TROLLEY PROBLEMS **Bob Stone**

I may disappoint some of you when I say that the trolley problem I shall be talking about is not this one – hard though it is, even after inspection of the area at the edge of the Waitrose carpark, to see how this can possibly work. Perhaps someone can tell me over coffee?

Here is the kind of trolley that bothers philosophers, starting with Philippa Foot, who introduced this genre of thought experiments nearly 50 years ago in the area of ethics. She asks you (Foot 584) to imagine that you are driving the trolley, and it is heading for five people who are on the track and unable to move out of the way before being killed; but you know that by pressing a button you can redirect the trolley to a branch line where one person is on the track, thus killing him/her but saving the other five. Should you press the button? The general response she received is Yes; people would rather one was killed than five.

The more usual version, which was introduced by Judith Jarvis Thomson 30 years ago, has you as a bystander (Thomson 1397); you see that the trolley is heading for the five, who will certainly be killed, but you are standing near a switch, and you know that by flicking it you can redirect the trolley to the branch line so that it kills the one instead of the five. Should you flick the switch? Again, the large majority of people who are asked say that you are certainly entitled to do so, and most think that you ought to.

Thomson also introduces a quite new element (1409). This time you are standing on a footbridge over the track, and you can see that the trolley (unmanned) is about to pass under the bridge and then run over and kill five people on the track. You would like to throw yourself in its path to stop it, but you know your weight is not enough to do the job; but, as luck would have it, there is a large gentleman standing next to you who would do the job, and you just have time to topple him over the parapet so that he falls in front of the trolley. He will of course be killed, but the trolley will be stopped and the five on the track will be saved. Should you give him the push? This problem, in those politically incorrect days of the 1980s, she called the Fat Man. Almost everyone who is asked their opinion says No, you should not push him.

The interesting issue that arises from these two examples is that, while on the surface they seem identical, people who are asked their opinion give a clear Yes to the first and a clear No to the second. Each action – flicking the switch and pushing the man – will result in one person being killed by the trolley and five saved, whereas not to perform it will result in five dead and one alive. A no-brainer, you would think; *of course* it's better that one should die rather than five, and therefore there is a moral obligation on you to ensure that better result. So why do we all feel that throwing the switch would probably be OK, but pushing the fat man would not?

Thomson takes it more or less for granted that the popular view is correct, and she suggests various moral principles which might entail that the switch-flicking is right and the man-pushing wrong. The attempt is quite entertaining: every time she comes up with a promising theory, she can think of, or someone else has thought of, an ever weirder thought experiment which appears to refute it. For example – Philippa Foot's final position – killing is always worse than letting die (Thomson 1396). So pushing the man off the bridge is off limits, however many people it might save from death. We normally think that the negative duty not to cause harm is more important than the positive duty to help.

But this general principle is challenged by a thought experiment suggested by James Rachels in a paper supporting the "equivalence thesis", which holds that killing and letting die are morally equivalent. Imagine two women who both want to poison their respective uncles. One makes some poison and administers it, the uncle dies; the other has made the poison, but by chance she sees her uncle drinking something else she knows to be poisonous and, though having the antidote about her person, deliberately fails to administer it; the uncle dies. Is the second woman significantly less guilty than the first? In any case, apart from giving us an excuse for laziness and ducking out of difficult moral decisions, where is the logic in the moral distinction between causing death by a deliberate action and allowing death to occur by a deliberate omission?

Another of Thomson's suggestions (1401) is one version of Kant's categorical imperative, that we shouldn't use a human being as a means to some end (Kant 4.429). So, in the case of the bystander and the switch, the bystander is not going to use the person on the branch-line as a means to saving the other five; he merely foresees that that man will die as a result of his using the *diversion to the branch line* as a means to saving the five. On the other hand, if you push the man off the bridge you are using his body as a means to stop the train killing the other five – therefore wrong.

But, what about a slightly different switch scenario, as illustrated here, where the branch line is part of a loop (Thomson 1402). If you divert the trolley along it, the body of the man on the track, who will be killed by the impact, will stop the trolley, which will then be unable to continue on its way and kill the other five, which it would otherwise have done. This would entail using that one man to save the other five, but it doesn't seem automatically wrong that you should do so; at least it seems a good deal more permissible than pushing a man off a bridge.

She also explores the idea of people having inviolable rights (1403), such as the fat man's right to life, which we must not infringe by pushing him off the bridge; when we flick the switch, on the other hand, what we are doing is not *in itself* infringing anyone's right to life – or isn't it? The man on the branch line might have a view on the matter.

Again another criterion may be the *intention* of the agent as opposed to the *foreseen consequence* (Foot 582): pushing the fat man off the bridge is an intentional attempt to kill him, whereas flicking the switch – at least in the original scenario – is not intended to kill the man on the track, even though we foresee that he will die; we'd be only too happy if he suddenly escaped.

Thomson's favoured idea is that, if there is some danger that threatens death anyway, it is OK to divert it from killing five to killing one (as with the switch, in either scenario), but not OK to create a new threat in order to release five from a different threat (1413). I think you could argue that pushing the fat man off the bridge, so that he is killed by the trolley and the other five are not, might be permissible on that principle; but I don't think Thomson does.

What it does get round is my favourite of all ethical thought experiments: the surgeon at the hospital (Thomson 1396). When I turn up at Worcester Royal Hospital to visit a friend, a surgeon spots me from his office window walking up the path. He has five critically ill patients, who will all die within a very short time unless they receive transplants of heart, lungs, kidney, liver. When he sees me, he thinks, "Now there's a

healthy young man, whose records, which I can see on this computer, say that all his organs are in perfect working order. My patients could all survive after all." Is he entitled to kill me when I walk in and to distribute my organs among the 5 patients in such a way that they are all saved from the immediate death that was coming otherwise and will live long, healthy lives? Instead of five dying, with all the sadness that that will entail, only one will die. Seems an open and shut case. But it does violate the principle above, as he is not diverting an existing threat from them to me, and he does seem to be infringing my rights, and he is using me as a means, and he is killing me intentionally rather than merely foreseeing that I will die, and killing me is a worse crime than letting five people die. Luckily, for whatever reason, a surgeon's right to kill me is not supported by anyone at all, especially not me. But are we all wrong, when it seems his killing me would cause the world more happiness than sadness?

It is easy to go mad reading the various attempts to think of moral theories to justify this or that distinction. But it may be that they are all missing the point. One thing that the thought experiments all highlight is the difference between a consequentialist judgment of which action, or non-action, will cause the most good or the least harm, and a concern with rule-keeping, in which we care about such things as duties and rights, and think that certain actions are just plain wrong, whatever the consequences. That is often summarised as the difference between a utilitarian and a deontological approach. In fact such thought experiments are often used as a critique of out-and-out utilitarianism. But there is another distinction, which is in the area of psychology: there is a conflict between, first, what we *calculate*, apparently rationally, will be the consequences of an action and, second, our *intuition* about the rightness and wrongness of the action.

This dichotomy is characterized by Joshua Greene, a neuroscientist, as the 'dual-process theory' (Greene 11). His interest is in seeing which parts of the brain appear to be involved in the calculating approach and which parts in the intuitive reaction. Recently we have been able to do research on healthy subjects by using functional magnetic resonance imaging on their brains. This (normally known as fMRI) means that we can scan the brain to detect areas which at any moment have an increased blood flow, which signifies greater activity among the neurons.

Greene (12) used various trolley problems, among others, as examples of tests to give his subjects; he then observed which bits of their brain were activated when they were giving the two types of answer. With the proviso that (a) fMRI is quite a crude method, and (b) there is no nice simple correlation between individual parts of the brain and individual mental states, the upshot is that the dual-process theory was pretty well confirmed.

So, for instance, when the problem involved doing harm of an upfront and personal nature, for example pushing the man off the bridge with one's hands, areas of the brain associated with emotion were engaged; but when the harm was not so personal, but at more of a distance, as in the switch-flicking case, it was the turn of the brain areas connected with working memory and cognitive control, the thinking areas (Greene 13). When there was a conflict such that different subjects gave different answers, those giving the utilitarian answer – i.e. brace yourself and do the nasty deed – spent longer giving the answer than those who intuitively decided against the nasty deed (Greene 14); which is what you'd expect, as it involves more thinking.

Well, how do we adjudicate between the intuitive, deontological approach and the calculating, utilitarian one? Peter Singer has a strong view about the question. Why, he wonders, do we use intuitions as data against which to judge moral theories? He refers both to the discussions and research I've already described about trolley problems, and to the idea of John Rawls in his book 'A Theory of Justice' that we need to reach a 'reflective equilibrium' between our best moral theory and our initial, intuitive moral judgments (Singer 344). As in forming a scientific theory, we should, says Rawls, when finding a clash between the theory and the data – in this case intuitions – modify one or the other until we reach that equilibrium. But Singer insists that this is quite wrong. You could, by looking at our intuitions, produce a psychological or anthropological theory of why we think as we do, but that is not a *moral* theory. A moral theory is a theory about what we *ought* to do, and moral intuitions are not data for it (345-6).

Now there seems little doubt that intuitive emotional responses arise from a mixture of evolution, upbringing and culture. Given that our genes want to reproduce themselves, we can easily see why we – their carriers – have a natural urge to survive, to find sexual partners, to favour our children and other relations; and even crude notions of reciprocity among members of a group are useful for the benefit of all members, and therefore of the genes within; apparently monkeys pick parasites out of each others' backs, and ostracize those who 'cheat' by not returning the favour (Singer 336). And in the small groups in which we used to live for our first few hundred thousand years, when it would generally be to our disadvantage to do violence to other group members if the community was to survive, we developed intuitions against doing personal, up-close violence. The opportunity to do indirect violence at a distance, such as flicking a switch, did not figure prominently in those communities (Singer 348).

Even if you don't go a bundle with such evolutionary explanations, remember the way you were brought up as children – or at least those of you who are typical of the sort of people who join the Oxford Philsoc. You were told very firmly never to hurt anyone; the only chance you ever had to hurt anyone was to punch them or trip them up, rather than to operate switches to divert trolleys. What is more, these rules were first presented to you as having some kind of absolute, maybe divinely sanctioned validity, rather than accompanied by logical justification, so that your attitude towards them is one of piety, and breaking a rule makes you feel guilty. That is the most effective of getting you to obey them. So, whether you prefer evolution or upbringing, it is not surprising that we all have an intuitive distaste for doing upfront personal harm to others, and a less strong distaste for causing harm at a distance or, for that matter, for allowing harm to happen without doing anything about it.

Singer's conclusion from all this (349) is something I entirely agree with: if our intuitive aversion to doing certain types of direct personal harm is (a) an emotional response, as brain studies show, and (b) caused by our evolutionary history or (I'll add for him) by our upbringing and culture, it has no normative force whatsoever. It has no moral significance, no more moral significance than all those other intuitive emotional responses we have – many of which, if I'm anything to go by, would, if acted on, land me in prison. That doesn't mean that it is OK to push the fat man off the bridge; but *whether* it is OK needs to be worked out properly, and the fact that we intuitively recoil from it is of supreme irrelevance. Any attempt to *justify* the intuitive aversion is mere rationalisation after the event. So all those attempts to work out a

moral theory, which caters for different intuitive reactions to different trolley problems, are barking up the wrong tree.

Where does that leave us if we want to decide what is right and wrong? Singer, as is well known, advocates consequentialism as the only rational way of deciding on rightness and wrongness; in other words, an action is right if the consequences of doing it are better than those that would result if we didn't do it. So in all these trolley problems the right way to act is that which our *reason* tells us will lead to the better outcome, rather than that action or inaction which appeals to our intuitions, or which we feel comfortable doing.

That seems eminently sensible to me . . . except for one thing. Although it may be possible to work out by reason what the consequences of performing an action will be, and what the consequences of not performing it will be, how do we know that one set consequences is better than another? In the switch problem, for example, how do we know that, other things being equal, five deaths are worse than one? Or that, generally speaking, death is something to be avoided where possible? As Singer points out (349), philosophers have been trying for centuries to find some fundamental moral principle on which to base our ethics; most agree that all have failed. So, if it is irrational to follow our emotions, or intuitions, and if there is no reason to justify any moral preferences, we seem to be floundering in a sea of nihilism.

There is a further problem: as neuroscientists such as Antonio Damasio have shown by their work with brain-damaged patients, emotions are *necessary* for people to come to a decision. Those whose reasoning powers are entirely unimpaired, but whose emotions are no longer triggered after a brain injury, simply cannot make up their minds what to do. So, even if we did think up some entirely rational moral principle, it seems that no one would be able to decide to act on it.

The utilitarian proposal is that the basis of moral actions is, or should be, to promote the general happiness. Singer admits that this is a kind of intuition – what he calls a 'rational intuition' (351). That may sound like an oxymoron, or a cop-out; but it could be the only option, or the only honest option, left to us. Leaving aside those people, if any, who simply do not care about being moral, most of us probably do in fact take the general happiness as the basis of our thinking about right and wrong. That's why we have the rules we have, and why we want to adhere to them. We assume that it is a bad thing if someone is killed, and that therefore five deaths are worse than one; we assume it is better if a person, or people in general, are happy than if they are unhappy; we assume it is wrong to hurt people for fun. This "sentiment of generalised benevolence", to quote JJC Smart (Smart and Williams 7), "is surely present in any group with whom it is profitable to discuss ethical questions" – among whom I include members of the Oxford Philsoc.

Of course there are, as in *any* conceivable moral system, huge problems in, for instance, deciding exactly what types of happiness are important, or judging one person's happiness, or life, against another's; so there is no easy calculus for deciding how to act in any possible situation. But in general, we assume that, if by doing something, we will cause an increase in happiness, it is the right thing to do. There is no rational justification for this as an absolute principle; but it is what most decent people can be assumed to be committed to, and in their enlightened moments to *want*. It certainly seems a better bet as a moral starting-point than the idea of doing what we

feel in our 'gut' is the right thing from moment to moment, when our guts have been shaped by a history that may or not have got it right.

So let us run with that idea for now. The big problem, as I see it, is what count as the 'consequences' of my action. There seem to me to be three levels of consequence that an action may have:

1) First, there are the immediate, obvious consequences. In the case of the Fat Man, the effect of pushing him is that he will fall off the bridge and probably die; that is clearly a bad thing on the utilitarian – or any other – basis. It seems likely that our intuitions are activated most powerfully by this level of consequence, because when we are growing up, (a) that is the only one we are likely to foresee, and (b) it is the only one we have any real control of.

2) Second there is the knock-on effect. If I push the Fat Man off the bridge, the immediate effect is his death, but the knock-on effect is the saving of the other five people. At first sight, the rational response is to push the man, since five deaths are worse than one. But hang on! In this case, as in probably the vast majority of real-life cases, our calculation of the knock-on effects is guesswork. It is part of the Fat Man scenario that the pusher *knows for certain* that (a) his body will fall precisely on to the track, (b) it is bulky enough to stop the trolley, (c) the pusher himself is too light to do so. It is inconceivable that even Isaac Newton could work all that out by just being there. In a sense, the scenario is sheer fantasy, as it assumes a knowledge of the future that in real life we simply do not have. In a real world situation, where our knowledge of the immediate consequences of our actions far exceeds our knowledge of the knock-on effects, if people kept doing acts of violence in order to achieve longer-term good, they would probably miscalculate as often as not; if, on the other hand, we concentrate simply on avoiding those harmful effects that we know will result, rather than trying to foretell the future, the overall result of us all behaving like that will probably be on balance happier. This is an argument for rule-utilitarianism; if we take it as a rule that we shouldn't ever push people off bridges, occasionally we may be wrong, but the vast majority of times we'll be right and in the long term the world will be a happier place.

But what about those cases where we really can be certain that the result of a horrible deed will be to produce greater happiness? Take the surgeon, for example, who really *does know* that by killing me he will save the lives of five – or even two or three – patients. The foregoing argument does not apply to him, and so it seems he should go ahead and shoot.

3) We now come to the third level of consequence. As well as the immediate effect of his action (me dead) and the knock-on effect (five lives saved), there is also the effect on the community of such a practice being either commonly engaged in or even regarded as permissible. Even those five patients, who will die without my organs, will agree that they would not like to live in a society where surgeons routinely, or indeed ever, kill people at random in order to harvest their organs; we should not be *happy* in such a society. We'd have to creep around the place in terror of being randomly killed by someone who had some greater cause in view and would for that reason get away with murder. We should be far *un*happier with that scenario than with the possibility of dying for want of an organ donor.

So a properly rational utilitarian view, where we look at the overall consequences not just of our actions but of our rules and habits, indeed of our general moral mindset,

does seem to justify at least some of the attitudes that our emotional intuitions inspire us to hold. Pushing the fat man off the bridge is not actually the rational thing to do; but the fact that we have an intuitive aversion to doing it is not the reason.

But a final warning, in case you think I have simply rationalised after the event and saved moral intuitions. Bernard Williams (Smart and Williams 98-99) asks you to imagine finding yourself by accident in a South American town, where there are twenty local Indians tied up against the wall – about to be shot – and you are regarded as an honoured guest. You are told by the man in charge that, as a guest, you may kill one of the Indians yourself, in which case the other 19 will be freed instead of being shot. You can see that there are no alternatives, such as overpowering or persuading the shooters. Do you kill the one?

The obvious utilitarian answer is Yes, though you will have an intuitive horror of so doing. Is it right to dismiss that aversion as squeamishness, to dismiss any concerns you have for your own psychological state afterwards and for your integrity, etc etc, as self-indulgence rather than true moral sensitivity? And what if I change the scenario slightly, and say that you are asked to kill not one of the twenty Indians but a girl from the town, and what's more to stab her to death with a knife – than which you cannot get more upfront and personal? Or if I increase the number of Indians at risk from 20 to 1000? Is there any honest and moral way of getting out of the utilitarian argument for killing? I leave that with you.

References

Damasio, Antonio: *Descartes' Error*, Vintage 2006
Foot, Philippa: 'The Problem of Abortion and the Doctrine of the Double Effect', in *Ethical Theory*, ed. R. Shafer-Landau, pp 582-589, Blackwell 2007
Greene, Joshua D: *The Cognitive Neuroscience of Moral Judgment*, found at http://wjh.harvard.edu/~mcl/mcl/pubs/Greene-CogNeurosciences-Chapter-Consolidated.pdf (c 2009) – a summary of recent work, including his own, in the field: accessed on internet August 2016
Jarvis Thomson, Judith: 'The Trolley Problem' in *The Yale Law Journal* Vol 94, No 6 (May 1985), pp 1395-1415
Kant, Immanuel: *Groundwork of the Metaphysics of Morals*, tr M Gregor, CUP 1998
Rachels, James: 'Killing and Letting Die', in *Encyclopedia of Ethics*, 2nd edition, ed Lawrence Becker and Charlotte Becker, New York 2001, vol 2, pp 947-950
Singer, Peter: 'Ethics and Intuitions' in *The Journal of Ethics* (2005) 9: pp 331–352
Smart, JJC and **Williams, Bernard**: *Utilitarianism – For and Against*, CUP 1973