Three Ambiguities in the Knobe Effect

John Michael McGuire

Hanyang University
mcguire@hanyang.ac.kr

The Knobe effect is widely regarded as one of the first and most important findings in the field of experimental philosophy. A good deal of research in this field over the past decade has been concerned with *explaining* the Knobe effect. However, much of this research has been vitiated by neglect for the more fundamental matter of *defining* "the Knobe effect." In this article I address the definitional question and argue that the Knobe effect is in fact plagued by three ambiguities which have received insufficient attention. In the first place, I show that the term has both a narrow and a broad interpretation. In its narrow sense, the term refers to an effect that moral considerations allegedly have on ascriptions of intentional action; in its broad sense, it refers to an effect that evaluative considerations allegedly have on all folk psychological ascriptions. Secondly, I show that the narrow reading of "the Knobe effect" is itself ambiguous between one interpretation on which the moral considerations in question refer to conscious moral judgments and another interpretation on which they refer to non-conscious reactions to norm violations. Thirdly, I argue that the Knobe effect can be interpreted either as a hypothesis concerning how people ordinarily use certain folk psychological concepts or as a hypothesis concerning how people use those concepts only in the context of hypothetical thought-experiments. While the vast majority of researchers have assumed the former view, recent experimental research supports the latter view, suggesting that the Knobe effect is in fact an experimental artifact.

Key words: *Joshua Knobe, the Knobe effect, the side-effect effect, intentional action, folk-psychological ascriptions*

Journal of Cognitive Science 14: 1-26, 2014

1. Introduction

In 2003 Joshua Knobe presented empirical evidence suggesting that ordinary people's judgments of intentionality are systematically influenced by moral considerations (Knobe 2003a; 2003b). It has long been understood that judgments of intentionality do, and should, influence moral evaluations of human agents and their actions; what was surprising about Knobe's experimental findings is that they provided evidence that the influence might also work in the opposite direction—that moral evaluations of agents and/or their actions might systematically affect people's judgments concerning whether or not those actions were performed intentionally. This intriguing suggestion set off an explosion of research among philosophers, psychologists, and cognitive scientists that has over the course of the past ten years corroborated, challenged, and extended Knobe's findings in several ways.

At the heart of this research is a phenomenon known as either "the side-effect effect" (Leslie, Knobe & Cohen, 2006; Beebe & Buckwalter, 2010; Uttich & Lombrozo, 2010; Wellen & Danks, 2012) or, more commonly, "the Knobe effect" (Nichols & Ulatowski, 2007; Mallon, 2008; Machery, 2008; Holton, 2010; Levy, 2011; Feltz, Harris & Perez, 2012).[1] Most of the research that has been carried out in response to Knobe's ground-breaking findings falls into one of two general categories. On the one hand, some researchers have attempted to *explain* the Knobe effect by identifying the key factor(s) that best account for it (Adams & Steadman 2004a, 2004b; Knobe & Mendlow, 2004; Machery, 2008; Nichols and Ulatowski, 2007; Guglielmo & Malle, 2010; Holton, 2010). On the other hand, other researchers have attempted to *extend* the Knobe effect, or in other words show that the effect in question manifests itself in a much broader range of

[1] Neither of these two terms is entirely felicitous, but the former term, "the side-effect effect," is especially problematic since, as we will see below, the effect in question is not confined to side-effects. For this reason I will in what follows use the term "the Knobe effect," noting that the person after whom the effect in question is named (Joshua Knobe) understandably does not use this term himself.

phenomena than Knobe (2003a) first identified (Knobe, 2004b; Nadelhoffer, 2005; Knobe & Fraser, 2008; Cushman et al., 2008; Pettit & Knobe, 2009; Ulatowski, 2012).

These two research projects have not been completely complimentary. Indeed, as we will see below, research suggesting that the Knobe effect is a far more general phenomenon than originally thought has been used by Knobe and others to argue against certain explanatory accounts of the effect, especially those that focus narrowly on features of the concept of intentional action (Pettit & Knobe, 2009; Holton, 2010). This dynamic between these two projects reveals a rather striking fact—that despite all the research that has been carried out on the Knobe effect in the past decade, there is a remarkable lack of clarity on what exactly the Knobe effect is. This is partly because some researchers have defined "the Knobe effect" in a way that later research has shown to be too restrictive, but also because many of the researchers working in this field have simply failed to define "the Knobe effect" at all. Instead, it has been taken largely for granted that the effect itself is well understood even if the correct theoretical explanation for the effect remains elusive or contentious. However, in what follows I will take up the definitional question and show that defining "the Knobe effect" is not nearly as straightforward as many have supposed. Indeed, I will argue that the Knobe effect is plagued by three ambiguities that have received insufficient attention.

The order in which these ambiguities will be discussed is as follows. In Section 2, I will show that "the Knobe effect" has both a narrow and a broad interpretation. In its narrow sense, the term refers to an effect that moral considerations allegedly have on ascriptions of intentional action; in its broad sense, the term refers to an effect that evaluative considerations allegedly have on all folk psychological ascriptions. As we will see, different researchers use the term in these different senses without any acknowledgement of the ambiguity. In Section 3, I will show that the narrow or more restrictive reading of the Knobe effect is itself ambiguous and that it has been defined in two very different ways. On the one hand it has been characterized as an effect that conscious moral judgments allegedly have on ascriptions of intentional action; on the other hand, it has been characterized as an effect that non-conscious reactions to norm

violations have on ascriptions of intentional action. While the former characterization is by far the more common one, I will show that there is a clear counterexample to that interpretation of the Knobe effect. Finally, in Section 4, I address the question of whether "the Knobe effect," should ultimately be understood as a hypothesis about how people actually use the concept of intentional action and perhaps other folk psychological concepts or rather a hypothesis about how people use those concepts only in hypothetical though-experiments. While the majority of researchers have uncritically assumed the former view, I will show that recent experimental research on this question supports the latter view.

2. Intentional action or folk psychology?

The obvious place to begin a discussion of the Knobe effect is with the experimental findings that launched this area of research. In order to determine whether moral considerations influence ordinary people's judgments of intentional action, Knobe (2003a) recruited 78 research subjects, randomly assigned them to either one of two groups ("Harm" or "Help"), and asked them to read the corresponding (Harm or Help) versions of the following scenario.

Chairman Scenario

> 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 [help] the environment." The chairman of the board answered, "I don't care at all about harming [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 harmed [helped]. (Knobe 2003a, p. 191)

After reading the scenarios, subjects in both groups were asked to answer the appropriate versions of the following two questions: (a) On a scale from 0 to 6, how much blame [praise] does the chairman deserve for what he did? (b) Did the chairman intentionally harm [help] the environment?

Knobe (2003a) found that while 82% of subjects in the Harm group said that the chairman intentionally *harmed* the environment, only 23% of subjects in the Help group said that the chairman intentionally *helped* the environment. This result is statistically significant, as is the difference between the two group's responses to the second question. While subjects in the Harm group reacted with high levels of *blame* for what the chairman did; subjects in the Help group gave the chairman low levels of *praise*.

To test "the generality of the effect" found in this experiment, Knobe (2003a) ran a second experiment using a structurally similar scenario involving a military commander rather than a corporate executive. This second experiment yielded similar results to the first. On the basis of the results of these two experiments, Knobe (2003a) reported two striking asymmetries, one concerning subjects' ascriptions of intentional action, the other concerning subjects' judgments of praise and blame. Since these two asymmetries are central to the discussion that follows, let us designate and define them as follows:

> **Intentionality Asymmetry:** People are "considerably more willing to say that a side-effect was brought about intentionally when they regard that side-effect as bad than when they regard it as good." (Knobe 2003a, p. 193)
>
> **Praise/Blame Asymmetry:** People are "considerably more willing to blame [an] agent for bad side effects than to praise [an] agent for good side effects." (Knobe 2003a, p.193).

It is important to be clear from the outset that these two asymmetries are neither observed phenomena nor, in the first instance, explanatory hypotheses; rather, they are descriptive hypotheses or generalizations about human behavior based on observed phenomena, such as the experimental data reported by Knobe (2003a). As hypotheses they can be either true or false, and the way to determine their truth-value is to test the predictions they make concerning unobserved phenomena. The Intentionality Asymmetry, for instance, generates predictions about how people will judge the intentionality of other morally asymmetrical side-effect actions besides

the ones that Knobe (2003a) investigated.

In addition to providing experimental evidence in support of these two asymmetries, Knobe (2003a) suggested that the latter asymmetry may be "at the root of," or in other words, account for the former asymmetry. However, the main concern of that article was not so much to explain the Intentionality Asymmetry as it was to provide evidence in support of it. In subsequent work Knobe has provided explanations of this asymmetry which we will consider below (Knobe & Mendlow, 2004; Knobe 2006, 2007, 2010; Pettit & Knobe, 2009). In the present context it is sufficient to note that the Intentionality Asymmetry, which was first documented in Knobe (2003a), has been corroborated by many other researchers (McCann, 2005; Nichols & Ulatowski, 2007; Cushman & Mele, 2008; Phelan & Sarkissian, 2009; Guglielmo & Malle, 2010). Evidence in support of the Intentionality Asymmetry comes from a variety of sources including children (Leslie et al., 2006) as well as people from non-western cultures (Knobe & Burra, 2006). At this point in time the Intentionality Asymmetry is widely considered to be real and "remarkably robust" (Nichols & Ulatowski, 2007, p. 355).

But how exactly does the Intentionality Asymmetry relate to the Knobe effect? Some researchers working in this area (e.g. Nichols & Ulatowski, 2007, Mallon 2008; Machery, 2008) have understood the Knobe effect in terms of the Intentionality Asymmetry. However, it seems that that these two concepts cannot be identified, for a good deal of research subsequent to Knobe (2003a) suggests that the Knobe effect is much broader in scope than the Intentionality Asymmetry. In the first place, unlike the Intentionality Asymmetry, Knobe insists that "the effect" in question is not limited to cases involving side-effects (2010, p.318). Indeed, both Knobe (2003b) and Nadelhoffer (2005) have demonstrated that the same sort asymmetry in ascriptions of intentional action that Knobe (2003a) documented can be generated in experimental conditions using scenarios that do not involve side-effect actions at all. Secondly, unlike the Intentionality Asymmetry, the effect in question is not limited to ascriptions of intentional action. Thus, Knobe writes that:

> The effect does not appear to be limited to the concept intentionally, nor even to closely related concepts such as intention or intending. Rather, it seems that we are tapping into a much more general tendency, whereby moral judgments impact the application of a whole range of different concepts used to pick out mental states and processes. (2010, p. 318)

In support of this claim, Pettit and Knobe (2009) have demonstrated that the effect is found in concepts such as "intention," "decide," "desire," "in favor of," and "advocating." Furthermore, Knobe (2010, p. 319) writes that "The scope of the effect does not stop there. It seems to apply to intuitions about the relations that obtain among the various actions that an agent performs," such as the relations captured by the expressions "in order to" and "by" (Knobe 2004a, 2007). Moreover, Knobe (2010, pp. 319-320) points out that "the very same effect arises in people's intuitions about causation" as well as their intuitions about "doing" versus "allowing." So it seems that the hypothesis that has come to be known as "the Knobe effect" cannot simply be defined in terms of the Intentionality Asymmetry, which relates much more specifically to side-effects and ascriptions of intentional action.

What then is "the Knobe effect?" Knobe (2010, p. 320) sums up his discussion of the various manifestations of the effect that he is interested in by saying that "Thus far, we have seen that people's ordinary application of a variety of different concepts can be influenced by moral considerations. The key question now is how to explain this effect" (Knobe, 2010, p. 320)." And Pettit and Knobe (2009) suggest that "there are no concepts anywhere in folk psychology" that are not susceptible to this effect. Accordingly, one might define the Knobe effect simply as the claim or hypothesis that moral considerations can influence people's use of folk psychological concepts. However, this definition is problematic for at least two reasons. First, the idea that moral considerations *can* influence people's use of folk psychological concepts is neither controversial nor interesting: clearly all kinds of things can and do influence people's use of psychological concepts. The Knobe effect, if it has any real significance, must say something more specific about the sort of effect that moral considerations have on people's use of those concepts. The Intentionality Asymmetry, for

instance, is interesting because it says something quite specific about the relation between ordinary people's moral judgments and their ascriptions of intentional action, something from which testable predictions can be derived. Similarly, if the Knobe effect has any significance, then it too should be capable of generating testable predictions, which is something that the above definition clearly does not do. Second, Knobe has indicated that the effect he is interested in is not limited to moral considerations and that extra-moral evaluations (e.g. aesthetic evaluations) seem to affect folk-psychological ascriptions in a similar way that moral evaluations do (Knobe & Mendlow, 2004; Knobe, 2004b). So the Knobe effect is ultimately a hypothesis about how the use of certain psychological concepts is affected, not specifically by moral considerations, but rather by "normative or evaluative considerations" broadly construed to include moral as well as aesthetic and other evaluative phenomena.

The following example will serve to illustrate why Knobe has abandoned the idea that it is specifically moral considerations that generate the sort of asymmetries in psychological ascriptions that he is interested in. Knobe (2004b) describes an experiment involving a scenario modelled on his original Chairman Scenario. Subjects in this study were randomly assigned to one of two groups ("Aesthetically Worse" and "Aesthetically Better") and asked to read the appropriate version of the following scenario.

Movie Executive Scenario

> The Vice-President of a movie studio was talking with the CEO. The Vice-President said: "We are thinking of implementing a new policy. If we implement the policy, it will definitely increase profits for our corporation, but it will also make our movies worse [better] from an artistic standpoint." The CEO said: "Look, I know that we'll be making the movies worse [better] from an artistic standpoint, but I don't care one bit about that. All I care about is making as much profit as I can. Let's implement the new policy!" They implemented the policy. As expected, the policy made the movies worse [better] from an artistic standpoint. (Knobe, 2004b, p. 274)

After reading the assigned scenario, subjects in both groups were asked to answer the appropriate versions of the following two questions: (a) "Did the CEO intentionally make the movies worse [better] from an artistic standpoint?" and (b) "How much blame or praise does the CEO deserve for what he did?" This second question was answered on a scale from -3 ("a lot of blame") to +3 ("a lot of praise").

According to Knobe (2004b), both questions produced statistically significant differences parallel to the asymmetries found in his original experiment on the Chairman Scenario. In particular, the mean rating of blame/praise for those in the Aesthetically Worse group was -1.7 (significant blame) while the mean for those in the Aesthetically Better group was 0.3 (little or no praise). Similarly, while 54% of subjects in the Aesthetically Worse group said that the CEO intentionally made the movies worse, only 18% of subjects in the Aesthetically Better group said that the CEO intentionally made the movies better. Knobe (2004b) therefore concludes that aesthetic evaluations appear to have the same kind of effect that moral evaluations do, although the size of the effect is smaller than the effect normally observed in moral cases.

So the Knobe effect—whatever it is—allegedly applies to virtually all folk psychological concepts and relates, not only to moral considerations, but rather to evaluative considerations broadly construed. Perhaps the Knobe effect can be defined in a way that is both general enough to include all of the phenomena that Knobe seems to think are governed by the same principles and yet specific enough to say something interesting and generate testable predictions. However, as far as I can tell, no such definition currently exists. Researchers writing about "the Knobe effect" have generally done one of two things: either they have not attempted to define the term (Nichols & Ulatowski, 2007; Holton, 2010) or else they have defined it by identifying it with the Intentionality Asymmetry (Mallon, 2008; Machery, 2008). However, there are problems with both of these approaches.

To appreciate the problems with the former approach, let us observe how different researchers who have used the term without defining it, end up using it in very different ways. Consider first Nichols and Ulatowski (2007). The term "the Knobe effect" appears in the title of the article and again

in the title of the second section, and the main goal of the article, it seems, is to present and defend a novel explanation of the Knobe effect. And yet nowhere in the course of the article is the term explicitly defined. Instead, the reader is left to infer from the introduction and second section of the article that "the Knobe effect," as Nichols and Ulatowski understand it, is the experimental finding that "people's intuitions about whether an outcome was intentionally produced seem to vary depending on the moral status of the outcome itself" (2007, p. 346). In other words, they basically understand the Knobe effect in terms of the Intentionality Asymmetry.

But now consider Holton (2010). The term "the Knobe effect" appears in the title of his article too, and also in the title of the first section, and the chief aim of the article, it seems, is to provide a principled explanation and justification of the Knobe effect. And yet, once again, nowhere in the course of the article is the term actually defined. Instead, Holton (2010) gives a brief and clearly incomplete summary of some of the experimental findings that have been made by Knobe and other researchers working in this field. He then writes:

> Various explanations of these results have been offered…But most have been piecemeal, accounting for one finding or another. Ideally we want an explanation that accounts for all of them in a unified way. That is what I try for here. (Knobe, 2010, p. 418)

So Holton understands "the Knobe effect" to refer to a much broader range of phenomena than what is stated in the Intentionality Asymmetry.

The first problem then with the failure to define "the Knobe effect" is that researchers who are explicitly writing about it are using that term in very different ways. Given all of the various explanations of the Knobe effect that have been offered, including those by Nichols & Ulatowski (2007) and Holton (2010), it is hard to see how there might be progress in explaining the Knobe effect when there is such unacknowledged disagreement on what the Knobe effect actually is. A second, but related, problem with the failure to address the definitional question is independent of the inconsistency between researchers and applies to those, like Pettit and Knobe (2009) and Holton (2010), who see this effect as a much broader phenomenon than

what is stated in the Intentionality Asymmetry. As a matter of principle, before one attempts a unified explanation for a diverse range of phenomena one needs to be clear on which phenomena the explanation is supposed to explain, and for that, a definition would seem to be needed. At the very least one needs to be quite clear on the conditions that must be met in order for a given phenomenon to count as a manifestation of the Knobe effect, otherwise there is no way of knowing exactly what one is attempting to explain.

On the other hand, those researchers who have offered an explicit definition of "the Knobe effect" have generally identified it with the Intentionality Asymmetry. The main problem with this approach is that, as we have seen, there is a large body of research indicating that the Intentionality Asymmetry is itself one manifestation of a much more general phenomenon, even if that broader phenomenon has yet to be clearly defined or characterized. Of course, as a term of art, one is free to define "the Knobe effect" as one pleases, either in the more restrictive sense (i.e. the Intentionality Asymmetry) or in the broader sense that Pettit & Knobe (2009) and Holton (2010) have in mind. However, in Knobe's view at least, there is a real danger in defining the explanandum at the heart of this research program in the narrow sense of the Intentionality Asymmetry. The danger is that by restricting one's attention to the Intentionality Asymmetry, one will fail to grasp the deeper principles at play in the phenomena under investigation. Indeed, that is precisely the criticism that Knobe has made of many of the other explanations of the Intentionality Asymmetry, such as those offered by Nichols & Ulatowski (2007) and Machery (2008).

For example, Nichols and Ulatowski (2007) explain the Intentionality Asymmetry in terms of an ambiguity in the word "intentional." There are, they maintain, two different concepts (foreknowledge and motive) associated with the word "intentional": some people consistently use one concept, some use the other, and a third group of people use either one depending on the situation. Whatever merit there is to this explanation—and there certainly is evidence in support of it—it cannot possibly apply to the other concepts subject to the Knobe effect, such as "desire," "approve of," "cause," and so on. It is for this reason that Pettit and Knobe (2009) insist that explanations such as the one offered by Nichols and Ulatowski

(2007) are off the mark: they focus on features of the concept of intentional action and miss the more general principles at play in the phenomenon they are attempting to understand.

Let us briefly note one further example of the point being made here. Guglielmo & Malle (2010) present a compelling case for the idea that the Intentionality Asymmetry can be explained in terms of an asymmetry in the perceived level of desire in the relevant agent in different scenarios. For example, they say that what drives the asymmetrical responses of the Harm and Help groups in Knobe's original study on the Chairman Scenario is that, even though the chairmen in both versions of this scenario claim that they "don't care at all" about the effect in question, subjects in the Harm group interpret the chairman as *wanting to harm* the environment much more than subjects in the Help group interpret the chairman as *wanting to help* the environment. According to Guglielmo & Malle (2010, p. 1643) "This is because people interpret not caring about negative outcomes as evidence of moderate desire, but not caring about positive outcomes as evidence of virtually no desire." So in their view, the two versions of the Chairman Scenario are not structurally identical, as many people have supposed. Rather, the experimental design masks real differences in perceived levels of desire for the side-effect in the agents in the two contrasting scenarios. Part of what makes the case put forward by Guglielmo & Malle (2010) compelling is that they present the results of experiments in which they adjust the wording in both (Harm and Help) versions of the Chairman Scenario to increase and decrease the perceived level of desire the chairman has for the given outcome. They found that judgments that the chairman brought about the outcome intentionally dropped to 40-59% in the negative case when the harming chairman regretted the negative side-effect, and they increased to 56% in the positive case when the helping chairman welcomed the positive side-effect. This suggests that judgments about the chairman's desire for the outcome in question really do influence subjects' judgments concerning the intentionality of the chairman's behaviour.

However, while Guglielmo & Malle (2010) present a compelling explanation of the Intentionality Asymmetry, from Knobe's point of view this explanation cannot be correct since it is not deep enough to account for all of the other asymmetries related to the Intentionality Asymmetry.

Indeed, Pettit and Knobe (2009) have provided independent confirmation of the fact that subjects in the Harm condition of the Chairman Scenario have significantly different judgments on whether or not the chairman *wanted* to bring about the side-effect in question in comparison to the subjects in the Help condition. However, from Knobe's point of view, this latter asymmetry is just as much in need of an explanation as the Intentionality Asymmetry. Furthermore, unlike the Intentionality Asymmetry, many of the asymmetries that Knobe believes are in need of explanation seem to have little or nothing to do with the concept of desire.

Consider, for instance, the alleged asymmetry in judgments concerning agent-causation. The following study, from Knobe & Fraser (2008) will serve as an example:

The Missing Pen Scenario

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens. On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message . . . but she has a problem. There are no pens left on her desk. (Knobe & Fraser, 2008)

Faced with this vignette, most subjects say that the professor *did cause* the problem but that the administrative assistant *did not cause* the problem. Whatever the explanation is for these asymmetrical judgments, it is unlikely that it has anything to do with differences in the perceived levels of desire for the outcome between the agents. That is, it is doubtful that people would interpret this scenario as indicating that either Professor Smith or the administrative assistant *wanted* to cause the problem. More importantly, even if people did have different judgments concerning levels of desire between these two agents, there is no reason to expect this to be relevant to

judgments about agent-causation. Unlike the concept of intentional action, which does seem to be related to the concept of desire; there is no semantic connection between desire and causation. If a man throws a ball that strikes a window and causes it to break, then the man caused the window to break regardless of whether or not he had any desire to break the window. Having a desire to break the window may well be a necessary condition of breaking the window *intentionally*, but it is clearly not a necessary condition of *causing* the window to break.

Let us sum up the discussion so far. While a great deal has been written on the Knobe effect in the past decade, the foregoing survey of the relevant literature reveals that these terms have been used in different ways by different researchers. Some researchers have used the term to refer to the Intentionality Asymmetry, which relates specifically to moral considerations, side-effects, and judgments of intentionality. Yet other researchers have used the term to refer to a much more general phenomenon of which the Intentionality Asymmetry is one manifestation. This more general phenomenon has nothing in particular to do with moral considerations, side-effect actions, or judgments of intentionality, but relates rather to the effect that evaluative considerations in general have on people's use of virtually all folk psychological concepts. So there are then two different senses to the term "the Knobe effect," a narrow and a broad sense. However, these two senses are not on a par in terms of clarity or usefulness. While "the Knobe effect" in the narrow sense (i.e. the Intentionality Asymmetry) is clear and precise enough to generate testable predictions, this is not the case with the broad sense of the term, which is too vague to be tested experimentally.

This ambiguity in "the Knobe effect," which seems to have gone largely unnoticed, is obviously a source of potential confusion, but beneath it lurks a deeper problem. Even once the ambiguity is made clear, researchers working on the Knobe effect have a difficult choice to make: either they can investigate the Intentionality Asymmetry, as some have done (Nichols & Ulatowski, 2007; Machery 2008, Wright & Bengson, 2009) or they can investigate the much more general phenomenon (Pettit & Knobe, 2009; Holton, 2010). Both approaches have their own unique problems. The problem with the former approach is that, if the Intentionality Asymmetry is indeed a manifestation of a much more general phenomenon, as Knobe

insists, then whatever explanation that is given for the Intentionality Asymmetry will likely miss the deeper principles at play in the phenomenon under investigation. The problem with the latter approach is that, no satisfactory definition of the more general phenomenon has been given yet, and it is unlikely that researchers working on this more general phenomenon can provide a satisfactory explanation of it, if they cannot even define it or describe it in a way that generates testable predictions. Indeed, in the absence of such a definition, it is impossible to know whether there even is such a thing as "the Knobe effect."

3. Conscious moral judgments or unconscious reactions to norm violations?

Since it is not possible to generate testable predictions for the Knobe effect in its more general sense, let us concentrate instead on the narrower sense of the term, the Intentionality Asymmetry, or what Wellen & Danks (2012, p. 2523) call the "canonical expression" of the Knobe effect. It may seem odd to call into question the truth of the Intentionality Asymmetry, for as we have already observed, there is a great deal of evidence in support of it. But the question I am raising now is not whether there is any evidence *in support of* the Intentionality Asymmetry, but whether there is any evidence *against* it. As it happens, there is, and while this recalcitrant evidence is not exactly a secret, it has I think received insufficient and inconsistent attention. Let us now consider what that evidence is and what implications it has for the understanding of the Knobe effect.

Knobe (2007) recruited 41 research subjects, randomly assigned them to either one of two groups ("Violate" and "Fulfill"), and asked them to read the appropriate version of the following scenario.

Nazi Germany Scenario

> 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 [fulfilling] the requirements of the racial identification law." The CEO said: "Look, I know that I'll be violating [fulfilling] 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. (Knobe, 2007, p. 106)

After reading the scenario, subjects were asked to answer the appropriate versions of the following two questions: (a) Did the CEO intentionally violate [fulfill] the requirements of the law? (b) On a scale from -3 ("a lot of blame") to +3 ("a lot of praise"), how much blame or praise does the CEO deserve for what he did?

Knobe (2007) reports that, with respect to the first question, 81% of subjects in the Violate condition said that the CEO intentionally violated the requirements of the law, whereas only 31% of subjects in the Fulfill condition said that he intentionally fulfilled the requirements of the law, a statistically significant difference. With respect to the second question about how much blame or praise the CEO deserved, Knobe (2007) reports that there was no significant difference between responses in the Violate and Fulfill conditions (M = -0.9; M = -1.7).

Notice how the results of this experiment differ in an important respect from the results of Knobe's original experiment on the Chairman Scenario. In the experiment on the Chairman Scenario, Knobe (2003a) found that subjects were more inclined to judge that the chairman acted intentionally in the morally bad (Harm) condition than in the morally good (Help) condition. On the basis of these results, as well as those of one other related experiment, Knobe (2003a) proposed the Intentionality Asymmetry, the hypothesis that people are "considerably more willing to say that a side-effect was brought about intentionally when they regard that side-effect as bad than when they regard it as good" (Knobe 2003a, p. 193). As noted in the foregoing, this hypothesis can be used to generate testable predictions. With respect to the Nazi Germany Scenario, the Intentionality Asymmetry

predicts that subjects in the morally bad (Fulfill) condition should be more inclined than subjects in the morally good (Violate) condition to judge that the CEO acted intentionally. However, the experimental results were exactly the opposite. So it seems that the experiment on the Nazi Germany Scenario constitutes a clear counterexample to the Intentionality Asymmetry.

Knobe (2007) presented a rather detailed psychological theory to explain why subjects respond as they do to the Nazi Germany Scenario. The theory is described in the following passage.

> Suppose that we are thinking about a society governed by some morally abhorrent law (say, a law according to which one is obliged to kill people of certain races). Now consider what might happen if we learned that a member of this society violated the law. As soon as we encountered this case, we would begin a rapid nonconscious process of evaluation that only made use of the most salient norms. If the law itself was made salient in the context, the law would be the most salient norm and the behavior would therefore be classified as a transgression. Subsequently, we might take a moment to reflect and consciously think about whether the agent's behavior was right or wrong. In that subsequent process, we would determine that the law itself was morally abhorrent and that there was nothing wrong with violating it. But this subsequent reflection would not alter our initial nonconscious judgment. That judgment would remain in place and would continue to influence our intentional action intuitions. (Knobe, 2007, p. 103)

At the heart of Knobe's theory then is a distinction between two very different types of psychological processes or judgments. On the one hand, there are the quick, automatic, and non-conscious responses that people have to norm violations or in other words "transgressions." On the other hand, there are the slower, conscious, and more nuanced moral judgments that people tend to make of agents and their actions. The distinction here is not merely a distinction between conscious and non-conscious processes; it is also a distinction between quick, single-factor judgments on the one hand and slower, deliberative, and more comprehensive judgments on the other. It is the former non-conscious "judgments" or reactions that Knobe

ultimately came to think influence people's judgments of the intentionality of side-effect actions (2007, p.101).

The experiment using the Nazi Germany Scenario demonstrates that the Intentionality Asymmetry is false, and Knobe seems to fully agree with this assessment. Thus, he says of that experiment, that "what we have here is a case in which subjects consciously believe that violating the requirements [of the 'racial identification law'] is actually a good thing and nonetheless end up concluding that the agent acted intentionally" (Knobe, 2007, p. 103). But while Knobe came to agree that the Intentionality Asymmetry is false, he did not abandon the core idea that the Intentionality Asymmetry expresses. Instead he reformulated the expression of that idea, which may be described as follows:

> **Intentionality Asymmetry*:** People are considerably more willing to say that a side-effect is brought about intentionally when they regard that side-effect as violating a salient norm than when they regard it as violating no norms.

Non-conscious judgments of norm violations and conscious moral judgments often coincide, but they can diverge, as they do in the case of the Nazi Germany Scenario. Because these two different types of judgments do sometimes diverge, the Intentionality Asymmetry and the Intentionality Asymmetry* are not synonymous.

However, the distinction between these two versions of the Intentionality Asymmetry has been largely lost in the ongoing research on the Knobe effect. Not only do other researchers continue to write as though the Intentionality Asymmetry is established fact (Mallon, 2008; Pinillos et al., 2011; Cova & Naar, 2012), so too does Knobe. For instance, Pettit & Knobe (2009) conclude their discussion of the pervasive impact of moral judgments on pro-attitudes as follows:

> In light of these results, we are inclined to think that the impact of moral judgment is pervasive, playing a role in the application of every concept that involves holding or displaying a positive attitude toward an outcome. That is, for all concepts of this basic type, we suspect that there is a psychological process that makes people more willing to

> apply the concept in cases of *morally bad side-effects* and less willing to apply the concept in cases of *morally good side-effects*. (Pettit & Kobe, 2009, p. 593, my italics)

And similarly, in his most recent and comprehensive review of this field of research, Knobe describes the "surprising" finding at the heart of this research as follows: "people's judgments about whether a given action truly is morally good or bad can actually affect their intuitions about what that action caused and what mental states the agent had" (Knobe, 2010, p, 315).

Pettit & Knobe (2009) and Knobe (2010) both advance a theory, which Knobe (2010) describes as a "competence theory," to explain the various asymmetries he thinks are governed by the same principles. And yet nowhere in either of these articles is there any explicit discussion of non-conscious transgression detection. The theory on offer seems to be a theory designed to explain why people's *moral judgments* seem to be influencing their folk psychological ascriptions. In other words, the competence theory that Knobe defends seems to be a theory that purports to explain the Intentionality Asymmetry, not the Intentionality Asymmetry*. But since the former hypothesis is false, it is not in need of any explanation. The latter hypothesis, on the other hand, may or may not be true, but there is little if any empirical evidence in support of it. The Intentionality Asymmetry* has received nothing like the attention or experimental scrutiny that its cousin has. Moreover, it is not a hypothesis that can be tested in the way that the Intentionality Asymmetry has been tested (i.e. using simple questionnaires). As we have seen, the notion of norm-violation detection involved in the Intentionality Asymmetry* is thought to operate beneath the level of conscious awareness and can diverge sharply from conscious moral judgments. Thus, testing this hypothesis is no simple matter.

In the previous section we noted that the Knobe effect is ambiguous between a narrow and a broad interpretation. The discussion in this section shows that the narrow sense of the Knobe effect is itself ambiguous between one interpretation (the Intentionality Asymmetry) that is widely recognized but false and another interpretation (the Intentionality Asymmetry*) that may or may not be true, but which has received very little attention or empirical support.

4. Source of insight into folk psychology or experimental artifact?

Let us return to the philosophical experiment that ushered in this new field of research. As we have noted, on the basis of experimental data concerning how subjects responded to the Chairman Scenario, as well as the results of one other related experiment, Knobe (2003a) inferred the Intentionality Asymmetry, the hypothesis that people are considerably more willing to say that a side-effect was brought about intentionally when they regard that side-effect as bad than when they regard it as good. In drawing this inference, Knobe made two crucial assumptions that can be, and should be, questioned.

In the first place, it was an assumption on Knobe's part to treat the responses subjects gave to his Chairman Scenario as evidence concerning how people *actually use* the concept of intentional action. Responding "yes" or "no" to a question concerning a hypothetical thought-experiment may very well be a flawed way to assess how people use a certain concept in realistic contexts, and it is people's actual use of these concepts, not merely how they use them in hypothetical thought-experiments, that is the target of this sort of research. Accordingly, one can reasonably ask whether the experimental data that Knobe and other researchers collected in support of the Intentionality Asymmetry are an accurate measure of how people actually use the concept of intentional action or rather an experimental artifact caused by the hypothetical nature of the stimuli. Recent experimental evidence suggests that the Intentionality Asymmetry is actually an experimental artifact and that people do not really use the concept of intentional action in realistic situations as they do in response to the Chairman Scenario and other hypothetical though-experiments (Feltz et al., 2011; Wellen & Danks, 2012).

A second and related question overlooked by Knobe (2003a) is whether subjects would use the concept of intentional action as the Intentionality Asymmetry predicts even if they were judging *their own* side-effect actions, as opposed to those of another agent. In other words, when judging the side-effects that they themselves bring about, are subjects still more inclined to judge those effects as intentional if they are morally bad than if

they are morally good? While the Intentionality Asymmetry predicts that they would, recent experimental research on this very question suggests otherwise.

Feltz et al. (2011) and Wellen & Danks (2012) independently carried out a series of innovative experiments involving real, rather than hypothetical, actions to address both of the foregoing issues. The results of the experiments carried out by these two groups of researchers are consistent with each other and fatal to the conventional understanding of the Knobe effect. On the first question, both groups of researchers found substantial differences between how subjects judged the intentionality of side-effect actions in *non-hypothetical situations* versus how they judged the intentionality of side-effect actions in *hypothetical thought-experiments*. Wellen & Danks (2012), in particular, found that when subjects are asked to judge (as observers) the intentionality of morally asymmetrical side-effect actions in non-hypothetical situations, there is no difference between the judgments concerning negative versus positive side-effects. In other words, they found that when subjects judge (as observers) in non-hypothetical situations, the Knobe effect vanishes. On the second question, both groups of researchers again produced results that were consistent with each other. In particular, both groups found that when subjects in non-hypothetical situations are asked to judge their own side-effect actions as opposed to those of another agent, a reverse Knobe effect appears. For example, Feltz et al. (2012, p. 682) found that "actors tended to judge harmful side-effects as less intentional than helpful side-effects." Wellen & Danks (2012) do not see this result as indicating anything of significance for the understanding of folk psychological concepts; rather, they believe that this reverse Knobe effect should be situated theoretically within a more general class of actor-observer biases.

The results of these experiments present a major challenge to the conventional understanding of the nature and significance of the Knobe effect. While a large number of competing theories have been offered to account for the Knobe effect, almost all researchers have assumed that the effect reveals something important about how people actually use the concept of intentional action. In defending a competence theory, Knobe in particular has championed the view that the effect in question is a

source of great insight into the competency underlying people's use of folk psychological concepts (Knobe & Mendlow, 2004; Pettit & Knobe, 2009; Knobe, 2010). Of course, not everyone has agreed with Knobe on this point. Other researchers, notably Adams & Steadman (2004a, 2004b) and Nadelhoffer (2004a, 2004b, 2006), have seen the Knobe effect asymmetries as the result of either pragmatic features of intentional language or certain psychological biases. However, even these critics of Knobe's competence theory agree that the Knobe effect reveals something important about the way people actually use folk psychological concepts, including the concept of intentional action. However, if Feltz et al. (2011) and Wellen & Danks (2012) are correct, then the Knobe effect does not reveal anything important about how people actually use the concept of intentional action, for their key finding is that when people are asked to judge the actions of others in non-hypothetical contexts, the Knobe effect asymmetries simply do not arise. Thus, these experimental results provide good reason to believe that, contrary to the conventional view, the Knobe effect is in fact an experimental artifact with no significance for the understanding of the concept of intentional action or other folk psychological concepts.

5. Conclusion

The Knobe effect is widely considered to be one of the first and most important findings in the field of experimental philosophy. While a great deal has been written on the Knobe effect in the past decade, most of this research has been concerned with either explaining or extending that effect. Comparatively little attention has been given to the more fundamental question of defining what exactly the Knobe effect is. In the foregoing I have addressed this definitional question, and I have argued that the Knobe effect is afflicted by at least three ambiguities that have received insufficient attention.

First, I have shown that "the Knobe effect" has both a narrow and a broad interpretation. In its narrow sense, the term refers to an effect that moral considerations allegedly have on ascriptions of intentional action; in its broad sense, the term refers to an effect that evaluative considerations in general allegedly have on all folk psychological ascriptions. As we have

seen, different researchers use "the Knobe effect" in these different senses without any clear acknowledgement of the ambiguity. Furthermore, I have argued that even once the ambiguity is made clear, researchers are left with a dilemma, for if they investigate the Knobe effect in its more restrictive sense they run the risk of missing the deeper or more general principles at play in the phenomenon under investigation. On the other hand, if they attempt to examine the Knobe effect in its more general sense, they are immediately confronted by the fact that "the Knobe effect" in this more general sense lacks any clear definition, and it is hard to see how researchers can go about explaining a phenomenon when they are not even clear on what it is that they are explaining.

Second, while the narrow reading of the Knobe effect (i.e. the Intentionality Asymmetry) does have a clear definition that generates testable predictions, I have shown that some of these predictions have turned out to be false. That the canonical expression of the Knobe effect is false is a point that has clearly received insufficient attention. And while Knobe did at least temporarily adjust his characterization of the Intentionality Asymmetry in light of this recalcitrant evidence, many researchers, including Knobe, continue to write as though the Intentionality Asymmetry is true. Moreover, Knobe's modified version of the Intentionality Asymmetry introduces a second ambiguity into the Knobe effect, one that applies specifically to the narrow interpretation of that term. As a result of Knobe's modification, the following question arises: Is the Intentionality Asymmetry a hypothesis about the effect that moral judgments allegedly have people's ascriptions of intentional action or is it rather a hypothesis about an alleged effect of non-conscious reactions to norm violations? While the former interpretation is by far the more common view, the Intentionality Asymmetry on this interpretation is strictly false and in need of no explanation.

Finally, I have pointed to a fundamental question that has been largely overlooked in the ongoing research concerning the Knobe effect. Assuming that the foregoing ambiguities could be resolved and that "the Knobe effect" could be defined in a way that is meaningful and true, would it reveal something important about how people actually use folk psychological concepts in real life or would it rather tell us something about

how people use these concepts in hypothetical thought-experiments, such as Knobe's classic experiment on the Chairman Scenario? While the vast majority of researchers and commentators have assumed that the former is the case, recent experimental research suggests that the truth is rather that the Knobe effect may be an experimental artifact with virtually no significance for the understanding of the concept of intentional action or other folk psychological concepts.

In general, if the arguments contained in this article are correct, then the Knobe effect is a troubled philosophical doctrine. Before any further work is done on explaining the Knobe effect, researchers working in this field ought to devote more time and effort to clarifying what it is they are attempting to explain and determining whether or not it really is in need of an explanation.[2]

References

Adams, F. & Steadman, A. (2004a). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64(2), 173–81.

Adams, F. & Steadman, A. (2004b). Intentional actions and moral considerations: Still pragmatic. *Analysis*, 64(3), 268–76.

Beebe, J. R. & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind and Language*, 25(4), 474-498.

Cova, F. & Naar, H. (2012). Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology*, 25(6), 837-854.

Cushman, F., Knobe, J. & Sinnott-Armstrong, W. (2008) Moral appraisals affect doing/allowing judgments. *Cognition*, 108(2), 353–80.

Feltz, A., Harris, M. & Perez, A. (2011). Perspectives in intentional action attribution. *Philosophical Psychology*, 25(5), 637-687.

Guglielmo, S. & Malle, B. F. (2010). Can unintended side-effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, 36(12), 1635-1647.

Holton, R. (2010). Norms and the Knobe effect. *Analysis*, 70(3), 417-424.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*,

[2] I would like to thank three anonymous referees from this journal for helpful comments on an earlier draft of this article.

63, 190–193.
Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16, 309–324.
Knobe, J. (2004a). Intention, intentional action and moral considerations. *Analysis*, 64, 181–187.
Knobe, J. (2004b). Folk Psychology and Folk Morality: Response to Critics. *Journal of Theoretical and Philosophical Psychology*, 24 (2), 270-279.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130, 203–31.
Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31, 90–107.
Knobe, J. (2010). Person as scientist, person as moralist. *Behavioural and Brain Sciences*, 33(4), 315-329.
Knobe, J. & Burra, A. (2006). Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition*, 6(1-2),113–132.
Knobe, J. & Fraser, B. (2008) Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 2: The cognitive science of morality: Intuition and diversity* (pp. 441–448). Cambridge, MA.: MIT Press.
Knobe, J. & Mendlow, G. (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24(2), 252–58.
Leslie, A. M., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17, 421–507.
Levy, N. (2011). Neuroethics: A new way of doing ethics, *AJOB Neuroscience*, 2(2), 3-9.
Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, 23(2), 165–89.
Mallon, R. (2008). Knobe vs. Machery: Testing the trade-off hypothesis. *Mind and Language*, 23(2), 247–55.
Nadelhoffer, T. (2004a). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology*, 24(2), 196-213.
Nadelhoffer, T. (2004b). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, 24(2):259-269.
Nadelhoffer, T. (2005). Skill, luck, control, and folk ascriptions of intentional action. *Philosophical Psychology*, 18(3), 343–54.
Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some

problems for jury impartiality. *Philosophical Explorations*, 9(2), 203–20.
Nichols, S. & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind and Language*, 22(4), 346–65.
Pettit, D. & Knobe, J. (2009) The pervasive impact of moral judgment. *Mind and Language*, 24(5), 586–604.
Phelan, M. & Sarkissian, H. (2008). The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138(2), 291–98.
Pinillos, N.A., Smith, N., Nair, G.S., Marchetto, P. & Mun, C. (2011). Philosophy's new challenge: experiments and intentional action. *Mind & Language*, 26(1), 115-139.
Ulatowski, J. (2012). Act individuation: an experimental approach. *Review of Philosophy and Psychology*, 3(2): 249-262.
Wellen, S. & Danks, D. (2012). Actor-observer asymmetries in judgments of intentional actions. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 2523-2528). Austin, TX: Cognitive Science Society.
Wright, J. C. & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind and Language*, 24(1), 24–50.