Thought-experiments, Disagreement and Moral Realism

Sebastian Köhler

This paper is published in Grazer Philosophische Studien published by Rodopi. This is a post-print; that is, it is the full and correct text. The definitive version is available at Rodopi. The full citation is: Köhler, Sebastian 2010. Thought-Experiments, Disagreement and Moral Realism. In *Grazer Philosophische Studien* 80: 245-252

Abstract: : Thought-experiments are an important method in moral philosophy. But, what do we actually learn about morality through thought-experiments? In this paper I argue against a view that moral judgements about thought-experiments are judgements about a specific type of fact: a view which holds that moral facts are objective (independent of our beliefs, attitudes or conventions), but which also respects the intuition that if a moral question has an answer, then, in principle, we can discover this answer. I will argue that moral judgements about thought-experiments cannot be judgements about facts so characterized. I argue this on the basis that empirical studies show that the moral judgements people make about thought-experiments, which presumably are among our best epistemological methods with regards to finding out about moral matters, do not behave as they should if they were about that kind of fact.

Thought-experiments are an important method in moral philosophy. But, what do we actually learn about morality through thought experiments? In this paper I will argue that moral judgements about thought experiments are not judgements about a specific type of fact. To do this I will first say some words about thought experiments. Then I will present the position I want to argue against. Lastly, I will present an argument against this position and defend it against possible replies.

Thought experiments are descriptions of scenarios (see Cooper 2005, 329) that are used to answer questions about certain states of affairs by isolating the relevant features.

The position I will argue against in this essay I call **moral realism** (**MR** for short). MR subscribes to the following two theses:

> **MR1:** There are objective (independent of our beliefs, attitudes or conventions) moral facts which our moral judgements are about.

MR2: If a moral question has an answer, then, in principle, we can discover this answer.[1]

Proponents of MR would plausibly have to say that judgements people make about moral thought experiments are about the type of fact defined by MR1 and MR2. I will argue that this can not be true, because MR can not explain certain kinds of disagreement. I will now present my argument:

Thought experiments in moral philosophy are supposed to establish ideal epistemic conditions for moral judgements. By "ideal epistemic conditions for judgements about phenomenon X" I mean those conditions which, given the epistemic situation (e.g. the individual's epistemic capacities), provide the best guarantee that judgements about phenomenon X are correct. What are ideal epistemic conditions with regards to moral judgements? These are conditions in which, firstly, the non-moral features of the situation in question are clear, so that moral misjudgements on the grounds of misjudgements about the non-moral facts are ruled out (which are often responsible for moral misjudgements (see Brink 1984, 117). Secondly, they guarantee that people are sufficiently well-informed, clear-headed and rational, so that they can be expected to fully grasp the relevant non-moral facts and make rational judgements on their basis (another source of misjudgements (see Loeb 1998, 283). Thirdly, they guarantee that judgements are as unbiased by "self-interest, self-deception, historical and cultural accident, hidden class bias" and other such factors as possible (Daniels 1979, 265).

With regards to other empirical or non-empirical phenomena we can discover, in principle, it is plausible that under parallel conditions there will be an agreement close to 100% on one option. Suppose a group of individuals was asked whether a certain object has a certain colour. Assume ideal epistemic conditions hold: favourable light conditions and other such conditions hold so that

[1] One word of defense with regards to MR2: MR2 is a fundamental assumption of our ordinary moral discourse (see Smith 1994, 5/6 or Loeb 1998, 290/291). Whenever we enter moral discussion, we presuppose that the issue in question has an answer and *we* are capable of finding it. We can, therefore, assume that MR2 is a plausible thesis a proponent of MR1 will want to accept.

the factual situation is clear, all individuals are sufficiently capable of discovering the fact in question, and no individual has a personal bias that could disturb her judgement with regards to the question. This means we assume that all individuals are in an ideal position to answer the question given their epistemic capacities. Under these circumstances, we would expect an agreement close to 100% on one answer to the question.

Now, let's take the case of a non-empirical issue. Suppose a group of individuals is faced with a mathematical problem. Again assume ideal epistemic conditions hold: the features of the problem are all clear, all of them are capable with regards to skill and their cognitive capacities to answer this kind of mathematical problem, and there is no personal bias that could disturb the judgement. Under these circumstances we would again expect an agreement of close to 100% on one answer to the mathematical problem.

Does the same hold with regards to moral judgements when made facing a thought experiment? Take two variants of a thought experiment in moral philosophy: They are versions of the "Trolley Case" (for the original version see Foot 1967, 23; my variant was introduced by Thomson. See Thomson 1985, 1403):

> **Trolley Case 1 [TC1]:** Anne sees a trolley running down track A. She can spot five people on A who are unable to get out of the way before the trolley will hit and kill them. She recognizes that it is impossible to stop the trolley on its way down A. However, she can divert the trolley on track B by flipping a switch. Although B loops back onto A, so that the five would be killed if the trolley would continue on its way, there is a single person X on track B unable to get from the track. If the trolley kills X, it will be stopped which prevents the death of the five.
>
> **Trolley Case 2 [TC2]:** TC2 is identical to TC1 except that behind X is a heavy stone which would prevent the trolley from running into the five, even if X were not on B.

The question people proposing these cases are asking is: "Is it permissible or forbidden to flip the switch?" It is clear that what is asked for is a moral judgement. Now, consider the following two judgements:

> **Judgement A [JA]:** In both cases it is permissible to flip the switch.
>
> **Judgement B [JB]**: In TC1 it is forbidden, while in TC2 it is permissible to flip the switch.

JA and JB are two of the responses people give to these cases. In an experiment by Hauser et al. 56% judged it permissible to flip the switch in TC1, while 44% judged it forbidden. In the case of TC2 72% judged it permissible to flip the switch, while 28% judged it not permissible (see Hauser et al. 2007, 6). This means that a significant number saw the intuitive difference proposed by JB, while another significant number did not see any difference, following JA.[2] This shows that there are people who support either judgement.

Now, if there are thought experiments in moral philosophy providing ideal conditions for judgements, it should be TC1 and TC2. It is obvious which properties the scenarios have, which are morally relevant, all possibly disturbing factors are filtered out. Nevertheless, JA and JB are both supported by a significant number of individuals. So, even in this example of thought experiments providing ideal epistemic conditions, no agreement close to 100% can be found. This disagreement can not be constructed as conceptual (which is Sosa's strategy to explain away disagreements with regards to thought experiments (see Sosa 2007, 102-104)), since it is a genuine *normative* disagreement. It can also not be explained away by saying that the issue has no answer or is indeterminate. "Where there is indeterminacy it should be possible, at least in principle, to recognize it as such […]." (Loeb 1998, 291). If we are able to answer all answerable moral questions in principle, we should also be able to discover in principle whether a question has no

[2] Since there is no intelligible reason to judge that flipping the switch is at least permissible in TC1, but forbidden in TC2 (why in the world should the existence of a stone make the action forbidden?), we can conclude that 56% of the individuals followed JA, while 16% to sides with JB. 28% of the individuals followed a position which we might call "judgement C" which proposes that we judge flipping the switch forbidden both in TC1 and TC2.

answer or an indeterminate answer. But, no one considering theses cases thinks that the issue might be indeterminate or have no answer, especially not those working professionally on them (see Kamm 1989 or Thomson 1985). So, it is implausible that the issue has no answer, or an indeterminate answer.

One could reply that the disagreement could be accounted for by arguing that the issue is *epistemically* so difficult that it is unlikely that everyone will be able to come to the right result, even given ideal epistemic conditions. Both cases involve a moral dilemma with two morally undesirable options and therefore demand a decision that is at least in some respects morally undesirable. If we consider how this might arguably complicate the epistemic issue, we can see why disagreement occurs. However, if we consider easier questions, for example whether it is permissible to torture innocent children for pleasure, we will find agreement close to 100%. The problem is not that there is no answer to the question, but that the answer is hard to figure out, even under ideal epistemic conditions.

Now, although I admit that the issue in question is harder than others, I disagree that its difficulty suffices to explain the disagreement. Of the difficult moral questions, it is still among the easier: Only a limited number of moral considerations could be relevant and it is clear which these are. And the considerations relevant are quite familiar to common moral thought: Under which circumstances, if ever, it is permissible to harm some for the sake of others. Since the features to be taken into account are limited to a few very familiar ones, it seems implausible that the evaluative question could be hard to answer. It is true that some moral questions might be so hard to answer that this could explain disagreement with regards to them. But, it seems that the issue here (and in many other interesting cases) is not among them. However, given what I have said about the epistemic situation with regards to TC1 and TC2, it is highly probable that in no such case we will find an agreement close to 100% on one of the possible alternatives. If no other explanation for this

can be brought forward, this indicates that thought experiments in moral philosophy are not about facts as specified by MR, since even given these epistemic conditions no agreement will be found.

What could a proponent of MR reply to this? He could point out that moral judgements are more like theoretical than basic judgements in science (see Loeb 1996, 288 or Daniels 1979, 395/396). "Theoretical judgements" are those which are justified just because of their role in a successful theory. "Basic judgements" can be justified on their own, because we have some kind of sufficient direct epistemic access to the relevant facts, provided there are not too many disturbing factors. Since moral judgements are theoretical we should not expect agreement close to 100%, since such can also not be expected with theoretical judgements in other sciences.

This reply however is not convincing. Even if most moral judgements could be characterized as theoretical, it is doubtful that moral judgements made about thought experiments are such. They are extremely spontaneous and people are often unable to provide sufficient justification for them (see Hauser et al. 2007, 14-15), but are nevertheless attributed at least initial credibility (as can be seen from various examples in moral philosophy. See e.g. Kamm 1989, Parfit 1997 or Rawls 1971 among many others). And, it does not seem that we need to take into account a moral theory to come to these judgements or take them seriously. If there are basic moral judgements, then such judgements should be counted among them. So, these judgements should not be regarded as theoretical. What else could a moral realist reply? He could propose that at least one of the conditions which constitute ideal epistemic conditions for moral judgements does not hold.

First, he could deny that the **first condition** holds: He could argue that thought experiments never provide sufficient information of the non-moral facts for a determinate moral evaluation. Given this, different moral evaluations of the same thought experiment will always be possible. An argument along these lines was proposed by Jonathan Dancy (see Dancy 1985, 144/145): Every imaginary case is indeterminate, because they do not exist beyond their description and every such description is incomplete. This incompleteness can only be unproblematic with regards to moral

evaluation, if we stipulate that the case has no further morally relevant properties. Now, the moral evaluation of a case results in part from its non-moral properties. However, all non-moral properties which do not *directly* determine the moral evaluation, might do so indirectly. "[S]uppose that the action is right […] in virtue of its generosity, thoughtfulness and kindness […]. But each of these three will in turn result from, or exist in virtue of, further properties." (Dancy 1985, 144) This means that every property of a scenario could be relevant for the moral evaluation. The question is, whether "if we have been told the properties which *de facto* constitute the reasons why the action is right, and nothing much more than these, can we reasonably be expected to form a sound view on the question whether the action is right or wrong?" (Dancy 1985, 145; italics by the author) Dancy's answer is "No". Since the moral properties of a case are influenced by those which they directly result from, but also by many other properties which are not determinately set by the description, the evaluation has to be "indeterminate because it is crucially dependent upon others matters which are so far indeterminate." (Dancy 1985, 145) Nor is it possible to remove this indeterminacy, since every description is incomplete.

However, this does not threaten my argument, since, even if Dancy's argument goes through, we could use thought experiments to figure out moral conclusions. We could still use them, for example, to establish whether a difference between two cases with regards to a single property would make a moral difference. This is exactly what is done in TC1 and TC2: Both cases are identical except for one detail. The question is whether this would make a moral difference. It is not relevant for this to have a complete description of the two scenarios, since we are only interested in the difference this change in a single property would make. It is not even relevant that we know which specific evaluation for each situation would be right, but only whether our evaluations of the two cases would differ and, therefore, that the property in question is morally relevant.[3] And it

[3] This is actually a very common use of thought experiments in moral philosophy. It can be found in the original trolley problem (see Foot 1967), but also e.g. in the discussions about egalitarianism (the levelling down objection is an example for this. See Parfit 1997, 210/211), utilitarianism (e.g. in the case of Nozick's pleasure machine. See Nozick 1974, 42-45) or freedom of choice (see e.g. Sen's example in Sen 1988, 271) to name only some.

seems hard to conceive that thought experiments do not provide ideal conditions to answer these kinds of questions. They provide descriptions of situations where it is clear which feature is the one to be looked at and where disturbing factors are effectively ruled out. If MR were true, then thought experiment like TC1 and TC2 would be the method to provide ideal epistemic conditions for such moral judgements. Nevertheless, disagreements can be found with regards to these uses which can not be explained by MR.

Secondly, the proponent of MR could argue that the **second condition** does not hold: Thought experiments can only guarantee that the first condition holds, not that the people making the judgement are sufficiently capable of making reliable judgements. However, we can not expect in cases like the one I presented above, where people with a variety of backgrounds and levels of education where supposed to make a judgement, that everyone will be sufficiently well-informed, clear-headed, rational, etc. So, we should not be surprised if those people disagree.

This line of argument however is not convincing: Following it we would have to say that with regards to TC1 and TC2 *all the people* subscribing to JA or *all the people* subscribing to JB were not well-informed, clear-headed or rational. But, this does not seem plausible: Among both groups are surely people who can be regarded as well-informed, clear-headed, rational, etc. on any reasonable formulation of these conditions. Furthermore, we should not forget that there are *professional* debates about TC1 and TC2 in which people take sides with JA or JB who are *surely* among the most well-informed on the issue, and are most likely clear-headed, rational, etc. on any reasonable formulation of the conditions (see e.g. Thomson 1985, 1402/1403, Kamm 1989, 230 or Richardson 2008, 88).

Lastly, a proponent of MR could claim that the **third condition** does not hold: There are disturbing psychological or sociological factors which are not effectively ruled out in these cases that influence moral judgements. This explains why we find disagreement when people face moral thought experiments.

Now, although such factors do have a disturbing influence on moral judgements, it seems implausible that they can explain the constant occurrence of disagreement, not only in everyday moral discussion, but also in professional philosophical debates. First, think about the disturbing influence of these factors with regards to the discovery of other kinds of facts. Given that there *are* moral facts, we are in principle capable of discovering these facts, have a method which provides us with all the relevant factual details, and a working scientific community exists which controls the results of particular experiments the influence of these factors can be expected to be ruled out at least in the long run (see Nagel 1979). This means that given such a community, an agreement close to 100% should be achieved at least in the long run using the best epistemic devices for such purposes. Since this community actually exists for moral questions and since thought experiments are such epistemic devices, it is implausible that the named psychological and sociological factors can explain why there is such constant disagreement, even when people are well-informed, clear-headed, rational, and use thought-experiments. Second, it should be noted that the number of people taking sides with each moral judgement is constant over a large number of sociological factors: it made no significant difference whether the subjects had been exposed to moral philosophy, which type of education they had received, how old they were or what gender they had (see Hauser et al. 2007, 11). In each case around 55% judged flipping the switch permissible in TC1 and around 70% judged it permissible in TC2. This indicates that the relevant subjects *were not* influenced in their judgements too strongly by certain sociological or psychological factors. So, these can equally be ruled out as disturbing the moral judgement.

I conclude that all replies of the proponent of MR fail and he has no convincing explanation for the constant disagreement with regards to the moral evaluation of thought experiments. The thesis of a proponent of MR with regards to thought experiments is therefore false: moral thought experiments are not about such facts defined by MR1 and MR2. But, if thought experiments are indeed among the best epistemic devices for moral matters, this casts strong doubt on MR itself.

Bibliography

1. Brink, David O. 1984: "Moral Realism and the Sceptical Arguments from Disagreement and Queerness". *Australasian Journal of Philosophy* 62, 111-125.
2. Cooper, Rachel 2005: "Thought Experiments". *Metaphilosophy* 36, 328-347.
3. Dancy, Jonathan 1985: "The Role of imaginary Cases in Ethics". *Pacific Philosophical Quartely* 66*,* 141-153.
4. Daniels, Norman 1979: "Wide Reflective Equilibrium and Theory Acceptance in Ethics". *Journal of Philosophy* 76, 256-282
5. Foot, Philippa 1967: "The Problem of Abortion and the Doctrine of Double Effect". In: Philippa Foot (Ed*.)*, *Virtues and Vices and other Essays in Moral Philosophy*. Oxford: Oxford University Press, 19-32.
6. Hauser, Marc, Cushman, Fiery, Young, Liane, Kang-Xing Jin, R. and Mikhail, John 2007: "A Dissociation Between Moral Judgments and Justifications". *Mind and Language* 22, 1-22.
7. Kamm, Frances M. 1989: "Harming Some to save Others". *Philosophical Studies* 57, 227-260.
8. Loeb, Don 1998: "Moral Realism and the Argument from Disagreement". *Philosophical Studies* 90, 281-303.
9. Nagel, Ernest 1979: "The Value-Oriented Bias of Social Inquiry". In: Michael Martin and Lee C. McIntyre (Ed.), *Readings in the Philosophy of Social Science*. Cambridge MA: The MIT Press, 571-584.
10. Nozick, Robert 1974: *Anarchy, State, and Utopia*. Oxford: Blackwell.
11. Parfit, Derek 1997: "Equality and Priority". *Ratio* 10, 202-221.

12. Rawls, John 1972: *A Theory of Justice*. Oxford: Clarendon Press.
13. Richardson, Henry S. 2008: "Discerning Subordination and Inviolability: A Comment on Kamm's *Intricate Ethics*". *Utilitas* 20, 81-91.
14. Sen, Amartya 1988: "Freedom of Choice. Concept and Content". *European Economic Review* 32, 269-294.
15. Smith, Michael 1994: *The Moral Problem*. Oxford: Blackwell.
16. Sosa, Ernest 2007: "Experimental Philosophy and Philosophical Intuition". P*hilosophical Studies* 132, 99-107.
17. Thomson, Judith Jarvis 1985: "The Trolley Problem". *The Yale Law Journal* 94, 1395-1415.