Intention, Intentional Action and Moral Considerations

One of the most complex and long-standing problems in the analysis of intentional action is concerned with the status of *side effects*. An outcome can be considered a side effect when

> (a) the agent chooses to perform a behavior that she knows will bring about this outcome but
>
> (b) the agent does not perform the behavior for the purpose of bringing about the outcome.

The key question is whether or not it can ever be correct to say that an agent brought about a side effect intentionally.

In a recent paper (Knobe 2003), I showed that people's intuitions about what is done intentionally in such cases are influenced in part by moral considerations. In particular, there seem to be circumstances in which people are considerably more inclined to say that an agent brought about a side effect intentionally when they regard that side effect as bad than when they regard it as a good.

Adams and Steadman (forthcoming) caution us against using this fact about people's intuitions to reach certain conclusions about the nature of people's underlying concept of intentional action. Suppose we distinguish between the concept of doing something *intentionally* and the concept of having an *intention*. In light of the facts we reported about people's intuitions, one might be tempted to conclude that people's concept of intentional action is such that a behavior can sometimes count as intentional even if the agent did

not specifically have an intention to perform it. Adams and Steadman argue that, in fact, there are good reasons not to draw this conclusion. Their claim is that the evidence we have amassed concerning people's use of words in ordinary language can be misleading when it comes to more fundamental questions about people's underlying concept of intentional action.

In arguing for this claim, Adams and Steadman make a number of criticisms of my earlier work. It seems to me that some of these criticisms are valid. The evidence I presented earlier is indeed open to alternative explanations, and it would be premature to infer, solely on the basis of this evidence, to any sweeping conclusions of people's concept of intentional action. In the present paper, I try to plug up the gaps in my prior work, drawing on some of the ideas that Adams and Steadman provide.

Prior Work

Before we turn to Adams and Steadman's arguments, it may be helpful briefly to review the evidence presented in my earlier paper. This evidence comes entirely from people's intuitions regarding specific cases. By considering pairs of cases that differ only in certain narrowly specified respects, we can get a clear sense for the factors that influence people's intuitions.

As one example, let us consider a quick story that I will call the *harm vignette*.

> The vice-president of a company went to the chairman of the board and said, 'We are thinking

> of starting a new program. It will help us increase profits, but it will also harm the environment.'
>
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

Confronted with this story, most people feel that the chairman harmed the environment intentionally.

But now suppose that we replace the word 'harm' with 'help.' We then arrive at what I will call the *help vignette*.

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
>
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was helped.

Confronted with this latter story, most people do not feel that the chairman helped the environment intentionally.

In my earlier paper (Knobe 2003), I demonstrated experimentally that these two vignettes elicit different intuitions — with 82% of subjects saying that the agent in the harm vignette intentionally harmed the environment and only 23% of subjects in the help condition saying that the agent intentionally helped the environment. Similar results were obtained for another, somewhat different pair of vignettes that had the same basic structure. These results seem to indicate that moral considerations play an important role in shaping people's use of the word 'intentionally' in ordinary language.

Adams and Steadman do not dispute this claim about ordinary language, but they do express concern that people's use of terms in ordinary language may be leading us astray us in our investigation of people's underlying concepts. Thus, they do not deny that people sometimes apply the *word* 'intentionally' in cases where an agent brings about some bad side effect, but they caution us against leaping too swiftly from this fact about the people's use of words to any broader conclusion about the nature of people's *concept* of intentional action and its relation to their concept of having an intention.

Pragmatics

Adams and Steadman's first argument is that people's use of the word 'intentionally' may reflect not only their concept of intentional action but also certain purely pragmatic factors. In particular, people's use of 'intentionally' may carry with it certain

implicatures about whether or not the agent is to *blame* for her behavior. Thus, when a speaker asks the question ‘Did she do that intentionally?’ it may sometimes be assumed that what the speaker really wants to know is whether or not the agent was to blame for her behavior. And, in a similar way, when a speaker says ‘She did not do that intentionally,’ it may be assumed — unless the speaker explicitly says otherwise — that the speaker believes that the agent was not to blame.

The key point here is that people’s use of the word ‘intentionally’ may not be an accurate reflection of their concept of intentional action. Perhaps people simply call the chairman’s behavior intentional because they want to avoid the conversational implicature that the chairman is not to blame.

This point seems to me to be a helpful one. One way to address it would be to find a second, entirely distinct method for determining whether people regard a given behavior as intentional — a method that did not rely in any way on people’s use of the word ‘intentionally’ and therefore did not involve us in the same pragmatic complications. Then we could check to see whether this other method yielded the same results. If we ended up obtaining different results when we used the new method, we might conclude that the results obtained by looking at people’s use of ‘intentionally’ were due primarily to pragmatic factors. But if we obtained the very same results using this new method, we would have good reason to conclude that the results we obtained when looking at people’s use of ‘intentionally’ did, in fact, reflect people’s concept of intentional action.

As it happens, I think that there is a method for determining whether or not people regard a given behavior as intentional without making any use of the word 'intentionally' (or any similar terms). This is to look at people's use of *reason explanations*. Here we will be relying on the widely accepted view that reason explanations are applicable only to intentional actions.[1] We will not be providing any independent argument for that view here. Instead, let us simply accept it as a working hypothesis. This hypothesis will be confirmed to the extent that it helps us to make sense of the phenomena we will be discussing below.

Assuming now that this view is correct, it seems that we can gain valuable evidence about whether or not people believe some given behavior to be intentional just by checking to see whether or not they accept reason explanations for that behavior.

So, for example, suppose that I want to get a beer and therefore start walking toward the refrigerator, when suddenly I trip and fall. Here it seems wrong to say, 'He tripped in order to get a beer.' Indeed, it seems wrong to use any sentence of the form 'He tripped in order to...' Presumably, the problem is that, since I did not trip intentionally, it seems wrong to explain my tripping using a reason.

---

[1] For arguments in favor this view, see Anscombe (1957), Goldman (1970), Malle, Knobe, O'Laughlin, Pearce, and Nelson (2000) and Mele (1992). I know of no arguments on the opposite side. (There has been controversy about the converse claim — that all intentional actions can be explained by reasons — but that converse claim does not concern us here.)

If, however, someone did sincerely utter a sentence of the form ‘He tripped in order to…,’ then we would have good evidence that this person regarded my tripping as an intentional action. People’s use of the phrase ‘in order to’ thereby provides us with a kind of indirect evidence about which behaviors they regard as intentional.

Perhaps we can use this kind of indirect evidence to reach a better understanding of the influence of moral considerations on people’s classification of behavior. We noted above that moral considerations sometimes influence people’s use of the word ‘intentionally.’ But now, armed with this indirect method for determining whether or not people regard a given behavior as intentional, we can check to see whether the effect continues to emerge even in a situation where people do not use the word ‘intentionally’ and do not engage in any act of explicitly asserting a behavior to be intentional.

Indeed, the effect does emerge even under these very different conditions. To see this, we need only consider the two vignettes presented above. It sounds at least somewhat correct to say ‘The chairman harmed the environment in order to increase profits.’ But it sounds very wrong to say ‘The chairman helped the environment in order to increase profits.’ Confronted with this latter sentence, one wants to respond: ‘Well, he might have *implemented the policy* in order to increase profits, but he didn’t actually *help the environment* in order to increase profits. In fact, he didn’t help the environment “in order to” accomplish any goal at all.’ Presumably,

our intuitions here are a reflection of our sense that the chairman's helping of the environment was not an intentional action.

To confirm that people do indeed have these intuitions, I ran a simple experiment. Subjects were 77 people spending time in a Manhattan public park. Each subject was randomly assigned either to the 'harm condition' or to the 'help condition.' Subjects in the harm condition received the harm vignette; those in the help condition received the help vignette. After reading their vignettes, subjects were given the sentence 'The chairman harmed [helped] the environment in order to increase profits.' They were then asked whether or not this sentence sounded right to them.

Subjects answered this question by providing ratings on a scale from -3 ('sounds wrong') to +3 ('sounds right'), with the 0 point marked 'in between.' The average rating for subjects in the harm condition was +.6; the average for subjects in the help condition was -1. This difference is statistically significant, t (77) = 2.65, p = .01.

Note that this new method allows us to evade the pragmatic complexities that afflicted our earlier experiments. Adams and Steadman are right to point out that, if a person says, 'The chairman did not harm the environment intentionally,' there may be an implicature that the chairman was not to blame for harming the environment. But no such implicature arises when a person says, 'It sounds wrong to me to utter the sentence "The chairman harmed the environment in order to increase profits."' There is, it seems, no *pragmatic* reason for people not to give such a response. And yet, it appears that people are significantly less inclined to give this

response than they are to give the analogous response when confronted with the help vignette. It therefore seems unlikely that the difference between people's responses to the harm vignette and their responses to the help vignette is due entirely to pragmatic factors. At this point, the most plausible hypothesis seems to be that the difference between the two vignettes is showing us something fundamental about people's concept of intentional action.

Intention and Intentional Action

Adams and Steadman's second argument — discussed only briefly at the end of their paper — is that nothing in our experiment permits us to test directly whether or not people thought that the chairman had an *intention* to harm the environment.

Since the chairman clearly was not trying to ensure that the environment be harmed, it seems natural for people to conclude that he had no intention of harming it, But this leaves us in the seemingly uncomfortable position of saying that people think he had no *intention* of harming the environment but nonetheless harmed it *intentionally*. Adams and Steadman suggest one possible way out of this position. Perhaps it turns out — contrary to what one would at first suppose — that people actually feel that the chairman did have an intention to harm the environment. Then the results obtained in our earlier experiment would, in fact, be consistent with the principle that an agent can only perform a

behavior intentionally if he or she had an intention to perform that behavior.[2]

To address this issue, I ran a second experiment. Subjects were 63 people spending time in a Manhattan public park. As in previous experiments, each subject was randomly assigned either to the harm condition or to the help condition. Subjects in the harm condition received the harm vignette; subjects in the help condition received the help vignette. Within each of these conditions, subjects were further divided into an 'intentionally' condition and an 'intention' condition. Subjects in the intentionally condition were asked whether or not the chairman harmed [or helped] the environment *intentionally*, whereas subjects in the intention condition were asked whether or not it was the chairman's *intention* to harm [or help] the environment.

Here are the percentages of subjects responding 'yes' to each of these questions:

	Harm	Help
'Intentionally'	87%	20%
'Intention'	29%	0%

[2] This principle has been quite controversial. For discussion, see Adams (1986), Bratman (1984; 1987), Harman (1976), McCann (1986; 1997) and Mele (forthcoming).

As in previous experiments, most people felt that the harm behavior was performed intentionally, whereas relatively few people felt that the help behavior was performed intentionally. This difference was statistically significant, $\chi^2(1, N = 30) = 13.4$, $p < .001$.

The more striking result, however, was that relatively few people said that it was the chairman's *intention* to harm the environment. Within the harm conditions, we therefore obtain a significant difference between people's responses for 'intention' and their responses for 'intentionally,' $\chi^2(1, N = 32) = 10.6$, $p < .01$.

In short, we seem to have identified a behavior such that (1) most people don't think that the agent had an *intention* to perform it but (2) most people do think that the agent performed it *intentionally*. This finding raises interesting questions about the relation between people's concept of intention and their concept of intentional action — questions that I hope to explore in future work.

Conclusion

Our aim has been to reach a better understanding of people's concept of intentional action, drawing on ordinary language as a key source of evidence. Adams and Steadman have contributed greatly to this effort by suggesting hypotheses that might not otherwise have received adequate consideration. An investigation of these hypotheses then led to certain new discoveries, relevant both to questions about the role of moral considerations in people's concept of intentional action and to questions about the relation between intentional action and the state of having an intention.

References

Adams, F. 1986. Intention and intentional action: the simple view. *Mind and Language* 1: 281-301.

Adams, F. and A. Steadman. forthcoming. Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*.

Anscombe, G. 1963. *Intention*. Ithaca: Cornell University Press.

Bratman, M. 1984. Two faces of intention. *Philosophical Review* 93: 375–405.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

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

Harman, G. 1976. Practical reasoning. *Review of Metaphysics* 29: 431–63.

Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 16**.**

Malle, B. F., J. Knobe, M. O'Laughlin, G. Pearce, & S . Nelson. (2000). Conceptual structure and social functions of behavior explanations. *Journal of Personality and Social Psychology* 79, 309-26.

McCann, H. 1986. Rationality and the range of intention. *Midwest Studies in Philosophy* 10: 191-211.

McCann, H. 1997. Settled objectives and rational constraints. In A. Mele, ed. *The Philosophy of Action*. Oxford: Oxford University Press.

Mele, A. 1992. Acting for reasons and acting intentionally, *Pacific Philosophical Quarterly* 73: 355-74.

Mele, A. forthcoming. Intention and intentional action. In Brian McLaughlin and Ansgar Beckermann, ed. *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press.