Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow

Thomas Nadelhoffer
Florida State University

In a series of recent papers both Joshua Knobe (2003a; 2003b; 2004) and I (2004a; 2004b; forthcoming) have published the results of some psychological experiments that show that moral considerations influence folk ascriptions of intentional action in both non-side effect and side effect cases.[1] More specifically, our data suggest that people are more likely to judge that a morally negative action or side effect was brought about intentionally than they are to judge that a structurally similar non-moral action or side effect was brought about intentionally. So, for instance, if two individuals *A* and *B* place a single bullet in a six shooter, spin the chamber, aim the gun, and pull the trigger, but *A* shoots a person and *B* shoots a target, people are more likely to say that *A* shot the person intentionally than they are to say that *B* shot the target intentionally—even though their respective chances of success (*viz.*, one-in-six) and their control over the outcome are identical in both cases. And while Knobe and I agree that our research creates difficulties for any analysis of the folk concept of intentional action that ignores the biasing effect of moral considerations, we disagree about how best to explain this effect.

I have suggested that the moral *blameworthiness* of an agent can influence folk intuitions about intentional action. In a recent response to my work, Knobe and Mendlow (2004) reject this claim on two separate grounds—one *a priori*, one empirical. By their lights, not only is my view conceptually confused, but it also allegedly fails to explain the results of a recent experiment they have conducted. On Knobe and Mendlow's view, it is

[1] See, also, Knobe and Burra (forthcoming) for a cross-cultural study of the folk concept of intentional action.

the moral *badness* of the actions or side effects, and not the blameworthiness of the agent, that explains the biasing effect that moral considerations have on our intuitions and judgments about intentionality. In this essay, I will respond to both of their criticisms.

In their reply to my work, Knobe and Mendlow suggest that I am committed to the position that people only regard side effects as intentional if the agents who bring them about are blameworthy—i.e. on their reading of my view, blame is a *necessary* condition of folk ascriptions of intentional action in side effect cases. And while it may appear that I endorse this view, nowhere have I actually made this claim. Nor do any of my remarks commit me to this position. Of course, I do frequently suggest that blame attribution helps explain the biasing effect that moral considerations have on lay persons' judgments about intentional action, but just because I claim that blame helps explain this biasing effect, it does not follow that I must believe that blame either *fully* explains this effect or explains it in all instances.

Indeed, my view is much more modest than the one Knobe and Mendlow ascribe to me. Rather than suggesting that blame is a necessary condition for folk ascriptions of intentional action, I merely claim that blame can—and sometimes does—act expansively on these ascriptions. And this, of course, leaves open the possibility that badness has an effect as well—something I would not deny. In any event, given that I am only committed to the view that people sometimes attribute blame to an agent before they ascribe intentionality to her actions or side effects, my position is less conceptually demanding than either the overstated view that Knobe and Mendlow attribute to me or their own competing view. After all, they claim that if we make attributions of blame before we make ascriptions of intentional action, then the concept of intentional action

necessarily becomes a "pointless mechanism." And since Knobe and Mendlow correctly believe that we have good reason to reject any view that entails that the mechanisms underlying folk psychology are not useful, they conclude that my view is to be rejected. On the surface, this line of reasoning seems intuitively plausible. But I am now going to show that the first premise Knobe and Mendlow use to motivate their argument is questionable. To see why, we must first examine their *a priori* objection to my view in more detail.

In order to show that blame cannot explain the biasing effect of moral considerations, Knobe and Mendlow reasonably assume that the folk concept of intentional action plays "some helpful role in people's lives." Having introduced this assumption, they then examine the two following possibilities—the first represents their view concerning the series of judgments involved in making moral evaluations and ascriptions of intentional action, the second represents my view:

(a) (1st) mental state attributions
(2nd) judgments about goodness or badness
(3rd) judgments about intentional or unintentional
(4th) attributions of praise or blame

(b) (1st) mental state attributions
(2nd) judgments about goodness or badness
(3rd) judgments about praise or blame
(4th) attributions of intentional or unintentional

By their lights, the main difference between these two models is that (a) allows folk ascriptions of intentional action to play a useful role—*viz.*, fixing praise or blame—whereas (b) does not. As they say:

> The second model is not compatible with the commonsense view that people invoke the concept of intentional action when they are determining whether or not to assign praise or blame. Instead, the model has people deciding whether or not to assign praise or blame before they have even determined whether or not the

> behavior was performed intentionally. Thus, this model attributes to the folk psychologist a seemingly pointless mechanism.

Thus, the crux of their *a priori* argument against my view is that I leave no room for the folk concept of intentional action to do any meaningful work in our daily lives. I am now going to show that this is not the case.

In order to undermine Knobe and Mendlow's objection, I need only show that the concept of intentional *can* have an important role to play even in cases where it is applied *after* a determination of blame has already been made. My proffered example may loosely be called a pseudo-evolutionary one. Imagine that long ago humans who were good at quickly detecting "harmers"—i.e. morally blameworthy individuals such as cheats, liars, thieves, rapists, murderers, and other scoundrels—were more apt to survive than those who did not. For humans living under these conditions, the best survival strategy would be to blame *first* and worry about exculpating or mitigating circumstances *later*.[2]

Given that this kind of survival strategy is at least possible—if not likely—it is clearly possible that our judgments about the blameworthiness of an action may come before our determination of whether the action was performed intentionally. More importantly, this would not undermine the usefulness of the concept of intentional action to the extent that this concept—along with a cluster of other closely related concepts such as purposely, knowingly, and accidentally—could still be used to amplify, verify, mitigate, or exculpate our antecedent attributions of moral responsibility. And even

[2] Given these conditions, one thing we might expect to find is a system of legal and moral responsibility that is fueled by strict liability—i.e. a system whereby you are guilty or blameworthy for a crime simply insofar as you were causally responsible for it, regardless of whether or not you meant to do it. Indeed, this is precisely what one finds when one looks at a number of early civilizations. It was only much later in history that an agent's intentions, desires, and beliefs started to serve as factors relevant to blame attribution.

though in these situations our notion of intentional action would admittedly not play its usual role of fixing blame, it would nevertheless have an important role to play—*viz.*, helping to ensure that our initial "harmer detection" reaction was not unjustified.[3] Thus, to the extent that it is possible for our concept of intentional action to play an important role in our everyday lives even in cases where our ascriptions of this concept occur after our attributions of blame, the *a priori* objection that we have been discussing fails.

We should now turn our attention to the empirical objection Knobe and Mendlow develop against my view. They start by pointing out that in cases involving *both* bad actions or side effects *and* a blameworthy moral agent, our respective positions yield identical predictions—*viz.,* that moral considerations will affect subjects' ascriptions of intentional action. Thus, they correctly suggest that new experiments are needed if we are to settle the disagreement between us. However, because they have mistakenly assumed that on my view blame is a necessary condition of folk ascriptions of intentional

[3] In "Culpable Control and the Psychology of Blame," Mark Alicke (2000) develops a model of blame attribution that fits quite nicely with the example I have just developed in response to Knobe and Mendlow's *a priori* objection. According to Alicke's Culpable Control Model of Blame (CCMB), "personal control judgments and blame attributions are influenced by relatively unconscious, spontaneous evaluations of the mental, behavioral, and consequence elements. Spontaneous evaluations are affective reactions to the harmful event and the people involved" (2000:558). On the CCMB, these spontaneous and relatively unconscious responses can be triggered by both the evidential factors such as personal control and other "extra-evidential factors" such as a person's appearance, reputation, social status, etc. Thus, Alicke suggests that judgments concerning personal control—and hence of moral blameworthiness—are unwittingly influenced by spontaneous affective reactions to the agents and actions involved. One way that spontaneous reactions influence our assessments of moral and causal responsibility is by altering perceptions of the evidence itself. When this happens, "observers who spontaneously evaluate the actor's behavior unfavorably may exaggerate evidence that established her causal or volitional control and de-emphasize exculpatory evidence" (2000: 566). Another way that these reactions affect observers' judgments is by engendering blame-validation processing that subsequently increases the observer's "proclivity to favor blame versus non-blame explanations for harmful events and to de-emphasize mitigating circumstances" (2000: 568-9). Thus, to the extent that the observer believes that the action in question is immoral, she will be inclined to look for explanations of the action that favor ascriptions of blame while at the same time overlooking explanations that do not. Owing to both spontaneous evaluations and blame-validation processing, observers tend to "over ascribe control of human agency and to confirm unfavorable expectations" (2000: 558). Minimally, Alicke's CCMB offers empirical support for my claim that it is possible for blame attributions to occur prior to ascriptions of intentional action. Moreover, it may also help explain the sort of moral biasing that Knobe and I have discovered in several of our recent experiments.

action in side effect cases, the study they run to disconfirm my view is not effective in settling the disputed question. Their reasoning goes as follows:

> [I]n cases where people regard the side-effect as bad and blame the agent for bringing it about, the two models make identical predictions. But in cases where people regard the side-effect as bad but do not blame the agent for bringing it about, the predictions diverge. If judgments of praise and blame are directly influencing people's application of the concept of intentional action, people should be overwhelmingly inclined to see these side-effects as unintentional. But if people's judgments about the goodness or badness of the behavior itself directly influence their application of the concept of intentional action, one would expect to find results much like those obtained in Nadelhofffer's experiments—with a substantial portion of subjects saying that the agent brought about the side-effects intentionally.

These remarks make it clear that they view the disagreement between us as purely disjunctive—i.e. either blame acts expansively or badness acts expansively. But as I have already suggested, this is inaccurate. The real debate between us is not about whether blame or badness fully explains the biasing effect that moral considerations have on folk ascriptions of intentional action, but rather, whether or not blame can partly explain this biasing effect in a way that does not at the same time render the concept of intentional action useless. Consequently, the results of the experiment that Knobe and Mendlow marshal against my view fail to get at the heart of our disagreement.

Nevertheless, given that Knobe and Mendlow view the debate between us in the disjunctive manner I have just described, they understandably conclude that the best way to test our respective models would be to run an experiment with a vignette that has a bad side effect but that does not have a morally blameworthy agent. So, that is precisely what they do. The subjects of their "mini-study" were 20 people spending time in a Manhattan park—each of whom received a vignette that involved a CEO who knowingly adopts a new business plan that has the negative side effect of decreasing sales in New Jersey.

Not surprisingly, to the extent that the CEO adopted the business plan knowing full well that it would decrease the sales in New Jersey, 75% of the subjects said that she intentionally brought this side effect about. Moreover, very few of the subjects judged that the CEO deserved to be blamed for decreasing sales in New Jersey. Knobe and Mendlow take this result to show that their view—i.e. that badness, and not blame, explains the biasing effect of moral considerations in side effect cases—is correct. But as I am now going to show, this conclusion is too hasty.

First, Knobe and Mendlow's CEO vignette is problematic to the extent that decreasing sales in New Jersey is not really a *moral* consideration at all. In fact, we have good reason to doubt that the subjects perceived the badness of this side effect in the way than Knobe and Mendlow assume. After all, to the extent that the subjects judged that the CEO was *not* blameworthy for *intentionally* decreasing sales in New Jersey, we can draw one of two conclusions. On the one hand, perhaps subjects did not judge that this side effect was very bad. Of course, this would explain why they did not think the CEO deserved to be blamed, but then the badness of the side effect would lose the explanatory force that Knobe and Mendlow attribute to it. On the other hand, perhaps the subjects did think it was bad to decrease sales in New Jersey, but they did not think that it was *morally* bad. This would also explain why they did not judge that the CEO deserved to be blamed, but it would thereby undermine the vignette's effectiveness against my view. Either way, it would have been nice if Knobe and Mendlow had asked the subjects to give a badness rating for the side effect just for good measure.

Leaving this first problem with their experiment aside, the second point I want to make is that to the extent that I readily admit that badness can—and often does—act

expansively on ascriptions of intentional action, I am not surprised by their findings. Indeed, given that I believe that blame influences people's judgments about intentionality, I am seemingly committed to the view that badness does so as well—otherwise, I would be forced to assume that there are blameworthy cases that strangely do not contain any morally bad features. And since I am hard pressed to imagine what these cases would look like, I assume that wherever you find a blame ascription, you will likely find an antecedent judgment about the moral badness of the action or side effect. Given that this is so, I entirely agree with Knobe and Mendlow's suggestion that judgments about badness can precede ascriptions of intentional action.

Finally, as we have already seen, to the extent that I do not suggest that blame is a necessary condition for the occurrence of moral biasing, the results of Knobe and Mendlow's CEO experiment fails to say anything at all about the truth of my view. And by my lights, not only do their results fail to falsify my view for the three reasons I have just suggested, but they fail to verify their own view as well. To see that this is the case, I am now going to briefly develop a potential alternative explanation of their data. Keep in mind, Knobe and Mendlow take their data to show that the badness of the side effect—and not the blameworthiness of the agent—explains the subjects' ascriptions of intentional action. But as far as I can tell, there are other explanations that do not depend on moral considerations at all.

So, for instance, according to one explanatory model, the following two conditions are necessary and jointly sufficient for an agent's being able to intentionally bring about some side effect, *y*, (a) she wants to perform some action, *x*, that produces *y*,

and (b) she knows (or at least believes) that by doing *x* she will bring *y* about.[4] If this model were correct, a side effect need not be either morally negative or positive in order to be brought about intentionally; it need only meet the two aforementioned cognitive and motivational conditions.[5] Thus, according to this model the primary reason that subjects judged that the CEO intentionally decreased sales in New Jersey in Knobe and Mendlow's CEO vignette is that she *wanted* to adopt the business plan even though she *knew* that it would have this negative side effect. If so, the perceived badness of decreasing the sales in New Jersey had little, if anything, to do with the results of the experiment. Notice that this model accounts for the experimental data just as well as Knobe and Mendlow's does.

Hence, if we are to decide between them, we must try to generate some new data that might settle the dispute. But first, we must get clear about what exactly Knobe and Mendlow's view really is—a view they summarize in the following way, "our hypothesis is that the intrinsic…badness of the side effect itself is what influences people's intuitions about whether it was intentional." For present purposes, I am going to assume that they believe that badness fully explains folk ascriptions of intentional action in side effect cases. Given that we construe—or perhaps misconstrue—their position in this way, we now have a method of testing to see whether it is true. Indeed, we can simply adopt the same strategy they adopted in their attempt to falsify my view. All we have to do is present subjects with two cases that have the same morally bad side effect but where

[4] It is not necessary that the agent wants to bring *y* about—wanting to bring about *x* will satisfy this motivational requirement for the intentionality of side effects. Indeed, if the agent wanted to bring y about as well, y may very well cease being a side effect at all. For a discussion of this peculiar fact about side effects, see Nadelhoffer (forthcoming).

[5] Of course, this model leaves open the possibility that moral considerations may *amplify* ascriptions of intentional action in side effect cases.

different knowledge constraints are placed on the respective agents. If the "desire plus knowledge" model that we have been discussing is correct, then a greater portion of the subjects who read a case where the agent knowingly brings about a bad side effect will judge that the side effect was intentionally brought about than subjects who read a case with the same bad side effect, but where the agent did not knowingly bring it about.[6]

To test this prediction, I conducted a mini-study of my own. Subjects were 44 undergraduates, each of whom received one of the following two vignettes:

> Case 1 (C1)
> John—the manager of a large company—is hosting a company cook out at his house on a Saturday afternoon. During the cook out he strikes up a conversation with Susan—an employee that he does not know very well. John casually asks Susan how she and her husband are doing. Unbeknownst to John, he could not have asked a worse question. It turns out that Susan's husband left her for another woman only days earlier. Understandably upset by John's question, Susan bursts into tears.
>
> Case 2 (C2):
> John—the manager of a large company—is hosting a company cook out at his house on a Saturday afternoon. During the cook out he strikes up a conversation with Susan—an employee that he has heard is having marital problems. This concerns John because he strongly believes that divorce is entirely immoral and unacceptable. If Susan and her husband are getting divorced, John wants to punish her by not giving her the promotion that she deserves. So, in order to get an update about her marriage, he asks Susan how she and her husband are doing, knowing that this will likely upset her. It turns out that Susan's marriage is much worse than John has heard. Indeed, Susan's husband left her for another woman just days earlier. Understandably upset by John's question, Susan bursts into tears.[7]

The subjects were then asked the two following questions: (1) Did John intentionally upset Susan? (2) How much blame does the John deserve for upsetting Susan (On a scale from 0 to 6—0 being no blame and 6 being a lot of blame)? And the results were as follows: Whereas *not a single one of the subjects* in Case 1 judged that John intentionally

[6] I am in no way committing myself to this hypothesis. I introduce it here solely to show that there are other reasonable explanations for Knobe and Mendlow's data.
[7] I would like to thank Al Mele for helping me come up with C2.

upset Susan (with an average blame rating of 1.41), 64% of the subjects in Case 2 judged that he intentionally upset her (with an average blame rating of 3.76).

Given that the side effect in both cases was the exact same—*viz.*, upsetting Susan—clearly the badness of the side effect cannot explain the stark difference in subjects' ascriptions of intentional action. To borrow a phrase from Knobe and Mendlow, this "spells trouble" for any view according to which moral badness fully explains ascriptions of intentional action in side effect cases.[8] However, several difficulties remain. First, the "desire plus knowledge" model has problems of its own. For one thing, it does not settle with the results of one of Knobe's earlier side effect experiments. After all, in the case involving a CEO who did not care that the business plan he was adopting would help the environment, both of the conditions of the "desire plus knowledge" model were met—*viz.*, the CEO wanted to adopted a business plan and he new that adopting the business plan would help the environment—yet subjects gave the CEO a low intentionality rating (Knobe 2003a). This shows that desire and knowledge cannot be jointly sufficient for ascriptions of intentional action after all—at least as far as the majority of lay persons is concerned.

Ironically, this brings us full circle to the extent that this is the very case that I was trying to explain in the paper of mine that is the target of Knobe and Mendlow's objections. And while this makes it seem as if we have not made much progress, we have actually discovered some interesting things about folk ascriptions of intentional action along the way. First, desire and knowledge are not jointly sufficient for folk

[8] Not only do the results of the two cook out cases support the view that judgments concerning whether an agent knowingly brought about a side effect act expansively on folk ascriptions of intentional action, but to the extent that there is a correlation in the these cases between blame attribution and ascriptions of intentional action, this study is also consistent with my claim that blame can also influence people's judgments about intentionality.

judgments about intentionality, although they may be necessary. Second, whereas blame is not necessary for folk ascriptions of intentional action, badness is not sufficient—although it may be necessary. And finally, badness is not sufficient for blame attribution— although it may be necessary as well.

Taking all of this into consideration, I want to tentatively put forward the following model to explain the biasing effect that moral considerations have on the folk concept of intentional action. 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.

If this model were correct, it would explain why the subjects in Knobe's earlier side effect experiment did not judge that the CEO intentionally helped the environment. After all, even though the CEO's business plan did bring about a positive side effect, he was not praiseworthy to the extent that he did not care that the business plan that he adopted would have this effect. It also explains why fewer subjects in my first cook out case judged that John intentionally upset Susan than subjects in the second case, *viz*., in Case 1, unlike in Case 2, John failed to satisfy conditions (1) and (2). Moreover, the only case that this model seemingly cannot explain is Knobe and Mendlow's CEO case—but to the extent that I have already successfully argued that this case does not involve any moral

considerations in the first place, it fails to undermine the model of the biasing effect of moral considerations I have just put forward.

Minimally, the moral of the present story seems to be that folk ascriptions of intentional action are sensitive to a number of considerations, including, but not limited to, moral goodness and badness, praise and blame, cognitive conditions such as knowing and believing, and motivational considerations such as desiring and wanting. The exact role that all of these various considerations play is still unclear. Thus, more research must be done before philosophers and psychologists will be in a position to develop a robust analysis of the folk concept of intentional action.[9] Hopefully, the issues addressed in the exchange between myself and Knobe and Mendlow will help motivate others to join in the effort of trying to discover just how nuanced folk ascriptions of intentional action really are.[10]

Bibliography:

Alicke, M.D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, Vol. 126, No.4: 556-574.

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***.**

Knobe, J. and Burra, A. (forthcoming). What is the relation between intention and intentional action? *Journal of Culture and Cognition.*

Knobe, J. and Mendlow, G. (forthcoming). The good, the bad and the blameworthy: understanding the role of evaluative considerations in folk psychology. *The Journal of Theoretical and Philosophical Psychology.*

Nadelhoffer. T. (2004a). "The Butler Problem Revisited." *Analysis* 64: 277 – 284.

_____. (2004b). Skill, Luck, and Folk Ascriptions of Intentional Action. *Philosophical Psychology*.

_____. (forthcoming). On Praise, Side Effects, and Folk Ascriptions of Intentionality. *The Journal of Theoretical and Philosophical Psychology*.

[9] It may be that such a robust analysis is impossible given the complexities of folk intuitions and judgments. But until we know more about these complexities we will unable to say for sure.

[10] I would like to thank Joshua Knobe, Al Mele, Eddy Nahmias, Jaspreet Singh, Virginia Tice and Seth Tyree for helpful suggestions on earlier drafts of this paper.