral. This is unsurprising, because trade-offs between foreseen costs and foreseen benefits are not found only in stories involving blameworthy or praiseworthy choices or in stories involving emotionally salient choices.

### 3.4 First Objection

To defend Knobe's theory against the trade-off hypothesis, one could deny that the pair consisting of the extra-dollar and the free-cup cases is really appropriate to support the trade-off hypothesis. For, one might argue, the extra-dollar case does not involve a foreseen side-effect of a desired benefit, as the harm case does, but, rather, a means for a desired end. That is, paying an extra-dollar is not a side-effect; rather, it is a means for an end, because it causally brings about the desired benefit. Thus, the pair consisting of the extra-dollar case and the free-cup case contrasts a cost that is a means with a benefit that is a side-effect. It might be that the extra-dollar case and the free-cup case elicit different judgments about intentionality because the extra-dollar case involves a means, while the free-cup case involves a positively valued side-effect, and not because paying an extra-dollar is conceived of as a cost, while getting a free cup is not. If this were the right interpretation, the asymmetry between people's judgments in the extra-dollar and the free-cup cases would not support the trade-off hypothesis. For, this asymmetry would not provide evidence that what explains the asymmetry between the judgments elicited by the

harm and help cases (and similar cases) is that harming the environment, but not helping the environment, is conceived of as a cost–or so the objection goes.

Further research should address this objection, by showing, for instance, that side-effects are judged to be intentionally brought about *only when* they are conceived of as costs. Knobe's views predict that foreseen side-effects that are judged to be bad will always be judged to be intentionally brought about. By contrast, the trade-off hypothesis predicts that foreseen side-effects that are judged to be bad, but that are not conceived of as costs, will not be judged to be intentionally brought about.

Even if further studies are needed, I believe that this first objection is not very plausible. The striking phenomenon is that people make similar judgments when the case involves a negatively valued side-effect such as harming the environment *and* when the case involves a negatively valued means such as paying an extra-dollar. In both cases, people tend to judge that a foreseen by-product of a goal (paying an extra-dollar and harming the environment) has been intentionally brought about. The simplest and the most plausible explanation of why people have similar intuitions in the harm case and in the extra-dollar case is that when people read these two cases, they conceptualize both paying an extra-dollar and harming the environment as being a cost that the agent incurs in order to get a desired benefit. Because they believe that costs incurred to get a benefit are intentional, people conclude that harming the environment and paying an extra-dollar are intentional.

### 3.5 Second Objection

To defend Knobe' theory against the trade-off hypothesis, one could also argue that Knobe might have predicted the pattern of judgments found in the free-cup and extra-dollar cases. For, remember, Knobe and Mendlow (2004) argue that when a foreseen side-effect is judged to be bad (including, but not exclusively, morally bad), we tend to judge that the agent intentionally brings about this side-effect. When a side-effect is judged to be good (including, but not exclusively, morally good), we tend to judge that the agent does not intentionally bring about this side-effect (Figure 1). Knobe might want to generalize this idea to other foreseen by-products of goals, such as having to pay an extra-dollar. Since paying an extra-dollar is plausibly judged to be bad, while getting a free-cup is probably judged to be good, Knobe and Mendlow's account predicts the pattern of judgments found in the free-cup and extra-dollar cases. Hence, the findings are consistent with Knobe and Mendlow's account—or so the objection goes.

To test this objection, I developed two new cases, the worker case and the dog case (see below), that are based on the famous trolley case. In both cases, an agent acts in a way that brings about a side-effect—respectively, causing the death of a worker or saving a dog. In the worker case, the side-effect—causing the death of a worker—can be thought of as a cost to be incurred to reap a greater benefit—saving five workers. In the dog case, the side-effect—saving a dog in addition to five workers—cannot be thought of in this way.

**The worker case**

John is standing near the tracks of a trolley. John notices that the brakes of the trolley have failed. Five workmen are working on the tracks with the backs

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

**The dog case**

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

For each case, subjects were asked one of two questions. The first question ('the intentionality question') bears on whether the agent intentionally brought about the side-

effect described in the probe. The intentionality questions were formulated as follows: 'Did John intentionally cause the death of the workman on the side tracks? Yes/No' and 'Did John intentionally save the dog? Yes/No.' The second question ('the appropriateness question') bears on whether it was appropriate for the agent to bring about the side-effect described in the probe. The appropriateness questions were formulated as follows: 'Was it appropriate for John to cause the death of the workman on the side tracks in order to save the five workmen? Yes/No' and 'Was it appropriate for John to save the dog in addition to the five workmen? Yes/No.'

Partly on the basis of previous studies (Hauser et al. forthcoming), I predicted that bringing about the side-effect would be judged to be appropriate in both cases. If this is the case, Knobe's account predicts that subjects should judge *in both cases* that the side-effect has not been intentionally brought about. By contrast, the trade-off hypothesis predicts that subjects will be more likely to judge that the side-effect has been intentionally brought about in the worker case than in the dog case. For, in the worker case, causing the death of the worker in order to save five other workers is a cost that the agent is willing to incur in order to reap a greater benefit—viz. saving five workers. According to this hypothesis, subjects should be more likely to judge that the agent intentionally brought about the side-effect in the worker case than the side-effect in the dog case, because they are more likely to conceptualize the side-effect in the worker case—causing the death of the worker on the side tracks—than the side-effect in the dog—saving a dog in addition to five workers—as a cost incurred in order to reap a greater benefit.

135 undergraduate students from the University of Pittsburgh took part in this second experiment. In classroom settings, subjects were randomly given one of four probes, the worker case with the intentionality question (condition 1), the worker case with the appropriateness question (condition 2), the dog case with the intentionality question (condition 3) and the dog case with the appropriateness question (condition 4). 45 subjects took part in condition 1, 31 in condition 2, 30 in condition 3, and 29 in condition 4. The scoring procedure was straightforward. The intentionality question and the appropriateness question were scored binomially. A negative answer was scored 1 and a positive answer was scored 0. Percentages are presented in tables 4 and 5.

Put tables 4 and 5 about here.

A chi-square test failed to yield any significant different between subjects' answers to the appropriateness question in the worker case (condition 2) and in the dog case (condition 4) ($\chi^2$ (1, $N$ = 60) = 2.01, $p$ > .1, n.s.) (Figure 7). Thus, across both conditions, subjects did not give significantly different answers to the question about the appropriateness of the action of the agent. In both the worker case and the dog case, subjects tended to judge that it was appropriate for the agent either 'to cause the death of the workman on the side tracks in order to save the five workmen' or 'to save the dog in addition to the five workmen.' For this reason, Knobe's account predicts that subjects' judgment about the intentionality of the side-effect should not vary across the worker case and the dog case.

Put Figure 7 about here.

However, a chi-square test yielded a highly significant difference between subjects' answers to the intentionality question in the worker case (condition 1) and in the

dog case (condition 3) ($\chi^2$ (1, $N$ = 75) = 7.64, $p < 0.01$) (see Figure 8). As predicted by the trade-off hypothesis, but not by Knobe's account, subjects were significantly more likely to judge that the agent intentionally caused the death of the workman on the side tracks in order to save the five workmen than they were to say that the agent intentionally saved the dog in addition to the five workmen.

Put Figure 8 about here.

This finding strongly supports the trade-off hypothesis over Knobe's account. Knobe's account predicted that people should judge that the side-effect was not intentionally been brought in the worker case and in the dog case, because subjects' answers to the appropriateness question did not significantly differ across the two cases. The trade-off hypothesis predicted that subjects would be more likely to judge that the side-effect had been intentionally brought about in the worker case than in the dog case, because in the former case, but not in the latter case, the side-effect could be conceptualized by subjects as a cost incurred to reap a greater benefit.

**Conclusion**

The Knobe effect has been viewed by philosophers and psychologists alike as an important finding about the nature of our folk concept of intentional action and of its place in folk morality (Hauser, 2006; Leslie et al., 2006; Nichols and Ulatowski, forthcoming). Although it is premature to draw any definitive conclusion, I doubt that this is the case. First, a stumbling block prevents the resolution of the philosophical debate spurred by Knobe's and others' findings. Knobe, Nichols, Adams, Nadelhoffer and others disagree about the significance of these empirical findings for the nature of our

conceptual competence with the folk concept of intentional action. In the absence of an agreed upon theory of conceptual competence, the philosophical debate about the significance of the Knobe effect for the folk concept of intentional action cannot be resolved. Unfortunately, the literature on concepts has failed to reach, and does not seem to be heading toward, a consensus on the nature of conceptual competence. Furthermore, contrary to the received wisdom among philosophers and psychologists, the Knobe effect probably has nothing to do with our folk morality. According to the trade-off hypothesis, the asymmetry between people's judgments about the intentionality of the side-effect elicited by the harm case and the help case (or other similar cases) is merely a product of how people conceptualize the side-effect in the harm case. I propose that people conceptualize the side-effect in the harm case (and similar cases) as a foreseen cost that the agent described in the probe incurs in order to reap a foreseen benefit. Because people take costs to be intentionally incurred in order to reap benefits, they answer that the side-effect has been intentionally brought about. The evidence presented in this article supports the trade-off hypothesis. If this hypothesis is correct, the Knobe effect does nothing to show that the folk concept of intentional action has been shaped by its role in our folk morality or in ascribing blame and praise.

Department of History and Philosophy of Science

University of Pittsburgh

**References**

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

Adams, F. and Steadman, A. 2004b: Intentional action and moral considerations: Still pragmatic. *Analysis*, 64, 264-267.

Block, N. 1986: Advertisement for a semantics for psychology. In P. A. French, T. E. Uehling Jr. and H. K. Wettstein (eds.), *Midwest studies in philosophy X: Studies in the philosophy of mind*. Minneapolis: University of Minnesota Press.

Chomsky, N. 1965: *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cohen, L. J. 1981: Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4, 317-370.

Dwyer, S. 1999: Moral competence. In K. Murasugi and R. Stainton (eds.) *Philosophy and Linguistics*. Boulder, CO: Westview Press, pp. 169-190.

Fodor, J. A. 1998: *Concepts, Where Cognitive Science Went Wrong.* New York: Oxford University Press.

Hauser, M. 2006: *Moral Minds: How Nature Designed a Universal Sense of Right and Wrong.* New York: Ecco Press/Harper Collins.

Hauser, M. D., Young, L. and Cushman, F. forthcoming: Reviving Rawls' linguistic analogy: Operative principles and the causal structure of moral actions. In W. Sinnott-Armstrong (ed.), *Moral Psychology and Biology*. New York: Oxford University Press.

Hauser, M. D., Cushman, F., Young, L., Kang-Xing Jin, R. and Mikhail, J. forthcoming: A dissociation between moral judgments and justifications. *Mind & Language*.

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

Knobe, J. 2003b: Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16, 309-324.

Knobe, J. 2004: Intention, intentional action and moral considerations. *Analysis*, 64, 181-187.

Knobe, J. 2006: The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130, 2, 203-231.

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

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

Leslie, A., Knobe, J. and Cohen, A. 2006: Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science*, 17, 421-427.

Malle, B. F. 2006: The relation between judgments of intentionality and morality. *Journal of Cognition and Culture*, 6, 61-86.

Mele, A. 2003: Intentional action: Controversies, data, and core hypotheses. *Philosophical Psychology*, 16, 325-340.

Mikhail, J. 2000: *Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice'*. Unpublished PhD, Cornell University, Ithaca.

Nadelhoffer, T. 2006a: Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 9, 203-219.

Nadelhoffer, T. 2006b: On trying to save the simple view. *Mind & Language*, 21, 565-586.

Nichols, S. and Ulatowski, J. forthcoming: Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*.

Peacocke, C. 1992: *A Study of Concepts.* Cambridge, MA.: MIT Press.

Phelan, M. T. and Sarkissian, H. forthcoming: The folk strike back; Or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*.

Young, L., Cushman, F., Adolphs, R., Tranel D. and Hauser, M. 2006: Does emotion mediate the effect f an action's moral status on its intentional status? *Journal of Cognition and Culture*, 1-2, 291-304.

Wright, J. and Bengson, J. ms: Asymmetries in Folk Judgments of Responsibility and Intentional Action.

**Table 1: Percentage of subjects giving a 'yes' answer to the harm case and to the help case (adapted from Knobe 2003a)**

| | **Percentage of 'yes' answer** |
|---|---|
| **Harm case** | 82% |
| **Help case** | 23% |

**Table 2: Percentage of subjects giving a 'yes' answer to the intentionality question in the extra-dollar case and the free-cup case**

| | **Percentage of 'yes' answer** |
|---|---|
| **Extra-dollar case** | 95% |
| **Free-cup case** | 45% |

**Table 3: Percentage of subjects giving a 'neutral' answer to the value question in the extra-dollar case and the free-cup case**

| | **Percentage of 'neutral' answer** |
|---|---|
| **Extra-dollar case** | 90% |
| **Free-cup case** | 81% |

**Table 4: Percentage of subjects giving a 'yes' answer to the appropriateness question in the worker case and the dog case**

| | **Percentage of 'yes' answer** |
|---|---|
| **Worker case** | 81% |
| **Dog case** | 93% |

**Table 5: Percentage of subjects giving a 'yes' answer to the intentionality question in the worker case and the dog case**

| | **Percentage of 'yes' answer** |
|---|---|
| **Worker case** | 56% |
| **Dog case** | 23% |

**Figure 1: The Categorization Procedure for Intentional Actions according to Knobe**

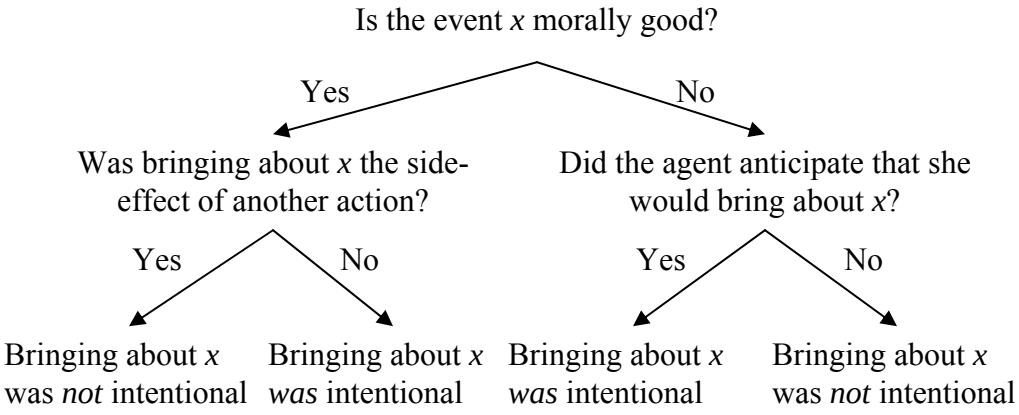

**Figure 2: Origins of our Judgments of Grammaticality**

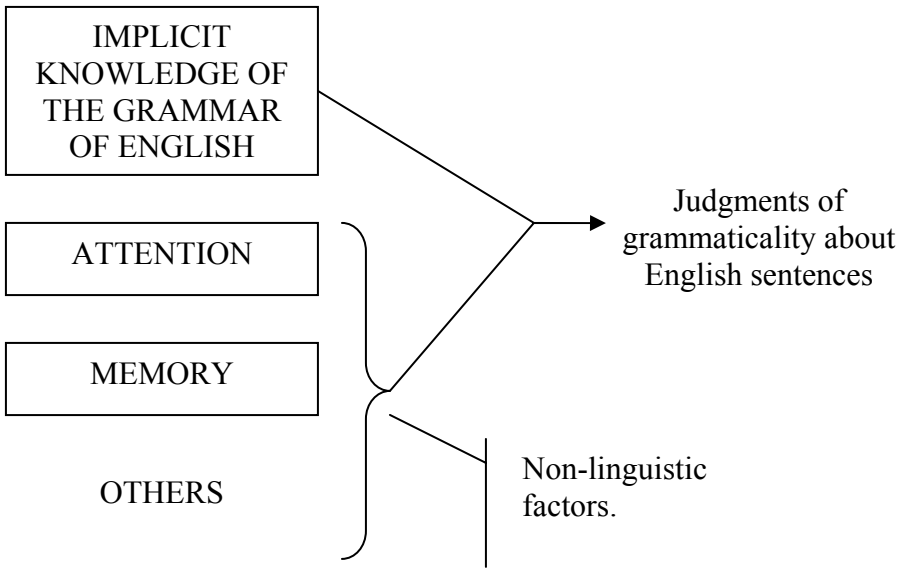

**Figure 3: People's Reasoning in the Harm and Help Cases, according to Knobe**

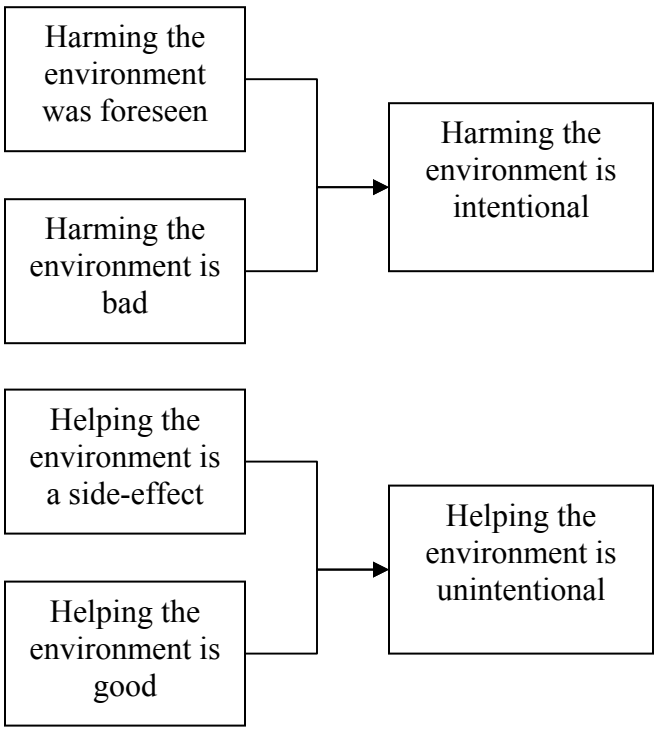

**Figure 4: People's Reasoning in the Harm and Help Cases, according to the Trade-Off Hypothesis**

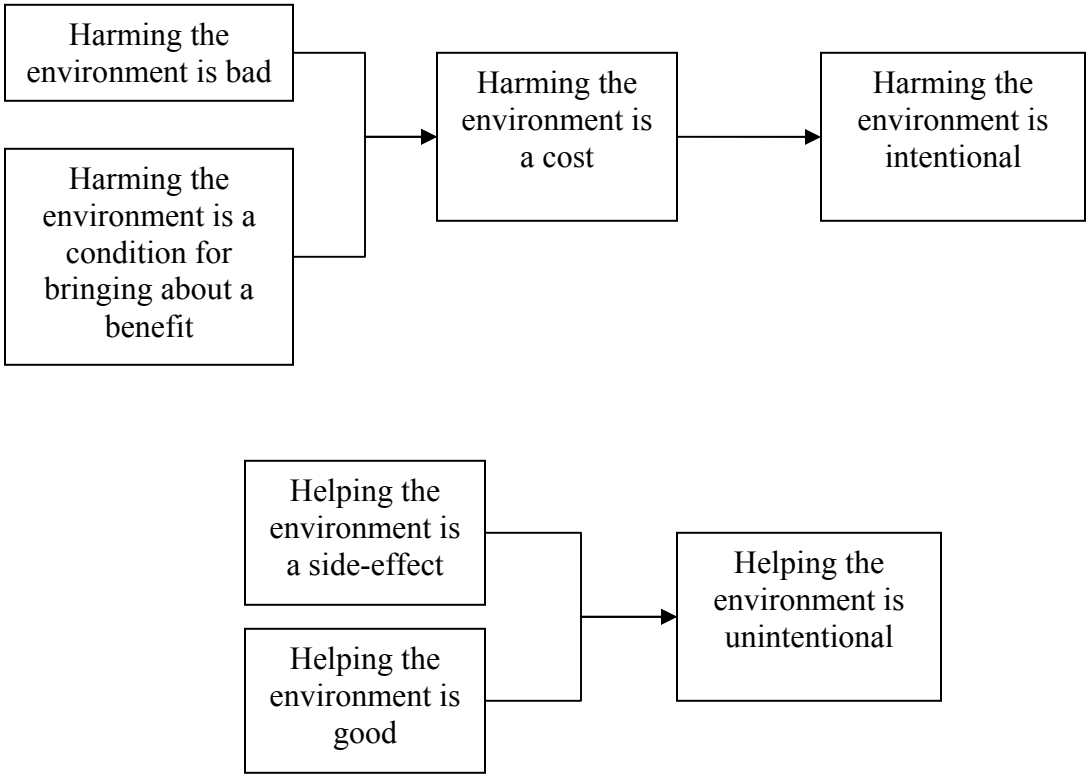

**Figure 5: Percentage of Yes for the intentionality question in the extra-dollar case and in the free-cup case**

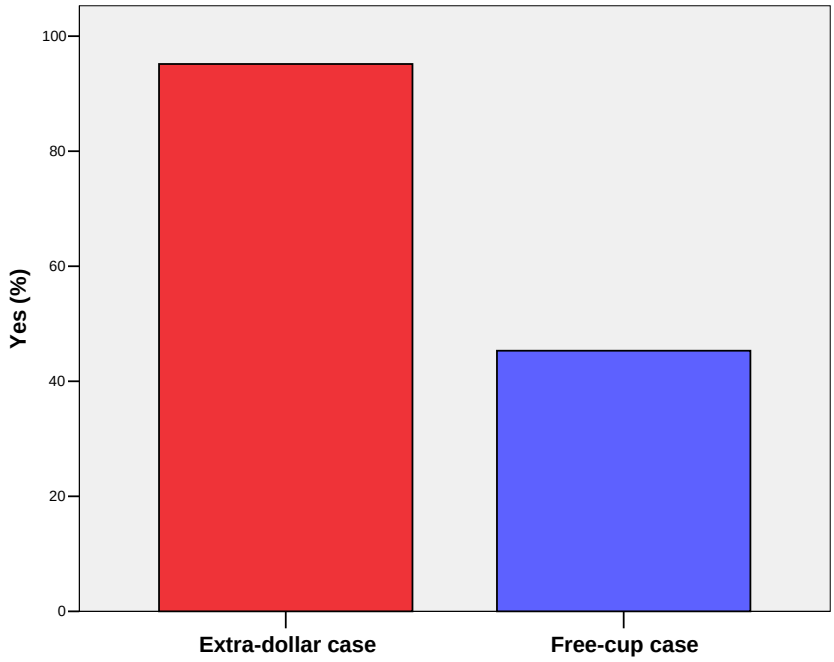

**Figure 6: Percentages for the value question in the extra-dollar case and in the free-cup case**

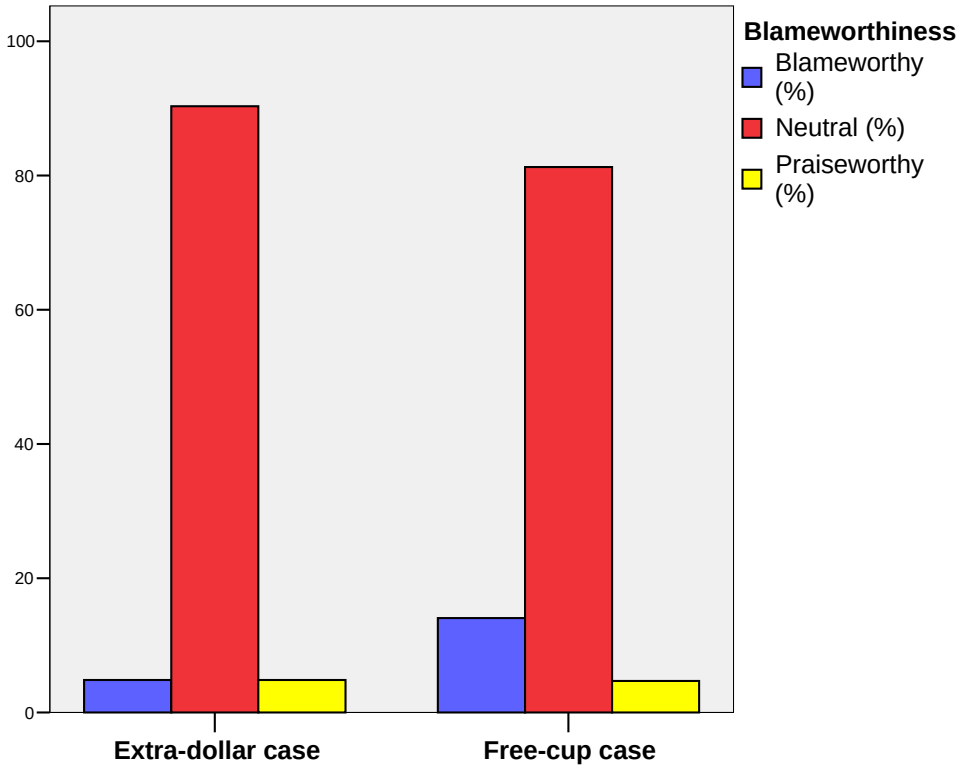

**Figure 7: Percentage of Yes for the appropriateness question in the worker case and in the dog case**

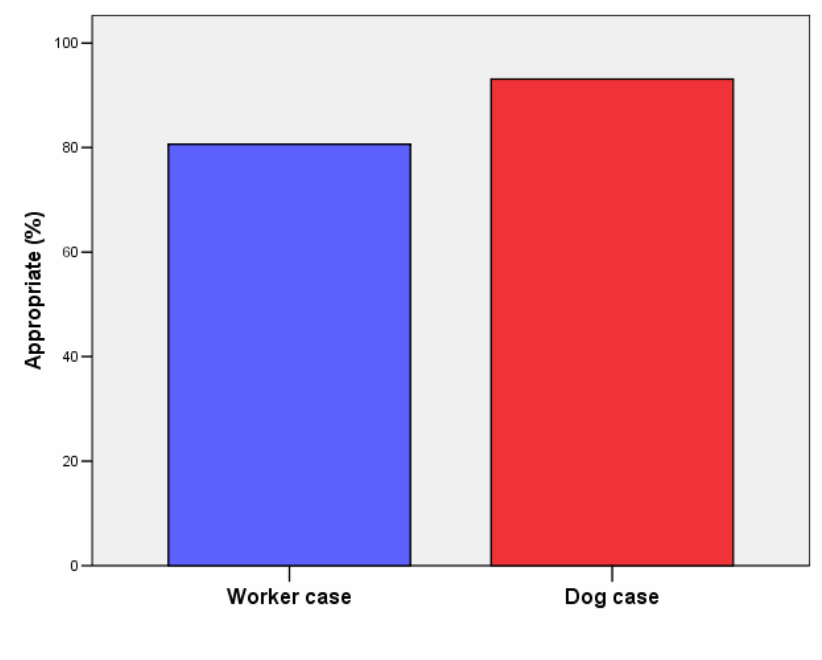

**Figure 8: Percentage of Yes for the intentionality question in the worker case and in the dog case**