Thought Experiments in Ethics

Richard Baron, www.rbphilo.com

This is a draft text of an informal talk given in June 2017. It should not be quoted or cited without the author's permission. The author may be contacted on richard.baron.789@gmail.com

1 The topic

Our topic is the use of thought experiments in ethics. These are invented case studies, tailored to have the specific features that philosophers want in order to make their arguments.

Typically someone will create a thought experiment, consider different decisions that an agent in it might take, and then consider the acceptability of those decisions. A thought experiment can be used to test an ethical theory by assuming that the agent takes decisions in accordance with that theory, and then asking whether those decisions would be acceptable ones. Acceptability may be intuitive acceptability, consonance with other theory-based principles, or the acceptability of some, perhaps non-obvious, implications of the decisions or of the methods of reaching them.

Our question is this: Is this really a good way to go about deciding which ethical theories are acceptable as they stand, and which ones should be modified or discarded? Or indeed to make us think of considerations we might have overlooked and improve our theories?

2 The plan of this talk

We shall first give some examples of thought experiments, and then set out some functions of thought experiments. Then we shall look at how cases are described, and at two common features of the contents of experiments: the exclusion of options which we might like to choose, and the use of extreme cases. We shall then consider the use of intuition. We shall not reach an overall judgement that thought experiments are a good idea, or that they are a bad idea. But we shall identify a number of things to look out for when they are used.

3 Examples

Carneades of Cyrene (now on the coast of Lybia) (214/3 to 129/8 BC), a Greek skeptic, offered us the plank. Two sailors have been shipwrecked. There is a plank in the water that can only support one of them. Sailor A gets there first and sits on it. Sailor B pushes Sailor A off, so A drowns and B survives.[1] Now suppose that B is charged with murder.

[1] Lactantius, *Divine Institutes*, Book 5, chapter 17.

We could take this one in at least two different (but potentially compatible) directions. (1) It would be ludicrous to charge B with murder, so any ethic which simply takes in what are usually the important facts, such as that one person deliberately ended another's life, without looking out for very special circumstances, and looks up the answer, in this case murder, in a rule book, is no good. (2) B acted wrongly, and an ethic should recognize this by requiring you to treat other people's interests equally with your own, or at least should not allow your interests always to trump other people's.

Immanuel Kant offered us the axe murderer, who comes to your door not to attack you but to ask whether the intended victim is hiding in your house (when he is). Kant argued that you should not lie, and noted that a lie could lead to murder – if the intended victim has escaped through a back door and gone off down the street.[2]

Robert Nozick offered us the experience machine.[3] Plug yourself into it, and you will appear to yourself to have the most wonderful experiences, and you won't know they are faked. Pleasure would be maximized, but this kind of life strikes us (outside the machine) as unsatisfactory. So there must be something wrong with any ethic which concentrates only on maximizing the quantity of pleasure.

Peter Unger offers us the vintage sedan.[4] (Peter Singer has made use of the same kind of argument.) You are not rich but you have one luxury, a lovingly restored vintage sedan car. Someone lies by the road with a badly injured leg. If you pick them up and take them to hospital, they will in due course make a full recovery. But the blood will ruin the upholstery in your car. If you leave them, someone else will find them later, and take them to hospital. They will survive, but lose the injured leg. A common response is that you should pick them up and ruin your car. But then why is it also a common response that you are entitled to ignore an appeal for funds (which you could easily afford) to support an aid programme in some far-off land, where your donation would save more than 30 children from early death?

John Harris offers us the survival lottery.[5] Everyone has to join in. If you have useful organs and there is a shortage of organ donors who have died naturally, your number may get drawn. If it is, you are killed and your organs are used to save two or more people from their fatal medical conditions. Thus the quantity of human happiness among people already alive will be increased. If you are a hard-line utilitarian, can you say that participation should not be compulsory?

A self-driving car is to be programmed to handle situations in which either its passengers or others must be killed. For example, there may be an unforeseen obstacle on the road. Driving into it will kill the passengers. Swerving to avoid it will kill more people on the pavement. How should such cars be programmed? And should people who buy such cars have any right to choose or change the programming?

[2] Kant, "On a Supposed Right to Lie Because of Philanthropic Concerns."

[3] Nozick. *Anarchy, State, and Utopia*, pages 42-45.

[4] Unger, *Living High and Letting Die: Our Illusion of Innocence*, pages 24-25.

[5] Harris, "The Survival Lottery." *Philosophy*, volume 50, number 191, 1975, pages 81-87.

Philippa Foot gave us the trolley problem.[6] A trolley (streetcar) is running out of control. If it continues on course it will kill five people stuck on the track. But you can switch the points, or push a fat man into its path, to lead to one death instead of five. If you are a utilitarian who cares only about the outcome, and not about who does what, you will do either of these. But should you?

Bernard Williams gave us Jim and the Indians.[7] Jim is on a botanical expedition in South America. He stumbles upon a village in an area where there have been protests against the government. The army are about to shoot 20 innocent people as a warning to others not to protest. But the captain invites Jim to kill one of the 20, then the others will be released. Otherwise all 20 will be killed. The people against the wall and the other villagers beg Jim to accept the offer. But should he? (Note the difference from the trolley problem. What happens next will not be inevitable. It will depend on the captain's decision as to whether to follow through with his threat to kill 20 if Jim refuses, or as to whether to keep his promise if Jim complies. Also note that it would be a bit much to say that this thought experiment disproved utilitarianism. The point for Williams was not that it would be wrong for Jim to kill one, but that utilitarianism must be missing something because it cannot account for the great difficulty of Jim's decision.)

4 Some functions of thought experiments

Thought experiments could be used to do any of the following:

- Challenge an ethical theory by showing that it leads to unacceptable recommendations.
- Give reassurance that a theory is acceptable by showing that it leads to acceptable recommendations.
- Explore ethical problems.

5 The description of cases

One notable feature of thought experiments is that cases are described very concisely, giving only a few of the facts that would be present in any real-world situation. This is inevitable, in that adding lots of detail would make the descriptions too long, and readers would find it hard to sort the significant features from the insignificant ones. But it should also raise a doubt about the usefulness of thought experiments.

Doubt arises because there is a danger that only details which happen to encourage a particular conclusion will be included, and other details which would encourage other conclusions may be omitted. For example, the self-driving car case, as usually stated,

[6] Foot, "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review*, number 5, 1967, pages 5-15.

[7] J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against*, pages 98-99.

does not go into details about the capacity of the hardware to detect and the software to consider various factors, such as the ages or the family circumstances of the potential victims. Nor is anything normally said about how human drivers would be likely to think and what they would be likely to do. So if we conclude that self-driving cars are ethically problematic and should not be introduced without precautions to stop them getting into such difficult situations, precautions such as strict limits on their speed, we may reach that conclusion without adequate consideration either of what they might do or of how the alternative – human drivers – would behave.

This looks a bit worrying. But whether it should worry us depends on what we want our thought experiments to do.

If the aim is to challenge a theory, then simplification, even simplification which biases our thought in a particular direction, may be perfectly acceptable. When a challenge is to be mounted, it is reasonable to make it a severe one, and even to select features of a case to include on the basis of where one suspects the weak points of a theory to lie.

Defenders of the theory may regard the challenge as unfair, but the onus is then on them to show how the selection of detail has made it unfair. For example, if the story of Jim and the Indians is used to show that merely counting the bodies is not the only thing to do, so that utilitiarianism is inadequate because it disregards the point of view of the person taking action and his sense of personal responsibility, the utilitarian could respond that by including an account of how Jim came to find himself in this situation, and indeed by giving him a name, the anti-utilitarian is unduly personalizing the situation, and not letting us see it from a detached point of view, while by suggesting that the captain will definitely shoot all 20 if Jim refuses, the anti-utilitarian is going the other way and ignoring the personal responsibility of the captain for his actions, so as to make Jim's position as painful as possible.

If, rather than challenging a theory, we want a thought experiment to show us that a given ethical theory really is adequate, we should be worried. The omitted details might have included some which, if given, would have taken away the reassurance that the thought experiment gave.

Finally, if a thought experiment is being used to explore a problem, there is no harm in omitting detail at first. But if a case as initially described would appear to give rise to an intractable ethical problem, we should not give up. It may be a matter of redescribing the case, and thereby making the problem more tractable.

6 The exclusion of options

One notable feature of thought experiments which challenge a person within them to decide what to do – and therefore challenge you, the reader, to say what you would do – is that they limit the options available. In the example of Carneades' plank, you are not allowed to say that the sailors should have taken half-hour turns on the plank, potentially saving both of them. In the trolley problem, you are not allowed to try something else, like shouting a warning to the people on the track. Jim is not allowed to try to negotiate with the captain. And so on.

Is this legitimate?

It is not clear that tightly limiting options is wholly legitimate when the objective is to challenge ethical theories. The closing off of options may put theories under more severe challenge than would actually arise except in the most extreme and most unlikely circumstances. It may be all very well to rule out Jim's negotiating with the captain in order to make the point that there are times when the only way to follow utilitarianism is to do something hideous. But that attack on utilitarianism is weakened if there usually would be an option to negotiate (particularly, in this example, since a refusal on the captain's part to negotiate would not be Jim's responsibility).

When the objective is to gain reassurance as to the qualities of an ethical theory, restricting options can have a valuable role to play. It may show that the theory stands up even under tough restrictions. For example, if we confront virtue ethics with Kant's axe murderer, and impose the restriction that the murderer's question must be answered truthfully or with a lie (no clever third option like saying ”He [the victim] is not *here*”, while stamping your foot), the virtue ethicist can say that honesty is a great virtue but that there are situations in which it can be outweighed by other virtues, such as the preservation of life.

It also seems to be perfectly legitimate to restrict options if the objective is to make us think hard about ethical problems. Closing off options puts us under increased pressure to think about what is really at stake. In addition, we can introduce and remove restrictions to see what difference they make.

7 The use of extreme cases

It is quite striking that thought experiments typically involve extreme cases, such as death, serious injury or, as with the experience machine, a life of pleasure without pain. Is this use of extreme cases a good idea?

If the objective is to challenge ethical theories, it has a certain logic to it. If a theory is put forward as the one correct way to think about ethical problems, we need to see how the theory holds up when the going gets tough. The result of failure on an extreme case need not however be the collapse of the theory. It may be that the theory can be retained for non-extreme cases, especially if we can give a reasonable idea of what would constitute an extreme case. Alternatively, it may be possible to modify the theory so that it does better in extreme cases.

Consider as an example a theory such as Immanuel Kant's which would rule out suicide, and by extension euthanasia at the patient's request, in all circumstances. (We mean actions directed to bringing about an early death. The use of painkillers which would incidentally shorten life could be a different matter.) When someone expresses such opposition, one can say ”Suppose that a soldier is badly, and almost certainly mortally, wounded on the battlefield. He begs his comrades to shoot him. Do you really think they should refuse?" (This is something that really does happen. This thought experiment is not hypothetical.) The opponent of all forms of euthanasia might say that they should refuse, but most of us would find it hard to accept that answer. Or they might say that this is an extreme case where the normal arguments against euthanasia do not apply with their usual force. That would amount to limiting the reach of the theory which underpinned their view to non-extreme cases. Or they might allow that there could be arguments in favour of allowing euthanasia which could occasionally outweigh their arguments against. That would amount to modifying the underlying theory, so that it ceased to be a theory which required an inference from absolute rules to an inevitable conclusion, and became instead a theory in which factors which would argue for different conclusions had to be balanced.

If the objective is to gain reassurance as to the qualities of an ethical theory, satisfactory performance of the theory when faced with extreme cases can give considerable reassurance, but extreme cases cannot do the job on their own. A theory might perform impressively in difficult circumstances, for example by insisting that we take account of all possible consequences of actions (as utilitarianism in its purest form would), but be impractical in daily life – or even lead us astray in daily life, when consequences were not so extreme and there was therefore space to attach weight to other factors, such as the exercise of virtues, even if doing so would lead to sub-optimal consequences.

We should also not necessarily think badly of a theory merely because it did not perform well when faced with extreme cases. It might be a perfectly good theory most of the time. If we only look at extreme cases, we might miss this fact.

Finally, when the objective is to explore ethical problems, there is no harm in using extreme cases, just as there is no harm in exploring any kind of problem by putting forward crazy ideas – the lateral thinking that is advocated by Edward de Bono. But we should use some more normal cases too.

8 The use of intuition

Philosophers tend to create a thought experiment, consider different decisions that an agent in it might take, and then consider the acceptability of those decisions. And a leading way to assess the acceptability of decisions is to use intuition. Would this or that decision feel right?

We should ask whether this is a good way to work. There are two central questions. Can the intuitions of the people who contribute to philosophical discussions be trusted? And should we prefer philosophical expertise of a non-intuitive sort to intuitions?

8.1 Can intuitions be trusted?

If we want to know whether intuitions can be trusted, we can sensibly start by looking at examples of intuitions.

Sometimes we will find that the same intuitive answer to a given question “Would this be an appropriate decision for an agent in this thought experiment to take?” will be given by a wide range of people, across different occupational and social groups and across cultures, with very few dissenters. Those examples will not take us far. We have no external point from which to decide whether the widespread intuition is in fact appropriate, except perhaps the ethical theories that we are trying to test by reference to intuitions – and to accept or dismiss intuitions on the ground that they fitted or conflicted with the theories we sought to test would be circular.

At other times we may find that intuitive answers to a given question “Would this be an appropriate decision for an agent in this thought experiment to take?” will differ as between people, either randomly or, more interestingly, systematically across different social groups or across cultures. That sort of difference should unnerve us. It would mean that not all intuitions could be relied upon to give answers as to what to do or as to whether given ethical theories were satisfactory. And it would leave us wondering whether anybody’s intuitions were a reliable source of answers. (It might be that on each occasion someone was right, but that no-one was consistently right.) In particular, it would not be at all clear that the intuitions of those who contributed to philosophical discussions should be trusted.

The concern has become more acute in recent years, with the rise of experimental philosophy. Thought experiments are put in front of a wider range of people than philosophy professors, and the views of these people are taken. What we find is that at least some of the time, there is variation between the views of different groups of people. (This sort of variation is by no means confined to ethics. Indeed, it is arguably greater in epistemology, for example when people are confronted with Gettier-like cases. That might reflect the fact that ethics must respect the practicalities of life, while epistemology is distant from everyday concerns, allowing greater freedom to think in one’s own way.)

For example, there is the basic trolley problem. The runaway trolley is about to kill five people, but you can divert it so it kills one person on another track. When the one person is a child, women are more reluctant than men to divert the trolley. When it is an adult stranger, men are more reluctant than women to do so.[8]

Even before we get on to which actions are acceptable in a situation that strikes us as obviously morally challenging, there is the question of what conduct might be condemned as morally problematic (as distinct from conduct being merely harmful in some way or other). And this has been found to vary between cultures – examples include desecration of the national flag, homosexual acts in private, and eating a pet dog that has died of natural causes.[9] It is not that all members of one culture will go one way and all members of another one will go another way. Rather, the percentages who will see such acts as morally problematic will vary as between cultures.

[8] Zamzow and Nichols, “Variations in Ethical Intuitions.” *Philosophical Issues*, volume 19, number 1, 2009, pages 368-388 (see pages 371-372).

[9] Haidt, Koller and Dias, “Affect, Culture, and Morality, or Is It Wrong to Eat your Dog?” *Journal of Personality and Social Psychology*, volume 65, number 4, 1993, pages 613-628.

But the picture is not consistently one of wild variation. Sometimes it is found that there is not much variation. Thus although one study found some variations, it also found that effects of gender, education, politics, and religion were only modest.[10]

Despite such reassurance, there is still enough evidence of divergent intuitions to make us at least a bit concerned about the use of intuitions to reach conclusions. But could we perhaps argue that philosophers had especially good intuitions?

We could argue that philosophers spend a lot of time thinking and arguing about thought experiments, and studying the theories which are confronted with those experiments. They have also learnt to favour precision, to look out for ambiguities, and to draw out logical implications with care. This expertise and these qualities of thought are admirable, and have indeed been admired in literature which advocates paying attention to philosophers' intuitions. (We should not however be too optimistic about philosophers' skills. They do appear to be influenced by the order in which different scenarios are presented, just as non-philosophers are influenced.[11])

On the other hand, precisely this acculturation might be a problem. They are unlikely suddenly to start seeing things in a different way, even when that would help to advance their understanding. There are also the problems that within a lot of universities, philosophy departments are still strongly white and male, and that alternative philosophical traditions are often in different languages which may be hard to learn (Chinese, for example). If we could say that philosophy professors within single departments routinely came from wide ranges of cultures and individual backgrounds, we might have more faith in them as sources of expert intuitions – but they do not.

There is a further argument against relying on intuitions, whether of philosophers or of people generally. This is that their source may be something upon which we cannot reasonably rely to get us the right answers, as opposed to the convenient answers. The source might for example be evolutionary pressures. This line of argument has been developed by Peter Singer.[12]

8.2 Options for the use of intuitions

If intuitions are sometimes problematic, how might we respond?

One option would be to rely only on those intuitions that were very widely shared across cultures, not because they would be guaranteed to be correct but because they had the best prospect of being good ones to follow.

[10] Banerjee, Huebner and Hauser, "Intuitive Moral Judgments are Robust across Demographic Variation in Gender, Education, Politics, and Religion: A Large-Scale Web-Based Study." *Journal of Cognition and Culture*, volume 10, number 3, 2010, pages 253-281. The study was open to people from a range of cultures, but all those studied did speak English.

[11] Schwitzgebel, Eric, and Fiery Cushman. (2012). "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." *Mind and Language*, volume 27, number 2, 2012, pages 135-153.

[12] Singer, "Ethics and Intuitions." *The Journal of Ethics*, volume 9, numbers 3-4, 2005, pages 331-352.

Another option would be to rely only on intuitions which we could defend in the face of critical challenge. For example, an intuition that it was good to help any people in serious need when the cost of giving help was low might be defended. But an intuition that this only applied to helping family members, friends, or people of the same nationality would be hard to defend on rational grounds, as distinct from explaining such selectivity on evolutionary grounds.

Alas, such restrictions on the use of intuitions might not leave us with enough to go on. We might be deprived of ways to solve some important ethical problems, or be left unable to subject plausible ethical theories to sufficiently extensive or severe tests.

8.3 Philosophical expertise

Could we fill in the gaps by relying on philosophical expertise of a non-intuitive sort? And indeed, could we rely on such expertise to dispense with intuitions altogether, and save ourselves the trouble of sorting the intuitions we should use from the ones we should not use?

One can certainly argue for the importance of expertise, by analogy with other academic disciplines. There is such a thing as expertise in physics, or in history, so why not in philosophy? And it is possible to respond to experimental philosophy results which seek to undermine claims to expertise.[13] Moreover, a defence of the role of expertise by reference to parallels with other academic disciplines can itself move us away from a tight focus on intuitions, because it can be argued that the expertise in question is one of methods rather than one of having good intuitions.[14]

Such arguments may go quite some way in epistemology or metaphysics. It is not clear that they go very far in ethics. We need to recognize that ethics is different. It is not, or at least not obviously, the study of a world which is in whatever state it is in, independently of how or what we think about the world. ("How" and "what" are importantly different, but not entirely independent of each other.)

There is also likely to be a continuing role for intuitions in ethics, at least at the stage of evaluation of ethical theories. An ethic that did not feel right, even after we had thought long and hard about it and its implications, would be one we would find very hard to adopt.

So intuitions are here to stay. We can control their use by investigating how they and their implications harmonize with one another, with facts about the world, and with theories that we are inclined to accept, by investigating the concepts that are used to formulate them, and by working out their sources. Ideally, they will be accommodated within a reflective equilibrium of all of our thoughts about how to live. In reality, we can expect a few imperfections in our thought. But that is life.

[13] Williamson, "Philosophical Expertise and the Burden of Proof." *Metaphilosophy*, volume 42, number 3, 2011, pages 215-229.

[14] Nado, "Philosophical Expertise and Scientific Expertise." *Philosophical Psychology*, volume 28, number 7, 2015, pages 1026-1044.