FOLK PSYCHOLOGY AND MORAL EVALUATION

JULIE YOO

Abstract: Assessments of an action done intentionally, as we might expect, influence judgments of moral responsibility. What we don't expect is the converse – judgments of moral responsibility influencing assessments of whether an action was done intentionally. Yet this is precisely how people decide, according to Knobe and Mendlow and Nadelhoffer. Known as the biasing effect, I evaluate whether the studies actually support it. I argue that the studies are at best inconclusive and that even if they demonstrated that people fall under the biasing effect, such tendencies ought to have no bearing upon philosophical analyses of the concept of intentional action.

Introduction

The concept of intentional action figures prominently in our understanding of agency. Drawing the distinction between intentional and unintentional action is crucial for the way in which we interpret and coordinate our social interactions. It is also indispensable to the concept of free will, as performing an action freely presupposes that one intentionally brings the action about. As we shall see, it is also crucial to our notion of moral responsibility.

Suppose Jill spills some boiling soup upon your lap, which causes you to experience considerable pain and discomfort. If you judged that Jill's spilling was not an intentional act, you would not hold her morally responsible for the spilling. You might be upset at what Jill did, understandably, but you would not blame her for it. Matters would be quite different, however, if you judged that her action was intentional. If you thought that Jill spilled the soup on purpose because she had it out for you, you would hold her culpable without hesitation. What is salient to note in the decision process in both scenarios is the order in which we make our judgments. We first determine whether the agent performed the action intentionally, then we gauge the agent's culpability accordingly. Moral evaluation comes at the heels of a prior intentionality ascription, not the other way around.[1]

[1] I use the term "intentionality" to mean "doing something intentionally," which is how the term is used in the present discussion, not how it is generally used, which refers the "aboutness" of contentful or representational mental states.

How, though, would you attribute intentionality when an action, performed in the service of a certain goal (fast service), concomitantly causes a foreseen side-effect (spilling soup) that the responsible agent neither desires nor intends? Suppose Jill owns the restaurant in which the unfortunate event takes place. She only cares about making a big profit, so she hurries and hustles to serve as many customers as she can, all the while knowing that her service will be poor and maybe even dangerous. In her frenzy, Jill spills your soup, with expected consequences. Would it be right to say that Jill *intentionally* spilled the soup?

Studies recently conducted by Knobe (2003a, 2003b, 2004), Knobe and Mendlow (2004) and Nadelhoffer (2004a, 2004b, 2004c) have been claimed to show that people sometimes judge in ways that would lead them to declare that Jill did indeed intentionally spill the soup. According to the studies, when the side effect is bad or its instigator blameworthy, people tend to judge that the agent intentionally brings it about, even though they are told that the agent has no desire for the side effect to occur. This is what Nadelhoffer calls the *biasing effect*, the influence of moral considerations upon attributions of intentionality.

If the biasing effect is real, then it reveals an interesting feature about people's conception of intentional action. Intentional action, we intuitively take it, is based upon the following (necessary) conditions:

(A) An agent S intentionally does F only if:
   i. S believes that she can bring F about,
   ii. S has a desire to bring about F,
   iii. S has control over the way in which F is brought about.

The intuitive model thus minimally involves knowledge, motivation, and control (Mele and Sverdlik 1996, Davidson 1963, Goldman 1970). But, as we shall soon see, in cases that elicit the biasing effect, some of these components are arguably absent. This means that we must either reconsider the components stated in the intuitive account of intentional action, or we have to reinterpret the studies to explain away the biasing effect. Indeed, if the biasing effect were to influence philosophical analyses of the concept of intentional action, then it would pose serious problems for some of the most well developed analyses – that an agent does F intentionally only if she *intends* to do F (Searle 1983), that she must *try* to do F (O'Shaugnessey 1973), or that she has a *reason* to F (Davidson 1963). Each of these critical conditions – intending to do F, trying to do F, having a reason to F – arguably fail to get met in those cases involving foreseen yet

undesired side effects.[2] So if people nonetheless attribute intentionality to foreseen yet undesired side effects, then the above analyses either call for rejection or, again, the data requires reinterpretation so that the biasing effect can be explained away.

For all we know the biasing effect is real. And if it is, then the model of the folk concept of intentional action will have to take the biasing effect into consideration. I shall argue, however, that the studies do not present a clear picture of how the folk go about attributing intentional action, and thus do not unequivocally substantiate the biasing effect. In a separate series of studies conducted by Sverdlik, it was found that people sometimes attribute intentionality in ways that are independent of judgments about the moral goodness or badness of the side effect or the integrity of the agent. If by *biasing effect* we mean strictly the influence of moral considerations upon intentionality ascriptions, then Sverdlik's results pose a serious threat to the hypothesis that our intentionality ascriptions fall under the sway of the biasing effect. This complicates considerably the project of delineating the conditions governing the application of the concept of intentional action.

Here is how my discussion will proceed. After presenting the results of the experiments, I will examine several models that attempt to capture the pattern of the folk understanding of intentional action. I will then conclude with a discussion about the normative implications of the biasing effect.

The Experiments and Their Results

Demonstrating what we're calling the biasing effect amounts to showing that attributions of intentionality are a function of a person's (tacit) appeal to normative considerations involving the agent and her actions. That said, there are several variables that characterize the biasing effect. The following, I think, are particularly relevant.

1. Which kinds of scenarios elicit the biasing effect? In side effect cases only or in side effect as well as non-side effect cases?

[2] This gets into thorny issues concerning how to individuation actions. If we individuate them "coarsely," so that the side effect F of an larger action G is a part G, then having a reason for G entails having a reason for F. *Mutatis mutandis* for the other analyses.

2. Is the bias effect weighted towards negative considerations or does it occur just as much with positive considerations?

3. What is the source of the biasing – the person's judgment about the blameworthiness/praiseworthiness of the agent, the goodness/badness side effect itself? Are other factors involved?

4. Is the biasing effect limited only to moral judgments or is it sensitive to non-moral judgments as well?

Which kinds of scenarios elicit the biasing effect?

The scenarios or vignettes depict an agent who intentionally performs an action with an eye to achieving a certain goal. Most of the scenarios involve actions that cause a side effect the agent both foresees but does not desire or intend to bring about. So we have vignettes involving agents who have a primary interest in starting a new corporate strategy that causes the side effect of harming the environment (Knobe 2003), helping a friend in a competition at the cost of compromising one's own chance of winning in the competition (Nadelhoffer 2004a), and increasing overall profit but decreasing sales in NJ (Knobe and Mendlow 2004). For convenience, I'll tag each of these vignettes as *harm environment*, *lose competition*, *decrease NJ sales*, respectively. In each of the experiments invoking these vignettes, the experimenters recorded a positive correlation between the negative side effect and the agent's intentionally bringing them about. In *harm environment*, 82% of the participants said that the agent harmed the environment intentionally; in *lose competition*, 55% of the participants said that the agent intentionally lost the competition; and for *decrease NJ sales*, 75% said that the agent intentionally decreased sales in NJ. (The numeric data make the experiments appear more precise than they really are; I think it is best to think of the figures as very rough representations of the tendencies of a small subpopulation of the folk.)

But some scenarios do not involve side effects. Nadelhoffer offers evidence of the biasing effect in non-side effect cases as well. In this amusing vignette, which we'll call *cause explosion*, an agent decides to cause a meltdown at the nuclear power plant from which he was fired. The reactor is controlled by a certain computer that can make the reactor explode, but only

if a 10-digit code is punched in, which the agent has no knowledge of. On a lark, the agent punches in a string of numbers that happens to be the right code. The reactor explodes and thousands of people are killed. Notice that the explosion was not a mere side effect but the primary goal towards which all the agent's other actions were aimed. Yet, even though the agent did not have full control over process that could lead to his nefarious goal, as the chances of punching in the right ten digit code are astronomically small, 83% of the participants said that the agent intentionally caused the explosion. Now, one might wonder whether the agent's moral deficiency had any role in guiding the intentionality attribution, since that is what the biasing effect is all about. To demonstrate the relevance of the judgment about the agent's moral integrity, Nadelhoffer tested out a comparable but slightly altered scenario engaging a morally neutral agent, and he found that 33% percent of the participants said that the agent intentionally brought about his primary goal, thereby lending prima facie support to the biasing effect in *cause explosion*. Thus, if all of the above studies are accurate, the biasing effect occurs in side effect and non-side effect cases alike.

Is the biasing effect weighted towards negative considerations?

In Knobe 2003, Knobe argued that the bias effect is indeed weighted towards negative considerations: people were much more like to attribute intentionality when the side effect caused by the agent was bad than when the effect was good. But, as Nadelhoffer argues, the weighting is falsely based upon misleading wording of the good side effect vignettes. Moreover, Nadehoffer continues, reactions to certain *non-side effect* scenarios actually demonstrate that the weighting is fairly even – negative and positive considerations can equally influence intentionality attributions.

To show this, Nadelhoffer appeals to a variation on *cause explosion* , which we'll call *prevent explosion*. In *prevent explosion*, an agent is again responsible for operating a computer that controls a nuclear reactor in danger of exploding, which can only be prevented by punching in a 10-digit code that the agent is clueless about. But on a lark, the agent punches in a string of numbers that happens to be the right code, and thereby prevents the explosion. 73% of the participants said that the agent intentionally prevented the explosion. On the basis of the fact

that most people would regard preventing a fatal explosion praiseworthy, Nadelhoffer explains the high intentionality attribution to the sensitivity people have towards the positive considerations surrounding the scenario. From this, Nadelhoffer concludes that negative as well as positive considerations can each underlie the biasing effect.

What is the source of the biasing?

Here we encounter conflicting hypotheses. Whereas Nadelhoffer targets judgments about the responsibility of the agent – her praiseworthiness or blameworthiness – Knobe and Mendlow argue that it is rather judgments about the value of the side effect – its goodness or badness. The two judgments can come apart. We can have a case where a side effect is bad but the agent who causes it is not blameworthy, and we can have a case where a side effect is not bad but the agent is blameworthy nonetheless. To illustrate the former, consider stepping on someone's foot by mistake; stepping on someone's foot is a bad thing, but if you did it by mistake, surely you cannot be blamed for it. For the latter, consider an attempted murder that does not (thankfully) succeed; not being killed is a good thing, for sure, but an agent who attempts to kill is certainly blameworthy.

To demonstrate that it is judgments about the value of the side effect rather than about responsibility, Knobe and Mendlow asked participants to assess a scenario in which a CEO chooses to adopt a strategy that will increase sales overall but decrease sales in a certain locale (NJ). When asked whether the agent intentionally brought about the side effect of decreasing sales in NJ, Knobe and Mendlow found that the subjects tended to agree with the intentionality attribution but withhold attributing blame to the CEO. So this study demonstrates a stronger positive correlation between intentionality attribution and the badness, and hence the value, of the side effect, than with the agent's blameworthiness. From this, Knobe and Mendlow surmise that the driving force behind the biasing effect is judgments about the value of the side effect itself, not the responsibility of the agent. Nadelhoffer, however, rightly points out that to the extent that the present concern lies with explaining the underpinnings of the moral considerations regarding intentionality attribution, the *decrease NJ sales* scenario is infelicitous since the side effect is not *morally* loaded. Thus, it is possible that in cases involving proper morally loaded

side effects, it is the moral responsibility of the agent takes the leading role in guiding intentionality attributions, and that the blameworthiness or praiseworthiness can be relevant after all.

The dispute concerning the source of the biasing effect may, however, be moot if the results of Sverdlik's studies are reliable. On the scenarios drawn by Sverdlik an agent is morally responsible for a bad effect but arguably does not produce it intentionally. In what will be called *wake neighbor*, an agent wants to mow his lawn on a weekend morning. He feels regretful about mowing his lawn because it will produce a noise so loud that it will inevitably wake up his neighbors, but he does it anyway. For this scenario, 77% of the subjects assigned blamed for waking the neighbors, but only 24% said that the agent woke his neighbors intentionally. This suggests that perceptions about the responsibility of the agent do not drive one's judgment about the intentionality of his action. So Nadelhoffer's hypothesis again is disconfirmed, but so is the one advanced by Knobe and Mendlow. It is safe to assume that people take being woken early on a weekend morning to the sound of a lawnmower a bad thing, but given that only 24% of the subjects attributed intentionality, it appears that perceptions about the value of the side effect do not drive judgments about intentionality either.

The separation between responsibility, side effect value, and intentionality is again illustrated in people's reactions to the scenario we can call *cause pain*. In this scenario, a dentist decides she must perform a surgical procedure on a patient that will generate the unfortunate side effect of causing pain. Feeling very regretful about causing pain, she goes out of her way to prevent as much pain as possible. When asked about responsibility, only 26% of the participants assigned blame for causing pain. However, 41% claimed that the dentist caused the pain intentionally. These results suggest a couple of things. One is that people can withhold intentionality even as they assign blame, as *wake neighbors* appears to demonstrate; and people can attribute intentionality even as they withhold blame, as *cause pain* appears to demonstrate. As Sverdlik cautiously summarizes, "subjects do not simply use either their belief that a side effect is bad, or that the agent is responsible for it, to determine their view about the intentionality of its production." (2)

But if neither perceptions of the agent's responsibility nor of the value of the side effect play a significant role in the biasing effect, what does? We aren't left with many options. One might try to make a case for judgments about the level of care exhibited by the agent regarding

the bad side effect she brings about, as Sverdlik sometimes seems to suggest, but we know that this alone cannot fully explain the biasing effect, since a comparable percentage of subjects attributed intentionality both to a caring agent and an uncaring agent.

Can non-moral judgments influence attributions of intentionality?

It is possible that considerations other than just moral judgments play a role in guiding our attributions of intentionality. While Nadelhoffer is right to complain that Knobe and Mendlow's decrease NJ sales does not have a morally significant side effect, and that their dismissal of Nadelhoffer's hypothesis about the relevance of judgments concerning responsibility is premature, Knobe and Mendlow could still argue that their study establishes something interesting, namely, the relevance of *prudential* norms in the explanation of the biasing effect (assuming that it is prudent not to decrease sales). Perhaps the biasing effect is even more "expansive," to use Nadelhoffer's term, in that intentionality attributions may be sensitive to an array of different types of normative judgment, not to moral judgment alone.

It is even possible that non-normative considerations enter into the picture as well. As Sverdlik hazards to guess, one's views about the agent's level of care concerning the bad side effect she will produce may also play a role in guiding intentionality attribution. In *harm environment*, the agent is callously indifferent to the harm his decision will cause the environment, and in the trial using this scenario run by Knobe, 82% attributed intentionality to the agent. In *wake neighbors*, on the other hand, the mower is described as feeling very reluctant to wake his neighbors, and only 24% attributed intentionality. The less the agent seems to care or feel badly about causing the bad side effect, the greater the tendency to judge that the effect is caused intentionally. But while the comparison of these two scenarios – *harm environment* and *wake neighbors* – appears to support this generalization, the support is tentative at best. Repetition of the trials generated some surprisingly different results that undermine the generalization: in a different run of the *harm environment* scenario, 44% attributed intentionality (way down from 82%), and in a different run of the *wake neighbors* scenario, 45% attributed intentionality in a different run of the *wake neighbors* scenario (up from 24%). The variation in results across trials should give one pause (more on this below), but one thing is clear: certain

comparisons show that one's view about how much an agent cares has little effect upon one's intentionality attribution.

As is now obvious, the results do present a clear picture of how the folk attribute intentionality. It is hard to know what to make of the data, especially because much of it seems to have been gathered under sub-optimal experimental conditions. There are three ways in which the experimental conditions could be improved. First, the scenarios to which people are asked to react must engage morally loaded situations, if by "biasing effect" we mean the influence of *moral* considerations upon the attribution of intentionality. The side effect must be morally bad or good; or the agent must be morally praiseworthy or blameworthy. Not all the scenarios, however, satisfy this condition. The *decrease NJ sales* vignette, for example, is a case in point. As Nadelhoffer rightfully complains, an agent's decreasing sales in a certain region is not a morally significant action. Thus the results garnered from reactions to that vignette cannot be used to establish anything about the biasing effect in the way defined above.

Second, the wording of the questions regarding the scenarios and of the scenarios themselves must be unambiguous. Otherwise, the results will not be reliable. However, certain crucial terms, such as "blame," or "responsibility," are ambiguous in that they have a moral sense and a non-moral causal sense. Sverdlik rightly points this out in his discussion. When people "blame" the agents for causing the bad side effects, they may have in mind the non-moral sense, thereby giving results that do not support the biasing effect in the way originally defined.

And third, the subjects need to be asked the same set of questions for each scenario to yield consistent results. In one of the studies conducted by Sverdlik, for instance, one group was asked three questions regarding *wake neighbor* – the first about whether the mower knew that the lawn mowing would wake the neighbors, the second about intentionality behind waking the neighbors, and the third about the responsibility of the mower. But on a separate trial using the same scenario, the group was asked only the last two questions. This means that the subjects across the different trials were not primed in the same ways. Failure to prime the subjects in the same ways may also explain why the figures for intentional attribution differed so significantly across separate trials of one and the same scenario. For instance, while Knobe found that 82% of the subjects attributed intentionality to the agent in *harm environment*, Sverdlik found that only 60% judged likewise on one rerun and 44% on another. Such variation in the results for a repeated trial ought to give one pause.

I think more illuminating results can be obtained if the subjects are primed into considering the three intuitive components of the concept of intentional action – the cognitive, the motivational, and the aspect regarding control –before they encounter the question about intentionality and the one about responsibility. By making the subjects mindful of these elements, we might be able to control more effectively for a tendency to over-attribute intentionality. Thus, there would be a question about whether the agent was fully aware of the inevitability of the side effect, a question about whether the agent wanted the side effect to occur, and whether the agent had control over the way in which the side effect was brought about.

A Model of the Concept of Intentional Action

One of the aims of studying the biasing effect is to construct a *model* of the concept of intentional action. I take the term, "model," from Nadelhoffer 2004b to mean a reconstruction of the conceptual components of a concept on the basis of its actual pattern of application. Constructing a model of a concept involves laying out its necessary and sufficient conditions as they are judged by the folk. Thus, a model of the concept of intentional action must draw its necessary and sufficient conditions from the pattern of actual intentional attribution represented by the data collected in the experiments. As we saw in the studies conducted by Sverdlik, it is neither clear whether the biasing effect includes more than moral considerations or whether it even occurs at all. So in all likelihood, no model can be constructed on the basis of what we have so far. Nonetheless, it would be a useful exercise to see how an account of intentional action could accommodate the biasing effect. In this section, I shall explain the motivation behind each of the conditions outlined in (A), the intuitive account of intentional action, presented at the beginning of this paper, and assess it in light of the biasing effect. I will then examine Nadelhoffer's latest model and assess how well it accommodates the intuitions behind the intuitive account and the biasing effect. I will end the section with a recommendation for a better model.

Recall the conditions on the intuitive account.

(A) An agent S intentionally does F only if:
  i. S believes that she can bring F about,
  ii. S has a desire to bring about F,
  iii. S has control over the way in which F is brought about.

Each of these conditions has strong intuitive appeal; if presented to the folk, the folk would mostly likely endorse the conditions.

Consider the first condition. Suppose Jill spills some soul on your lap, so clearly there was soup in the bowl. But suppose that Jill didn't know that there was any in the bowl, as she was convinced that the bowl was empty. One would be hard pressed to say that she spilled the soup intentionally, for without the belief that one's actions could lead to a certain consequence, one cannot form a plan to achieve the goal. Without a plan there is no intentionality. So Jill's act of spilling the soup cannot be considered intentional, and this would still true even if Jill wanted to spill soup on your lap, and everything about the situation was set up so that she had control over whether and how the spilling took place. The agent must believe that her plan is capable of achieving her goal in order for her action to count as intentional. This is the rationale behind the first condition.

To see rationale behind the second, suppose Jill abhors the idea of spilling soup, but Jack has a gun pointing at her head, threatening to kill her if she doesn't spill soup on your lap. Then even if she believes that she can spill the soup and she does indeed have control over each step of the process leading to the spilling, we wouldn't say that she spilled the soup intentionally, because the element of being motivated to engage in the action was patently missing.

The third condition is motivated by considerations concerning deviant causal paths. Suppose that Jill wanted to spill soup and believed that by tipping the bowl she could achieve her goal, but right at the moment the bowl begins to tip, another waiter bumps into the bowl, causing the bowl to fall onto your lap. Again, the spilling cannot be considered an intentional act on the part of Jill since it was not brought about by Jill in the right way. The agent has to have control over the process by which the goal is brought about. Without that control, it is just a matter of luck that the goal is achieved, and the point of calling an action intentional is precisely to rule out the element of luck.

The biasing effect, however, calls into question each of the conditions. Nadelhoffer's non-side effect scenarios – *prevent explosion* and *cause explosion* – lack both the condition concerning knowledge and the condition concerning control. Recall that the agent in these scenarios does not know how to achieve his goals, since he can achieve them only if he punches in a certain ten-digit code, which he does not know. Given the extremely low chance that he can guess the code correctly, the agent has very little reason to believe that he can achieve his goal. Furthermore, given that the realization of the goal is a matter of chance, the agent has very little control over bringing about the goal. Those subjects exhibiting the biasing effect attributed intentionality to the agent in spite of his failing to satisfy the two necessary conditions. The feature of non-side effect scenarios is very interesting, but I will put them to the side and focus upon a very curious feature of side effect scenarios.

The most jarring things about the biasing effect is the attribution of intentionality in the absence of motivation. The side effect scenarios target the condition concerning motivation. Here, the agent decides to go through with an action she knows will bring about a bad side effect, which she has no desire to bring about, and would even prefer not the bring about. And yet those under the sway of the biasing effect attributed intentionality to the agent anyway. Given the intuitive pull of the motivational aspect of intentional action, how can it be ignored? It seems that even Nadelhoffer, in his most recent model of intentional action, has failed to capture this crucial detail. Here is his proposal:

> (B) In cases involving side effects and either moral badness or goodness, an agent will be judged to have intentionally brought about a side effect, *y*, by performing some action, *x*, only if the following conditions are met:
>
> (1) The agent (a) wants to do *x,* (b) wants to bring about *y* by doing *x*, or (c) both (a) and (b).
>
> (2) The agent knows that doing *x* will likely bring about *y*.
>
> (3) If *y* is bad, the mental states (i.e. desires, beliefs, intention, etc.) of the agent must be such that the agent is blameworthy.
>
> (4) If *y* is good, the mental states (i.e. desires, beliefs, intentions, etc.) of the agent must be such that the agent is praiseworthy.

It is not clear how this partial model (partial because it only gives a few necessary conditions) is supposed to capture the fact that the agent does *not* want the side effect to come about. Remember that the biasing effect is the attribution of intentionality to a foreseen yet *undesired* side effect; in fact, the undesired part is what makes side effect cases interesting, for if the agent wanted the side effect to occur, then there would be no controversy about whether it is appropriate to attribute intentionality. If intentionality is attributed even in the face of flouting the motivation condition – (b) of (1) – then (B) cannot be the correct model. A better model needs to take this into consideration.

Our intuitive notion of intentional action includes the motivational component, but the biasing effect of moral considerations jettisons this condition, making a model of intentional action look more like this:

(C) An agent S intentionally does F only if:

i. S believes that she can bring F about,

ii. S has a desire to bring about F *unless S is morally blameworthy for F or if F is bad*,

iii. S has control over the way in which F is brought about.

iv. *If S is morally blameworthy for F or if F is bad, then S has no desire to bring about F.*

(C) is just (A) but with added conditions (represented by the italicized text). It does a better job than (B) of accounting for the fact that judgments made under the biasing effect countermand the motivational condition. In addition, (C) is more general in that it applies both to side effects and the actions that generate them. (C) is admittedly a far cry from representing an accurate model: for one, it is explicitly neutral about the source of the biasing effect, and so may be too broad, and for another, it does not acknowledge possible non-moral but otherwise normative influences, and so may be too narrow. It also illegitimately rules out the possibility of a blameworthy agent who desires to bring about a bad effect. Perhaps we can fiddle with (C) to fix these shortcomings, but I leave that for another occasion.

At this stage, I wish to consider a more pressing issue, and that is the lack of unanimity in any given experiment testing for the biasing effect (assuming that the biasing effect is indeed real). A fact about the results true of all of the studies is that some subjects withheld their

attributions of intentionality, even as their peers, who read exactly the same scenarios, fell under the sway of the biasing effect. Now why was this? Why was there no unanimity? One possibility is that those who are subject to the biasing effect operate with a more liberal conception of intentional action, more liberal in that it is defined by fewer conditions, whereas those who held out operate with a more stringent conception. Let us say that the following correspond to the two different conceptions of intentional action:

robust intentionality = knowledge, motivation, control
thin intentionality = knowledge, control

On one hypothesis, those who undergo the biasing effect operate with thin intentionality whereas the others operate with robust intentionality. On another, everyone operates with robust intentionality as their default conception, except when moral considerations arise as they do in side effect cases, in which case those who undergo the biasing effect switch to operating with thin intentionality while the others do not shift out of their default concept. Both hypotheses would explain the curious fact that some attribute intentionality while others withhold it to one and the same scenario. It would also explain why the "biased" attribute intentionality even in the face of knowing that the agent has no desire, interest, or pro-attitude towards the action they are said to intentionally bring about. Evidently, these subjects either never acknowledge a motivational component in connection with intentional action or they do acknowledge it but countermand it when side effect cases crop up.

Confirming or disconfirming either of these hypotheses would bring us a lot closer to a model of the concept of intentional action. Now, the project of coming up with such a model is certainly an interesting one from the point of view of social psychology. But does it bear upon philosophical analyses intentional action? I explain why it does not in the next section.

The Philosophical Implications of the Biasing Effect

So what is the value of an empirical model – a reconstruction of a folk definition – of a concept vis-à-vis the conceptual analyses of that concept? Traditionally, the aim of a conceptual analysis is to disclose the *nature* of the phenomenon the concept picks out, and this involves considering a wide range of counterfactual situations in which the concept applies, situations that are sometimes extremely far-fetched, in order to specify exhaustively the necessary and sufficient conditions of the concept. Although philosophers today are less sanguine about fulfilling the aim of conceptual analyses than their early modern and ancient Greek forebears, conceptual analysis is still a large part of what philosophers do today, and is even fashionable in certain circles (Chalmers 1996, Jackson 1993). An empirical model of a concept is a representation of patterns of actual concept application, and does not purport to reveal the nature of the phenomenon the concept picks out. To the extent that philosophy is engaged in revealing the nature of things, empirical models of their corresponding concepts do not deliver upon philosophical interests. Admittedly, there is no sharp division between an empirical model – folk definition – of a concept and a philosophical analysis of the concept, since conceptual analyses must draw from patterns of actual concept application. Nor is there a sharp division between philosophers and the folk, for that matter, since the folk include philosophers when philosophers have taken off their thinking caps, and philosophers include anyone who wishes to do philosophy. But folk definitions and conceptual analyses can certainly come apart, and when they do, conceptual analyses rule the day, not their competing empirical models.

Take, for instance, the concept of irony. The term, "ironic," is often misused by the folk who tend to use it to mean "amusing coincidence" (as in "My new neighbor who just moved in to my neighborhood used to be my neighbor when I lived in my old house. Isn't that ironic?") Were we to defer to the folk in defining the nature of irony, we would not necessarily get any closer to understanding what irony really is. This point is even more plain in the case of scientific concepts. There are, for instance, a sizeable number of people who believe that a whale is a fish, but no proper conceptual analysis of the concept of a whale would defer to the folk when it comes to fleshing it out. In the event that the biological taxonomy undergoes revision, it will certainly be for scientifically motivated reasons rather than to reflect the folk applications of the biological concepts.

Now, the difference between a conceptual analysis and empirical model of a folk psychological notion will not be as large as the difference just noted regarding natural

phenomena, since folk psychological notions are constituted by the conceptual practices of the folk, unlike natural phenomena. But to the extent that we wish to preserve the distinction between *correct* applications of a concept and *actual* applications of a concept, we need to be cautious about how we incorporate patterns of actual usage. As is usually the case with the application of concepts, the folk apply the concept when they shouldn't and fail to apply it when they should. The purpose of conceptual analyses is to correct for these errors of over-application and under-application, errors that get ironed out only after careful reflection upon a vast range of counterfactuals that are unlikely to have been fathomed by the folk. Thus, it is not obvious that folk judgments about concept application ought to constrain their conceptual analyses, even when the object of analysis is a folk notion. In the event that a conceptual analysis of a folk concept diverges from the intuitions of some of the folk, there is only one way to resolve the difference, and that is by getting the folk to explain their intuitions. But once the folk embark upon that endeavor, they are, in effect, engaging in the very philosophical enterprise that the empirical model was meant to supplant. Philosophy is inescapable if we want to get a grip to the nature of things.

References

Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.

Davidson, D. 1963. Actions, Reasons, and Causes. Reprinted in Davidson 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Goldman, A. 1970. *A Theory of Human Action*, Prentice-Hall, Englewood Cliffs, NJ.

Jackson, F. 1993. Armchair Metaphysics. In O'Leary-Hawthorne, J. and Michael, M. eds. *Philosophy in Mind*. Dordrecht: Kluwer. 23 – 42.

Knobe, J. 2003a. Intentional action and side effects in ordinary language. *Analysis* 63, 190 – 94.

_____. 2003b. Intentional action in folk psychology: an experimental investigation. *Philosophical Psychology* Vol.16, No. 2, 309 – 24.

_____. 2004. Intention, intentional action, and moral considerations. *Analysis*. 181 – 187.

Knobe and Mendlow. 2004. The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology. (this volume)

Mele, A. and Moser, P. 1994. Intentional Action. *Nous* 28: 39-68

Mele, A. and Sverdlik, S. (1996). Intention, Intentional Action and Moral Responsibility. *Philosophical Studies* 82: 265-87.

Nadelhoffer. T. 2004a. Skill, Luck, and Folk Ascriptions of Intentional Action. *Philosophical Psychology*.

_____. 2004b. On Praise, Side Effects, and Folk Ascriptions of Intentionality. (this volume)

_____. 2004c. Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow. (this volume)

O'Shaugnessey, B. (1973) Trying (as the Mental 'Pineal Gland'). *Journal of Philosophy* 70, 365 – 86.

Searle, J. 1983, *Intentionality*, Cambridge: Cambridge University Press.

Sverdlik . 2004. Intentionality and Moral Judgments in Commonsense Thought about Action. (this volume)