NORMS AND THE KNOBE EFFECT

RICHARD HOLTON, MIT

1 THE KNOBE EFFECT

In a justly famous study, Joshua Knobe found an asymmetry in the way we ascribe intentional action (Knobe 2003a). Consider an executive who, motivated entirely by the goal of maximizing profit, embarks on a policy that he knows will also cause environmental damage. Does he intentionally harm the environment? Most people hold that he does. In contrast, when considering an otherwise identical case in which the side effects would be *beneficial* to the environment, most people hold that the executive does not intentionally help the environment. A number of follow-up studies have found that the finding is robust, that it applies to children as young as four, and that it occurs in other languages and cultures (Knobe 2006).

If the finding is straightforward, explaining it is not. Considering it on its own one might conclude (i) that it is just the notion of intentional action that is affected, and (ii) that the crucial feature is moral judgment: actions judged to have morally bad side effects are judged intentional, whereas others are not. But subsequent work has shown that neither of these ideas is correct:

(i) The phenomenon applies to more than just intentional action. Ascriptions of whether agents *desire* to bring about the side-effects, *decide* to bring them about, *advocate* bringing them about, or are *in favor of* bringing them about, show a similar asymmetry. (Pettit and Knobe 2009) The asymmetry is also shown when subjects are asked whether agents bring about the side effects *in order to* increase profits (Knobe 2006). And even staying with the idea of intentional action, the effect comes up in surprising places: it arises in subjects' judgments of the intentionality of actions that are known to have little chance of success (Nadelhoffer 2004; Knobe 2006, 108–12; Pettit and Knobe 2009, 601–2).

(ii) In some cases morally good side effects are judged intentional, whereas morally bad are not. To take an example to which we shall return: a profit-driven executive whose actions have the side effect of violating a pernicious Nazi law—surely a good outcome in most subjects' eyes—is judged to have acted intentionally; whereas a similarly driven executive whose actions have the bad effect of fulfilling the requirements of that law is not judged to have acted intentionally (Knobe 2007; q.v. Knobe 2006, 111–12).

Various explanations of these results have been offered (see Pettit and Knobe 2009 for summary and references). But most have been piecemeal, accounting for one finding or another. Ideally we want an explanation that accounts for all of them in a unified way. This is what I try for here. More than that though, I don't simply aim to explain the findings: I aim to justify them. For I think that the subjects of the experiments are quite right in the ascriptions that they make. There *is* an asymmetry here.

Central to the account I shall offer is the idea of a norm; in particular, the idea of violation of a norm. It turns on two claims. The first is a claim about the origin of the asymmetry:

(1) There is a fundamental asymmetry concerning norms: to intentionally violate a norm all one needs to do is to knowingly violate it; whereas to intentionally conform to a norm one needs to be counterfactually guided by it.

This asymmetry is well-founded, since one violates a norm simply by ignoring it, and for that it doesn't have to act as a guide.

The second claim is that this asymmetry at the level of norms is transmitted up to our intentional ascriptions more generally:

(2) In making attitude ascriptions concerning a given action we are influenced by a number of factors; one central factor determining whether an action is judged intentional is whether in performing it the agent intentionally violates or intentionally conforms to a norm.

What makes all this plausible is the central place that norms have in our lives. We are norm driven creatures. From a very young age—by 2 or 3—children infer norms from the behaviour of others, conform their own behaviour to those norms, and then police them, criticizing and correcting those who deviate (Rakoczy *et al.*, 2008; Tomasello 2009, 34–9). Even apparently harmless behaviour can elicit a quick and angry response if it violates a norm: witness Garfinkel's celebrated 'breaching' experiments and the responses to such things as behaving like a lodger at home, erasing an opponent's move in tic-tac-toe (noughts and crosses), and being excessively literal-minded (Garfinkel 1967, Ch. 2). There are doubtless distinctions to be made between different types of norms—constitutive and cooperative, moral and conventional, legal, and so on—but the details are controversial, and will not be my focus here. All that I need is the idea that we are acutely sensitive to the existence of norms, very widely construed, and to whether they are being violated.

2 THE ASYMMETRY OF NORM VIOLATION AND NORM CONFORMITY

There is typically a gap between simply bringing an outcome about, and bringing the same outcome about intentionally. Dropping a catch is one thing; intentionally dropping a catch is another. In this case the gap is filled by something like an intention to drop the catch. When one intentionally drops a catch one has a plan to drop the catch that serves as a guide: one regulates one's behaviour to ensure that the ball is not caught. Much the same is true when one intentionally *conforms to* a norm. Agents who intentionally conform to a norm treat the norm as a regulatory guide. They will be ready to modify their behaviour (within reason) to ensure that the norm is followed.

When it comes to intentionally *violating* a norm though, things are very different. There is indeed a gap between simply violating a norm and intentionally violating a norm. But in this case, the gap is small and is filled by something like knowledge. I count as intentionally violating a norm if I knowingly violate it. I do not need, in addition, to treat violation of the norm as a regulatory guide. That is, I do not need to be ready to modify my behaviour to ensure that the norm is violated. It is certainly possible to behave in such a way: plausibly something along these lines is what Milton's Satan had in mind when he vowed 'Evil be thou my good' (Milton 1668, Book IV). Such satanic motivation is rare (Baumeister 1996), though

something with the same structure arises in cases of civil disobedience: here the agent's purpose is to violate the norm. But such intent is not necessary to intentionally violate a norm.[1]

Why does mere knowing violation count as intentional? The agent who knowingly violates a norm does have a kind of guiding intention, even though it is typically not the satanic one of acting contrary to the norm. Instead they intend to *disregard* the norm: to not let it stand in their way; to ignore it. Of course, such an attitude does not itself constitute intentional violation, for it is compatible with there being no violation at all. As we saw with the non-polluting executive, one can be ready to disregard a norm without ever needing to violate it. It is when the attitude of norm disregard leads to actual norm violation that we get intentional norm violation.

This difference between our concepts of intentional norm violation and of intentional norm conformity has been borne out experimentally. Recall Knobe's case of the profit-maximizing executive whose policies will have the side-effect of either violating or fulfilling the requirements of the Nazi racial identification law. In the case of violation, 81% judged the executive to have *intentionally violated* the requirements of the law. In contrast, in the case of fulfillment, only 30% thought he had *intentionally fulfilled* the requirements of the law. Intentional fulfillment is not quite the same as intentional conformity, but it is certainly very close.

One feature that this case brings out is that in making these different judgments, subjects do not have to think that violating the norm is a bad thing, or that conforming to it is a good thing. What matters is just that the norm is knowingly violated. This is not to deny that the issue of violation is connected to our broader evaluative and social concerns. Subsequent to the judgment that a norm has been violated come a host of further judgments: judgments about whether the violator has done wrong, whether they should be blamed, whether they should be punished, and so on. But these are further steps. The identification of intentional norm violation is independent of them.

3. The impact of norm violation on intentionality

In assessing others' attitudes we are moved by diverse factors: by 'worda ond worca' as *Beowulf* has it, by what is said and what is done (Heaney 2000). In judging whether someone believes a proposition, we are influenced both by what they say about the matter, and by how they behave; and the relevant patterns of speech and behaviour are themselves complex. It is not just what they say about the proposition itself, but also what they say about other propositions that somehow support it, or are in conflict with it; and the relevant behaviour is equally complex. Moreover, such complex factors are not merely evidential: it is not that belief is a simple thing which we can come to assess using these diverse factors. It is rather constituted by these factors.

---

[1] Note that there are other verbs for which there is no gap between performance and intentional performance. There is no difference between stealing and intentionally stealing, or between lying and intentionally lying. In these cases though this is because the verb is, so to speak, *intentionally thick*: the concepts of stealing and of lying already build in the idea that the act is done intentionally. It is possible that the idea of *following a norm* is like this—one only counts as following a norm if one intentionally follows it—which is why I have phrased my discussion in terms of conformity. Interestingly I don't think that we have a corresponding intentionally thick verb for the satanic notion of actively aiming to violate a norm; and this provides more evidence that it is the thinner notion of knowing violation that is our normal concern.

My second claim is that intentional behaviour is much the same. To judge an outcome as intentional is to see it has having various features that cluster around the ideas of intention and plan and decision, either actual or potential. Amongst these is the agent's attitude to the various norms that govern that outcome. If in performing an action the agent intentionally violates a norm against an outcome, then that is a factor in the outcome being brought about intentionally.

In saying that we are influenced by this factor, I do not mean that intentionally violating a norm against a certain outcome is either a sufficient or a necessary condition for bringing it about intentionally. I mean rather that it is a *contributory* factor. We are more prepared to judge the outcome intentional: we rank such a claim higher on a graded scale, and are more likely to endorse it if asked for an all-out judgement. The model of concepts in play here is thus that of prototype (Rosch, 1975), or, to use an older philosophical term, of the cluster concept.

The same is perhaps true about *conforming to* a norm: if in bringing about an outcome an agent intentionally conforms to a norm, plausibly that too will be a contributory factor in our judgment of whether the agent intentionally brought about that outcome. But since in the cases under consideration this is not the case, it will not have any impact.

4 EXPLAINING THE FINDINGS

I now turn to show how these considerations can explain the findings.

(i) The original case

The profit-motivated executive presses ahead with a policy that he knows will cause pollution; in so doing he knowingly, and hence intentionally violates the norm on not polluting. Since this is a contributory factor to intentionally polluting, subjects will tend to judge that he intentionally pollutes. In contrast the profit-motivated executive who presses ahead with a policy that he knows will benefit the environment does not intentionally conform to the norm on not polluting, since he is in no way guided by that norm (he explicitly disregards it). Hence this provides no reason to say that he intentionally benefits the environment; and so, in the absence of other reasons, subjects will not tend to say that he does.

(ii) Good norm violations

We have seen in the Nazi law case that the judgment that a norm has been violated is independent of judgment that that norm is good. Other cases exhibit the same structure. Pizarro, Bloom and Knobe found that agents who encouraged kissing between gay partners or who encouraged interracial sex were more likely to be judged to have done so intentionally than agents who encouraged kissing between partners of different sexes, or sex between partners of the same race (Knobe, 2007). They put this down to subjects' unconscious disapproval of the former activities; but it can instead be explained by the subjects' awareness of norms against these activities—whether they approve of them or not. (Knobe likewise explained the Nazi law example as resulting from an unconscious disapproval of violating the law; I find this implausible.)

(iii) Unexpected success

Typically it seems that we are reluctant to describe an act as intentional if the agent thought success highly unlikely. But this reluctance is also sensitive to the Knobe effect. Pettit and Knobe presented subjects with a pair of cases in which an agent is trying to activate a device that is controlled by a code (Pettit and Knobe 2009, 601–2). In the first case the device *detonates* a bomb; the agent aims to kill innocent tourists. In the second case the device *defuses* the bomb; the agent aims to save innocent tourists. In neither case does the agent know the code. He guesses, and happens to get it right. Subjects were more likely to judge that the agent in the first case intentionally detonated the bomb than they were to judge that the agent in the second case intentionally defused it. On the current approach the explanation follows from the fact that trying to violate a norm, even without the expectation that one will succeed, is itself a violation of that norm. (We don't think that there are two norms, a norm against killing, and a separate norm against trying to kill. Someone who tries to kill violates the norm against killling, whether or not they believe that they will succeed.) So if they do succeed, as in the detonation case, we have grounds for counting their action as intentional. In contrast where they do not violate a norm, as in the defusing case, there will be no such grounds.

(iv) Other propositional attitude verbs

The two step account can be extended to the other propositional attitude verbs— 'desiring', 'deciding', 'advocating', 'favouring' etc.—in a straightforward way. The idea, once again, is (1) that the fundamental asymmetry is at the level of the attitude to the norm, and (2) that this is then inherited by ascriptions that make no explicit reference to the norm. So:

(1) Just as intending to violate a norm requires less than intending to conform to a norm, so desiring to violate a norm requires less than desiring to conform to a norm, deciding to violate a norm requires less than deciding to conform to a norm, and so on. The polluting executive intentionally violates the norm. He also decides to violate the norm (he decides not to let the norm stand in his way); he advocates violating the norm (he advocates not letting the norm stand in his way); and he favours violating the norm (he favours not letting the norm stand in his way). In contrast the environment-benefitting executive does nothing that could count as deciding to conform to the norm on not harming the environment, or as advocating, or favouring, benefitting it.

(2) Just as the judgment that an agent intentionally violates a norm contributes to the judgment that the agent intentionally brings about an outcome, so these other judgments about attitudes to the norm contribute to judgments about the attitudes simplicier. Thus the judgment that A decides to violate a norm against an outcome contributes to the judgment that A decides to bring about that outcome; the judgment that A advocates violating a norm against an outcome contributes to the judgment that A advocates bringing about that outcome; and so on.

The case of desire is more complicated. The polluting executive does not have an intrinsic desire to violate the norm against pollution. But he does have some instrumental desire to violate it in the sense that he desires to violate it to achieve his desire to making a profit, which in turn will contribute to the judgment that he desires to hurt the environment. This suggests that subjects should find the ascription of a desire to hurt the environment less acceptable than

the ascriptions of decision, advocation or favouring. But there should still be an asymmetry, since there is nothing to be said for the claim that the environment-benefitting executive desires to conform to the norm. This is just what Pettit and Knobe found (p. 591).

A similar two step explanation applies to the locution 'in order to'. Knobe found that subjects are more likely to agree that a profit-motivated executive harmed the environment in order to increase profits than they are to say that a similarly profit-motivated executive helped the environment in order to increase profits. But in the first case the executive has intentionally violated a norm in order to increase profit; in the second case they have not intentionally conformed to a norm in order to do so. It is this fact that underpins the simple 'in order to' judgment.

CONCLUDING REMARKS

So far I have talked mainly of explanation. As I said at the beginning though, my aim is not just to explain but to justify. I think that the work has already been done. The asymmetry between intentional violation and intentional conformity is a real one, rooted in the very nature of the concepts. And it makes perfect sense that we incorporate our judgment that a norm was intentionally violated into our assessment of whether the outcome was intentionally brought about. Far from being a quirk of human psychology, the Knobe effect illustrates our abiding concern with norms.[2]

BIBLIOGRAPHY

Harold Garfinkel (1967) *Studies in Ethnomethodology* (Englewood Cliffs, NJ: Prentice-Hall).

Seamus Heaney, trans. (2000) *Beowulf* (New York: Norton).

Joshua Knobe (2003a), 'Intentional Action and Side Effects in Ordinary Language' *Analysis* 63, 190–3.

— (2003b) 'Intentional Action in Folk Psychology: An Experimental Investigation' *Philosophical Psychology* 16, 309–324.

— (2006) 'The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology', *Philosophical Studies* 130, 203–231.

— (2007) 'Reason Explanation in Folk Psychology' *Midwest Studies in Philosophy* 31, 90–107.

John Milton (1668)*Paradise Lost*.

Thomas Nadelhoffer (2004) 'The Butler Problem Revisited' *Analysis* 64, 277–284.

Dean Pettit and Joshua Knobe, (2009) 'The Pervasive Impact of Moral Judgement' *Mind & Language* 24, 586–604.

---

[2] Thanks to an audience at MIT; and to Joshua Knobe, whose input to this paper should qualify him as a co-author, were it not that he doubtless doesn't agree with it.

Hannes Rakoczy, Felix Warneken and Michael Tomasello (2008) 'The sources of normativity: Young children's awareness of the normative structure of games', *Developmental Psychology*, 44, 875-881.

Eleanor Rosch (1975), 'Cognitive Representations of Semantic Categories', *Journal of Experimental Psychology: General*, Vol. 104, 192–233.

Michael Tomasello (2009) *Why We Cooperate* (Cambridge, MA: MIT Press).