University of Delaware
Newark, DE 19716, USA
fa@udel.edu

References

Adams, F. 1986. Intention and intentional action: the simple view. *Mind and Language* 1: 281–301.
Adams, F. 1997. Cognitive trying. In *Contemporary Action Theory*, vol. 1, ed. G. Holmstron-Hintikka and R. Tuomela, 287–314. Dordrecht: Kluwer.
Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.
Cosmides, L. and J. Tooby. 1994. *The Adapted Mind*. Oxford: Oxford University Press.
Grice, H. P. 1989. *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.
Harman, G. 1976. Practical reasoning. *Review of Metaphysics* 29: 431–63.
Knobe, J. 2003a. Intentional action and side effects in ordinary language. *Analysis* 63: 190–94.
Knobe, J. 2003b. Intentional action in folk psychology: an experimental investigation. *Philosophical Psychology* 16: 309–24.
Malle, B. and J. Knobe. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33: 101–21.
Malle, B., L. Moses and D. Baldwin, eds. 2001. *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge, MA: MIT/Bradford.
McCann, H. 1986. Rationality and the range of intention. *Midwest Studies in Philosophy* 10: 191–211.
Mele, A. 1992. *Springs of Action*. Oxford: Oxford University Press.
Mele, A. 2001. Acting intentionally: probing folk notions. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. B. Malle, L. Moses and D. Baldwin, 27–43. Cambridge, MA: MIT/Bradford.
Mele, A. and S. Sverdlik. 1996. Intention, intentional action, and moral responsibility. *Philosophical Studies* 82: 265–87.

Intention, intentional action and moral considerations

Joshua Knobe

Adams and Steadman (2004) make a number of important criticisms of my work on the concept of intentional action. 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

Analysis 64.2, April 2004, pp. 181–87.

of this evidence, to any sweeping conclusions of people's concept of intentional action. In the present paper, I try to plug the gaps in my prior work, drawing on some of the ideas that Adams and Steadman provide.

1. *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 behaviour. Thus, 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 behaviour 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 behaviour 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 behaviour 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.

[1] For arguments in favour of 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.)

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

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.

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 behaviours 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 behaviour. 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 behaviour 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 behaviour to be intentional.

Indeed, the effect does emerge even under these very different conditions. To see this, we need only contrast the vignette about the chairman who harms the environment as a side effect with the vignette about the chairman who helps the environment as a side effect.[2] 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'

[2] These vignettes are presented in Knobe 2003: 191 and quoted in Adams and Steadman 2004.

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, $t\ (77) = 2.65$, $p = 0.01$.

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 use the sentence "The chairman harmed the environment in order to increase profits." ' There is no common practice of using this sort of utterance to express one's views about praise and blame, and if a speaker did use an utterance like this one, the audience would be highly unlikely to interpret it as a roundabout way of saying that the agent was not blameworthy. Most likely, the audience would interpret such an utterance in a more literal way: as a claim that a particular English sentence does not sound right. Thus, there seems to be 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.

2. *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 behaviour intentionally if he or she had an intention to perform that behaviour.[3]

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%

As in previous experiments, most people felt that the harm behaviour was performed intentionally, whereas relatively few people felt that the help behaviour was performed intentionally. This difference was statistically significant, $\chi^2(1, N = 30) = 13.4$, $p < 0.001$.

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,' $\chi^2(1, N = 32) = 10.6$, $p < 0.01$.

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

In short, we seem to have identified a behaviour 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.

3. *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.[4]

Princeton University
Princeton, New Jersey 08544-1006, USA
jknobe@princeton.edu

References

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

Adams, F. and A. Steadman. 2004. Intentional action in ordinary language: core concept or pragmatic understanding? *Analysis* 64: 173–81.

Anscombe, G. E. M. 1957. *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* 63: 190–94.

Malle, B. F., J. Knobe, M. O'Laughlin, G. Pearce, and 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. *American Philosophical Quarterly* 28: 24–36.

[4] I am grateful to Fred Adams, Michael Bratman, Gilbert Harman sand Alfred Mele for helpful comments on an earlier draft.

Mele, A. 1992. Acting for reasons and acting intentionally. *Pacific Philosophical Quarterly* 73: 355–74.
Mele, A. Forthcoming. Intention and intentional action. *The Oxford Handbook of Philosophy of Mind*, ed. B. McLaughlin and A. Beckermann. Oxford: Oxford University Press.

Newcomb's problem: the causalists get rich

Phyllis McKay

Suppose you are offered two boxes, one red and one black. You have to decide whether to take one box, or both boxes. There is always £1,000 in the black box, but what is in the red box varies. A predictor will have put £1,000,000 in cash in the red box if she judges that you will take only the red box. But if she judges that you will take both boxes, she will put nothing in the red box. We suppose that this predictor is a spookily good judge of character, and very likely to be right about you. We could suppose that this is a long-running game show, and the predictor has never yet been wrong. In the past, if the contestant has taken the red box, it turned out to contain £1,000,000. But when contestants in the past have taken both boxes, the red box contained nothing.

It is usually thought that evidential and causal decision-theory give different prescriptions for action for such a choice. Evidentialists advocate taking one box, since this choice is evidence for there *already being* £1,000,000 in the red box and will render the best option of getting £1,000,000 most likely. On the causalist view, you should take both boxes. The deposit in the boxes has *already happened*, and can no longer be affected by you – or by anyone at all. In the best case, you will have £1,001,000, and in the worst case you will at least have £1,000.

To make the right choice, you must decide whether you are in a *genuine*, or merely an *apparent* Newcomb case. Which case you are in is determined by what the shadowy figure of the predictor does or can do. In recent years, the predictor has got more attention, and this attention is along the right lines, but has not yet identified the right question, which is this: can your action now in choosing one or both of the boxes have a *causal* influence on the predictor's decision to put £1,000,000 in the red box?

The answer usually given is 'no', because the action of the predictor is in the past, and backwards causation is impossible. If this is the answer, you

Analysis 64.2, April 2004, pp. 187–89.