Analysing Thought Experiments

Jan Willem Wieland & Matthijs Endt

Philosophers such as Gettier, Frankfurt, and Thomson are famous for their thought experiments. This makes one wonder: how did they invent their cases? Were they just lucky to devise a good case, or did they follow some basic rules that are available to all of us? In this paper, we argue for the latter answer by presenting a guidebook for analysing thought experiments. Our guidebook clearly specifies which factors should be included in a thought experiment, and which factors should be left out. This will help students to see through the fantastical elements of TEs, to learn the practice, and to check whether philosophers are doing things right. We illustrate our account in some detail using examples from Thomson's thought experiments.

Keywords: thought experiment; counterexample; sufficient; necessary

1. Introduction

Thought experiments (henceforth 'TEs') are widely employed in all domains of philosophy. In ethics we have the Frankfurt cases, in epistemology we have the Gettier cases, in metaphysics we have Black's spheres, and in the philosophy of language we have Kripke's Gödel. And there are plenty more beyond these.[1]

Typically, TEs are imaginary scenarios in which a certain theory is tested, i.e. the theory's verdict about a given scenario is compared to our intuitive verdict about that scenario. As an illustration of this, consider the well-known trolley case (first proposed by Foot 1967 and further developed by Thomson 1976). This is an imaginary scenario devised to test the theory that killing others is worse than letting them die (call this theory 'T'). In this scenario, you are a passenger on an out-of-control trolley and have only two options: you can either let five people die, who happen to be on the track in front of you, or else you can turn the trolley and thus kill just one person who happens to be on the side track. According to T, it is not

[1] Respectively: Frankfurt (1969); Gettier (1963); Black (1952); Kripke (1970: 83-4). On the basis of their TEs, Frankfurt challenges a view on responsibility, Gettier a view on knowledge, Black a view on identity, and Kripke a view on reference. In this paper, we'll focus on TEs in philosophy, leaving aside TEs in the sciences.

permissible for you to turn the trolley. Intuitively, however, it does seem permissible (if not obligatory). Moreover, if in this case killing one person isn't worse than letting five people die, we have a counterexample to T.

In this paper, we'll be interested in TEs insofar as they are able to expose counterexamples in this way.[2] Counterexamples can be of two kinds. They can challenge a *necessary* condition, or a *sufficient* one. And, as we all know, the difference matters (though the way in which it matters in the context of devising TEs might not be apparent at this point).[3]

Sufficiency. A is sufficient for B when B always occurs when A does. A counterexample to such a claim always has the same form: **A without B**. After all, if we ever do have an A without a B, then B does not always occur when A does, and A is insufficient for B.

Example: suppose your opponent claims that 10 beers are sufficient for a hangover. Thus, whenever you drink 10 beers, you are supposed to get a hangover. How can you find a counterexample? Find a case where one drinks 10 beers, but gets no hangover. Or suppose your opponent claims that justified true belief (JTB) is sufficient for knowledge. Thus, whenever there's JTB, there should be knowledge. How can you find a counterexample? Find a case where there's JTB without knowledge.[4]

Necessity. A is necessary for B, in contrast, when B can only occur if A does (so A always occurs when B occurs). A counterexample to such a claim always has the same form: **B without A**. After all, if we ever do have a B without an A, then A does not always occur when B does, and A is not necessary for B.

Example: suppose your opponent claims that 5 beers are necessary for a hangover. Thus, whenever you get a hangover, you had 5 beers. How can you find a counterexample? Find a case where one gets a hangover without having had at least 5 beers. Or suppose your opponent claims that alternative possibilities are necessary for responsibility. Thus, whenever there's responsibility, there should be

[2] There are also more open-ended TEs, which merely invite one to think about a certain case (which could be analogous to a real-life situation). In addition, one might suggest that TEs not only have the negative function of falsifying theories, but also the positive one of aiding in the discovery (and confirmation of) new theories. We critically discuss the latter suggestion in §3.

[3] In this paper, we'll adopt the standard view, according to which necessary and sufficient conditions are understood in terms of truth-functions (cf. Brennan 2003).

[4] Throughout the paper, we make references to the TEs by Gettier and Frankfurt. Ideally, this paper is read in tandem with their texts. We can't discuss all TEs in all detail, though we'll do this for Thomson's TEs in §4.

alternative possibilities. How can you find a counterexample? Find a case where there's responsibility without alternative possibilities.

Philosophical accounts are all about A and B, where B is the phenomenon of interest (such as knowledge or responsibility), and A is the factor that is supposed to be necessary and/or sufficient for that phenomenon (such as JTB or alternative possibilities) and thus supposed to provide a partial analysis of it. This idea is meant to be fully general and apply not only to the Frankfurt and Gettier cases, but also to Black's spheres and Kripke's Gödel, for example. In Black's TE, indiscernibility is rendered insufficient for identity. And in Kripke's TE, descriptions are rendered insufficient for reference.

To summarize. In order to challenge A's sufficiency, you need A without B. In order to challenge A's necessity, you need B without A. But this is just the first step. Knowing these things won't make you a Gettier or a Frankfurt. After all, knowing that you need A without B or B without A is not the same as knowing *how to get these*. The main purpose of this paper is to make that further step.[5]

Our plan in this paper is as follows. In §2, we'll summarize some main current accounts of TEs, and point out that the question we're raising here has hardly been addressed. In §3, we'll present our guidebook on how TEs can be made. Our guidebook is designed for students. It'll help them to see through the fantastical elements of TEs, to learn the practice, and to check whether they and others are doing things right. Surprisingly enough, the guidelines we propose have never been made precise and put on paper. And yet, philosophers are supposed to comply with them every day. In §4, we'll put our account to work, and illustrate it in some detail on the basis of Thomson's TEs.

2. Existing accounts

Surveying some important recent accounts of TEs, we find that none of them really address the issue of how TEs can be made (apart from a recent study we'll discuss at the end).

Williamson's influential (2007) proposal endorses the counterexample function of TEs just introduced. As Williamson writes, the function of TEs is "to show that a certain case could arise, and that if it did, the two things would come apart, from which it follows that the two things could come apart. That refutes the modal claim that they could not come apart" (2007: 185). For example, Gettier's TEs involve

[5] For the purpose of this paper, all other issues about TEs will be left aside (including the psychology of TEs, or the nature and role of intuitions).

possible cases where JTB and knowledge come apart, and hence refute the modal claim that they necessarily go together.

Häggqvist (2009) makes a comparable suggestion. According to him, the function of TEs "may seem vaguely Popperian in its emphasis on counterinstances to the target thesis" (2009: 64). For example, the trolley case forms a counterinstance to the target thesis that killing others is worse than them letting die ('T'). According to Häggqvist, furthermore, many TEs can be reconstructed along the following lines: (1) a certain counterfactual scenario C is possible; (2) theory T predicts a result W in C; (3) but, intuitively, W is false in C; hence: (4) T is false.[6] In terms of the trolley case, the reconstruction would run as follows:[7]

(1) Possibly, the trolley passenger has a choice between killing one person and letting five people die.

(2) If T were true, then, if the passenger had a choice between killing one person and letting five people die, she should choose to let the five die.

(3) But if she has this choice, then she may choose to kill the one.

(4) Therefore: T is false.

As Häggqvist explains, such conclusions can be resisted in exactly three ways, corresponding to the three premises: one could try to deny (1), (2) or (3).[8] To deny (1) is to say that C is impossible (the 'impossibility defence'). In the trolley case, this amounts to claiming that the passenger couldn't have such a choice between killing one person and letting five people die. To deny (2) is to say that T does not predict W in C (the 'irrelevance defence'). In the trolley case, this amounts to claiming that, according to T, the passenger should not choose to let the five die (but kill the one). To deny (3) is to say that W is actually true in C ('biting the bullet'). In the trolley case, this amounts to claiming that the passenger should choose to let the five die.[9]

As we see it, these accounts provide useful metaphilosophical insights, and explain what TEs do (namely, provide counterexamples) and how one might respond to them. Yet they are silent about how TEs can be devised. In terms of Häggqvist's account, the question is how one can devise TEs for which (1)-(3) are true so that

[6] This model is inspired by Sorensen (1992: ch. 6). Following Häggqvist (2009: 60-2), we'll assume that TEs are experiments (consisting of steps you have to go through), which are not identical, but still closely related, to reconstructions of them in the form of arguments.

[7] Our reconstruction closely follows the logical structure of Häggqvist's premises.

[8] Gendler (2000: 22) identifies three comparable options, focussing more on the analogical aspects of TEs.

[9] Clearly, to say that these three options are always *possible* does not mean that they are always *plausible* (cf. Häggqvist 2009: 67).

they constitute a counterinstance to T. At this point, one might be skeptical and suspect that philosophers don't say much about the issue simply because there's nothing to be said. There might be no general story to tell about devising TEs. In the following, we will, to a reasonable extent, challenge this skepticism.

To begin with, it is useful to point out that TEs are, after all, *experiments*. Generally, in experiments one isolates a certain variable A in order to prove something about its relation to another variable B (the phenomenon of interest). As Sorensen puts it: "An experiment is a procedure for answering or raising a question about the relationship between variables by varying one (or more) of them and tracking any response by the other or others … For example, Benjamin Franklin answered 'Which colors absorb the most heat?' by laying out squares of cloth on the snow. After a few hours in the sunshine the squares had sunk into the snow at depths that increased with the darkness of each square." (1992: 186)

The same holds for TEs. In TEs, one also isolates a certain variable A in order to prove something about its relation to another variable B. In Gettier's TE, we want to understand the link between JTB and knowledge, and to this end we isolate these factors in the imagination. In Frankfurt's TE, we want to know the link between alternative possibilities and responsibility, and to this end we isolate these factors in the imagination. This does not mean, of course, that TEs are exactly the same as other kinds of experiments. For one thing, they are conducted in the imagination rather than the external world.

Sorensen is not alone in making this comparison. Consider for example the following passages: "A primary goal of abstraction and schematization in moral reflection is to create the analogue of 'controlled experiments' in science: one wants to hold all other factors fixed, and test one particular factor for ethical relevance." (Fischer 1995: 10) And: "In philosophical cases, when we explore conceptual dependencies, we do thought experiments [and] test various hypotheses by considering cases in which we systematically vary the possible contributing factors." (Gendler 2000: 27)

But the question is what these claims amount to, concretely. If one wants to devise a TE in order to test the relation between different factors (such as between JTB and knowledge), how should one proceed? Specifically: *which factors should be held fixed, and which should be varied?*

If TEs are among the philosopher's main methods (as we think they are), then this underexplored issue is clearly pertinent. To our knowledge, Praëm & Steglich-Petersen (2015) are the first to touch upon the question we're raising here. Strictly speaking, they are not so much interested in how to devise TEs as in theory discovery (on the basis of TEs), but in passing they do make some suggestions concerning how TEs might be devised. Just as we have done here, Praëm &

Steglich-Petersen make use of the distinction between necessary and sufficient conditions.

Concerning TEs that refute sufficient conditions, they say that these can be devised by identifying a case where “the presence of some concrete factor hinders the conditions of the target analysis from resulting in the target property” (2015: 2837). To devise a Gettier TE, for example, one may add the factor ‘false grounds,’ which hinders the conditions of the target analysis (JTB) from resulting in the target property (knowledge). In Gettier’s example, Smith believes that the man who will get the job has ten coins in his pocket, which is true, and he is justified to believe this, but he believes it on false grounds, and this hinders the case from resulting in knowledge.[10]

Concerning TEs that refute necessary conditions, Praëm & Steglich-Petersen say that these can be devised by identifying a case where the presence of some concrete factor “renders a condition [of the target analysis] irrelevant” for the target property (2015: 2838). To devise a Frankfurt TE, for example, one may add the factor ‘the agent’s will,’ which renders a condition of the target analysis (alternative possibilities) irrelevant with respect to the target property (responsibility). Jones kills Smith of his own will, and this renders irrelevant the fact that he lacks alternative possibilities (due to the possible intervention of the manipulator Black).

In our view, these guidelines are on the right track, but the problem is that they’re rather metaphorical. What does it mean to add factors that ‘hinder’ something from resulting into something else, or ‘render’ something irrelevant? Our aim in the following is to build on Praëm & Steglich-Petersen’s suggestions and make the steps for devising TEs more precise. These instructions, moreover, will enable students to analyse existing TEs (as we’ll illustrate in the subsequent section).

3. Guidebook

As we’ll propose, TEs can be devised in four steps. Here’s the first step:

***Step 1*: Identify the theory of your opponent.**

We’ll assume that you’re looking for a counterexample against your opponent’s theory (though, of course, it’s also possible to see whether there’s a counterexample against your own theory). The first step, then, is to determine the phenomenon your

[10] Cf. Clarke (1963), and the subsequent debate (summarized in Praëm & Steglich-Petersen 2015: 2832-5).

opponent is interested in (knowledge, responsibility, freedom, identity, etc.), and whether she proposes a necessary or sufficient condition for this phenomenon (or both).[11] Let's consider sufficient conditions first. In that case, your opponent claims: A is sufficient for B. To obtain a counterexample to this, you'll need an A without B. That's what you already know:[12]

***Step 2a*: Keep A in your case, but not B.**

So, you also want to remove B from your case. But you can't do that directly. Your opponent might not be impressed by a case where A remains and you merely state that you've removed B. For your opponent would just say if there's A, there's B. After all, that's her theory. Of course, your opponent might be curious and wonder why you think there's no B, but so long as you don't have any further story on why there's no B, the challenge is not very interesting. Rather, you'd have to do the following:

***Step 3a*: Remove another factor, distinct from A and B, that can plausibly be considered necessary for B.**

If you do this, B will disappear. For if this factor is no longer present, and it's necessary for B, B is no longer present. Even your opponent will feel the pull of this (at least, if she agrees that your added factor does indeed seem necessary).

Let's illustrate this on the basis of Gettier's TE. Step 1: the opponent's claim is that JTB is sufficient for knowledge. To challenge this claim, what you'll need is JTB without knowledge. Step 2a: keep JTB. Now remove knowledge by doing the following. Step 3a: remove another factor that can plausibly be considered necessary for knowledge. A factor that's removed from Gettier's TE is 'true grounds'. Smith believes that the man who will get the job has ten coins in his pocket. He is justified in believing this proposition since he derives it from his belief that Jones is the man who will get the job and that Jones has ten coins in his pocket (for which he has strong evidence). Moreover, his belief is true since the man who will get the job (Smith himself, not Jones) has ten coins in his pocket. But, the grounds on which he holds this belief are not true: Jones isn't the man who gets the job. Since the factor of true grounds appears to be necessary for knowledge, and it has been removed from the case, knowledge is removed as well.

[11] Typically, philosophical theories are modalized: "necessarily, B only if A" or "necessarily, B if A", in which case counterexamples need not be actual cases, but might be merely possible. This qualification will remain implicit in what follows.

[12] We'll use step 'a' for sufficient conditions, and 'b' for necessary conditions.

The factor of true grounds was identified by Clarke (1963), and one might want to object that Gettier couldn't have made his TE in this way because Clarke's paper was written only *after* Gettier's paper. True enough. Importantly, however, our guidebook does not describe how philosophers *in fact* made their TEs, but rather how anyone can devise them (that is, in a non-lucky way, as we'll explain below).

The guidelines differ if your opponent endorses a necessary condition. This time, A is considered necessary for B. To obtain a counterexample to this claim, you'll need a B without A. That's what you already know:

***Step 2b*: Keep B in your case, but not A.**

So, you also want to keep B in your case. As before, you can't do that directly. Your opponent will likely not be impressed by a case where there's no A, and you merely state that B is still present. For your opponent can just say if there's no A, there's no B. After all, that's her theory. Rather, you'd have to do the following:

***Step 3b*: Add another factor, distinct from A and B, that can plausibly be considered sufficient for B.**

If you do this, B will be present. For if that additional factor is sufficient for B, then A is no longer needed for B. Even your opponent will feel the pull of this (at least, if she agrees that your added factor does indeed seem sufficient).

Let's illustrate this on the basis of Frankfurt. Step 1: the opponent's claim is that alternative possibilities are necessary for responsibility. To challenge this claim, what you'll need is a responsibility case without alternative possibilities. Step 2b: remove alternative possibilities. Now keep responsibility by doing the following. Step 3b: add another factor, or make it more salient, that can plausibly be considered sufficient for responsibility. A factor that Frankfurt added, or made more salient, is the agent's own will. Jones wanted to kill Smith, and did what he wanted. He didn't know that doctor Black would have intervened, had he decided to refrain from executing his plan (and that he didn't have alternative possibilities). But he didn't refrain, and Black didn't have to raise his hand. Here, Jones commits a wrongful act and can be held responsible for it, even though he didn't have alternative possibilities, since he acted of his own will.[13]

To summarize:

[13] "If he does it on his own, however, his moral responsibility for doing it is not affected by the fact that Black was lurking in the background with sinister intent." (Frankfurt 1969: 836)

> *Sufficiency*: You'll need to keep A fixed (i.e. the factor that is supposed to be sufficient for B), and remove another factor C that can plausibly be considered necessary for B.
>
> *Necessity*: You'll need to keep B fixed (i.e. the factor for which A is supposed to be necessary), and add another factor C that can plausibly be considered sufficient for B.

Our sufficiency guideline can be seen as a more precise rendering of Praëm & Steglich-Petersen's suggestion that the TE deviser 'add some concrete factor that hinders the conditions of the target analysis in resulting in the target property', and our necessity guideline can be seen as a more precise rendering of their suggestion that the deviser 'add some concrete factor that renders a condition of the target analysis irrelevant for the target property'.

Some brief clarifications are in order. First, the guidelines just presented do not imply, of course, that the new factors 'true grounds' (in Gettier's case) and 'the agent's will' (in Frankfurt's case) are not themselves vulnerable to further counterexamples. On the contrary, they have generated whole debates.

Second, you might think there's no real difference between the sufficiency and necessity guidelines. After all, A is sufficient for B if and only if B is necessary for A.[14] JTB is sufficient for knowledge if and only if knowledge is necessary for JTB. Alternative possibilities are necessary for responsibility if and only if responsibility is sufficient for alternative possibilities. Even so, the difference shouldn't be underestimated. For in the sufficiency case you must *remove* the phenomenon of interest, such as knowledge, from the TE (Gettier's case is a case *without* knowledge), while in the necessity case you must *keep* the phenomenon of interest, such as responsibility, in the TE (Frankfurt's case is a case *with* responsibility).

Third, step 3 (a or b) can be tricky. It's the most difficult step. If you have no idea what alternative factors the phenomenon of interest might be sensitive to, then relevant cases can be found only on the basis of guesswork (by playing with certain variables in a random way).[15] In other words, you do need an alternative theory to devise TEs in a non-lucky way. To devise a TE against a theory of the form 'B if A', one needs an alternative theory of the form 'B only if (also) C', and to devise a TE

[14] For discussion of this reciprocity claim, cf. Gomes (2009).

[15] One might suspect that guesswork is how it *always* goes. We disagree: philosophers might *think* they're just playing around with variables in a random way, though, if all goes well, they're actually carrying out these steps.

against a theory of the form ‘B only if A’ one needs an alternative theory of the form ‘B if C’ (we’ll provide further illustrations of such theories soon).

This point is overlooked by Praëm & Steglich-Petersen (2015). Their view is that TEs aid in theory discovery, rather than that alternative theories are needed to devise TEs in the first place, as we’re claiming here. As they themselves acknowledge (p. 2836), one can’t have one’s cake and eat it too: either theories come first, or else TEs do. One already needs a theory to devise a TE, or else one already needs a TE to discover a theory. On our proposal, when it comes to constructing counterexamples in a non-lucky way, theories do have to come first. Perhaps one doesn’t need a full and explicit formulation of the theory in the form ‘B only if (also) C’ or ‘B if C’, though one does have to be sufficiently sensitive to the possible relevance of that alternative factor C.[16]

Fourth, some TEs include all sorts of concrete details, while others are fairly abstract. The TEs by Gettier and Frankfurt are of the latter sort. The cases by Thomson that we’ll consider next have more details (although, in a certain way, they’re still quite abstract). Why are TEs so ‘streamlined’ (as Fischer 1995 puts it)? In principle, we don’t need to know the names of all the protagonists in a case. We don’t need to know what music they like, or who their friends are. We don’t need to know all these things, that is, so long as the phenomenon of interest is insensitive to them. What matters, though, is that you do the following:

> ***Step 4*: Provide enough information that readers can correctly interpret the factors of the relevant theories.**

As before, this step differs in the two cases. In the sufficiency case, you’ll need to provide enough information of the presence of A (‘justified true belief’) and of the absence of the alternative factor C (‘true grounds’). In the necessity case, you’ll need to provide enough information of the absence of A (‘no alternate possibilities’) and of the presence of the alternative factor C (‘the agent’s will’). All further information is irrelevant[17] and should be left out, and that’s why TEs are so streamlined.

Particularly, and importantly, information about the phenomenon of interest B (‘knowledge’ or ‘responsibility’) should *not* be included in the case description. That is, the reader of the TE is invited to answer to the question “B?”, and the answer should not be stipulated by the case description. If all goes well, in the sufficiency

[16] Still, we’d agree that things might be different for more open-ended TEs that aren’t immediately meant as counterexamples.

[17] And may even trigger effects on our judgments which might be irrelevant (cf. Nichols & Knobe 2007).

case, the reader agrees that C is necessary for B, and hence that the absence of C leads her to answer that B is absent as well ('Smith doesn't have knowledge'), and in the necessity case, hopefully the reader agrees that C is sufficient for B, and hence that the presence of C leads her to answer that B is present as well ('Jones is responsible'). Hence, providing enough details is not just to decorate the cases, it's needed to lead the reader through these steps.[18]

That's it. That is how TEs are to be devised. We don't claim to have solved all important issues concerning devising TEs, though we do claim to have made some important steps in this underaddressed topic. In the following, we'll devote some time to further clarifying our guidebook by considering some illustrative examples from Thomson's TEs. In our view, this is important: substantial metaphilosophical results make sense only on the basis of first-order details. Moreover, we'll show how our guidebook can actually be put to use to analyse and understand existing TEs.

4. Thomson's TEs

Judith Jarvis Thomson's article 'Killing, Letting Die, and the Trolley Problem' (1976) is concerned with the 'killing is worse than letting die' theory. Thomson finds this theory underlying debates on abortion, euthanasia, and the distribution of scarce medical resources. But it is not the only theory she tests. Throughout her paper, she is concerned with theories that answer the question: when is it morally permissible to intervene in a natural course of events and deflect a threat (or a good)?

Thomson's method for finding the right theory consists of TEs, *and only TEs*. In fact, Thomson presents at least 24 cases in her paper (involving trolleys, surgeons, avalanches, atomic bombs, and health-pebbles). Here is one example that most readers probably won't have heard of:[19]

> *Health-Pebble*. Suppose there are six men who are dying. Five are standing in one clump on the beach, one is standing further along. Floating in on the tide is a marvelous pebble, the Health-Pebble, I'll call it: it cures what ails you. The one needs for a cure the whole Health-Pebble; each of the five needs only a fifth of it. Now in fact that Health-Pebble is drifting towards the

[18] If you don't provide enough information, people might fill in any missing details in deviant ways (cf. Sosa 2009). One step to countering this problem is to present one's case not only in a concrete way, but also in a more abstract way where all the relevant factors are clearly identified (cf. Grundmann & Horvath 2014).

[19] Not to speak of all the variants of this case.

> one, so that if nothing is done to alter its course, the one will get it. We happen to be swimming nearby, and are in a position to deflect it towards the five. Is it permissible for us to do this? (1976: 209)

One may wonder: what's going on? If you're not a trained philosopher, and not used to TEs, you may see this case as an outlandish piece of fantasy that can tell us nothing about real life (such as about the distribution of scarce medical resources). Nothing could be further from the truth. The case is carefully devised, and all details are well chosen. In the following, we won't explain this for *all* of Thomson's cases. Rather, we'll confine ourselves to her main trolley cases: Passenger, Mayor, and Fat Man.[20] In each case, the question will be: *how is the TE devised?* On the basis of our guidebook, we'll be able to reconstruct Thomson's steps, and this will help students to appreciate TEs as counterexamples (rather than pieces of fantasy).

Let's start with the variant we're all familiar with:

> *Passenger*. Frank is a passenger on a trolley whose driver has just shouted that the trolley's brakes have failed, and who then died of the shock. On the track ahead are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Frank can turn the trolley onto it. Unfortunately there is one person on the right-hand track. Frank can turn the trolley, killing the one; or he can refrain from turning the trolley, letting the five die. (1976: 207)

How is Passenger devised? There are four steps. Step 1: according to Thomson's opponent, it's permissible to deflect a threat *only if* deflecting it doesn't involve killing people (call this view 'T1'). Step 2b: given that we're dealing with a necessary condition, one has to keep 'deflecting the threat is permissible' in the case, but remove 'deflecting it doesn't involve killing people.' Step 3b: add another factor that can plausibly be considered sufficient for 'deflecting the threat is permissible,' which is in this case the factor 'the threat's new target is smaller in number than other targets' (cf. Thomson 1976: 208). Step 4: add details about these factors, such as a trolley, a passenger, and a number of people on the two tracks.[21]

At this point, you might think that Thomson endorses the alternative view, according to which it is permissible to deflect a threat *if* its new target is smaller in number than other targets ('T2'). For in contrast to T1, this alternative view T2 doesn't

[20] In our view, *all* other cases, including Health-Pebble just cited, can be taken as analogs or variants of these three cases.

[21] Change 'threats' to 'goods' and you have an analysis of the Health-Pebble case.

predict the incorrect outcome, i.e. that Frank is not permitted to intervene and turn the trolley.[22] After all, the new target (one person on the side track) is smaller in number than the other target (five people in front of him). However, Thomson herself does not ultimately endorse T2, in light of the following TE:

> *Mayor*. The five on the track ahead are regular track workmen, repairing the track – they have been warned of the dangers of their job, and are paid specially high salaries to compensate. The right-hand track is a dead end, unused in ten years. The Mayor, representing the City, has set out picnic tables on it, and invited the convalescents at the nearby City Hospital to have lunch there, guaranteeing them safety from trolleys. The one on the right-hand track is a convalescent having his lunch there; it would never have occurred to him to have his lunch there but for the Mayor's invitation and guarantee of safety. (1976: 210)

Thus, the mayor has invited a person onto the side track, promising him absolute safety. Thomson contends that, in this scenario, we're not allowed to turn the runaway trolley, despite the fact that by turning it, we'd deflect it onto a smaller target. For, according to Thomson, the one has a better claim against the trolley.

How is Mayor devised? Step 1: according to the view Thomson wishes to challenge, it is permissible to deflect a threat *if* its new target is smaller in number than other targets. Step 2a: given that we're dealing with a sufficient condition (rather than a necessary one, as in the previous case), one has to keep 'the threat's new target is smaller in number than other targets' in the case, but remove 'it is permissible to deflect the threat'. Step 3a: remove another factor that can plausibly be considered necessary for 'it is permissible to deflect the threat', which is in this case the factor 'the threat's new target has no better claim against the threat' (cf. Thomson 1976: 209-10). Step 4: add details about these factors, such as the workmen, the mayor, and his promise.

Again, you might think that Thomson endorses the alternative view, according to which it is permissible to deflect a threat if (i) its new target is smaller in number than other targets, and (ii) the new target has no better claim against the threat ('T3'). For unlike T2, this alternative view T3 doesn't predict the incorrect outcome, i.e. that one is permitted to intervene and turn the trolley. After all, the new target has a better claim against the threat (due to the explicit promise by the mayor,

[22] Of course, not everyone agrees with Thomson's intuitive verdicts, here and below. Basically, if you disagree with Thomson, you're committed to denying, in our terms, the success of step 3 (a or b) of the guidebook.

and the high salaries of the workmen). But, again, Thomson herself does not ultimately endorse T3, in light of the following TE:

> *Fat Man*. George is on a footbridge over the trolley tracks. He knows trolleys, and can see that the one approaching the bridge is out of control. On the track behind the bridge there are five people; the banks are so steep that they will not be able to get off the track in time. George knows that the only way to stop an out-of-control trolley is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a fat man, also watching the trolley from the footbridge. George can shove the fat man onto the track in the path of the trolley, killing the fat man; or he can refrain from doing this, letting the five die. (1976: 207-8)[23]

This time, one has a choice between letting the five on the track die, or throwing and sacrificing an arbitrary bystander from the bridge (and save the five). According to T3, this is permissible: after all, one life is fewer than five, and none of the parties has a better claim against the trolley. According to Thomson, though, it's impermissible to throw the fat man from the bridge.

How is Fat Man devised? Step 1: according to the view Thomson wishes to challenge, it is permissible to deflect a threat *if* (i) its new target is smaller in number than other targets, and (ii) the new target has no better claim against the threat. Step 2a: given that we're again dealing with a sufficient condition, one has to keep conditions (i) and (ii) in the case, but remove 'it is permissible to deflect the threat'. Step 3a: remove another factor that can plausibly be considered necessary for 'it is permissible to deflect the threat', which is in this case the factor 'one can do something to the threat, rather than to a person' (cf. Thomson 1976: 215-6). This new factor distinguishes between *two* ways of deflecting threats. That is, one can deflect a threat by doing something to the threat (such as turning the trolley), or by doing something to a person (such as pushing the fat man), and the latter is considered impermissible. Step 4: add details about these factors, such as the footbridge and the fat man.

[23] The analogy with goods (rather than threats) can be found here: "Suppose that the one actually owns the Health-Pebble which is floating in on the tide. (It fell off his boat.) And I said that in that case, he has more claim on it than any of the five has, so that we may not deflect it away from him and towards the five. Let's suppose that deflecting isn't in question any more: the pebble has already floated in, and the one has it. Let's suppose he's already put it in his mouth. Or that he's already swallowed it. We certainly may not cut him open to get it out? Even if it's not yet digested, and can still be used to save five." (1976: 213)

The end-result of these TEs is the following account ('T4'):[24]

For all threats (or goods) X, it is permissible for one to deflect X if and only if
(i) X's new target is smaller in number than other targets;
(ii) X's new target has no better claim against x; and
(iii) one can do something to X, rather than to a person.

On this account, it's impermissible to throw the fat man from the bridge because doing so would involve doing something to a person, and this, Thomson assumes, is impermissible.

Thomson's overall strategy should be clear. She's interested in the link between moral permissibility and a range of potentially relevant factors, and to this end she isolates these factors in the imagination. The initial theory about moral permissibility that she finds in existing debates, i.e. T1, is repeatedly tweaked and retested, and in this way she builds up her theory until she arrives at T4. We should emphasize that T1-T4 cannot be found in this form and order in Thomson's paper, yet we believe they are faithful reconstructions of what she is after. T4 yields the correct verdict in all her three cases, as can be seen from the following table (where '+' stands for 'it's permissible to intervene', '–' for 'it's impermissible to intervene', and the grey boxes indicate the counterexamples).

TE	*T1*	*T2*	*T3*	*T4*
Passenger	–	+	+	+
Mayor	–	+	–	–
Fat Man	–	+	+	–

T1 has one counterexample: it predicts that it is impermissible to intervene in Passenger, while this is intuitively permissible. T2 has two counterexamples: it predicts that it is permissible to intervene in Mayor and Fat Man, while this is intuitively impermissible. T3 has one counterexample: it predicts that it is permissible

[24] This theory may seem complicated, though it is easy to work with on the basis of the following 'flowchart': Can you divert the threat by manipulating the threat itself? If No: do not intervene! If Yes: does one party have a better claim against the threat? If Yes: divert the threat from the party with the best claim! If No: maximize survivors!

to intervene in Fat Man, while this is intuitively impermissible. Finally, T4 has no counterexamples: it predicts the right result in all her cases.[25]

T4 is rather sophisticated, and an interesting result in itself. For it may be tempting to think that Passenger poses a challenge to deontological theories, while Fat Man poses a challenge to consequentialist views, and that these TEs demonstrate that the intuitions in one of these two cases must be abandoned. To say that one may intervene in Passenger, but not in Fat Man, would be 'inconsistent.' T4 shows that such an inference is too quick.[26] This theory is neither purely consequentialist nor purely deontological, but incorporates elements from both theories, and tells a consistent story about why one may intervene in Passenger, but not in Fat Man. Of course, more can be said about the factors 'having a better claim against the threat' and 'doing something to the threat rather than a person' (and more could be said in defence of them), yet doing so is not necessary for our purposes here. Our main goal has been only to illustrate our guidebook for analysing TEs.

5. Coda

We agree with Cappelen (2012: 131) that the details of Thomson's cases matter. Philosophers and non-philosophers alike often refer to trolley cases without having any idea about the function and details of these cases. According to Cappelen, furthermore, there aren't many things to say about TEs in general except that they are "devices for drawing our attention to philosophically interesting features of the world and for asking questions about those features" (2012: 132). Here, however, we disagree. We side with many others in claiming that, typically, TEs in philosophy function to expose counterexamples. The question was how such TEs are devised in order to fulfil this function, and in this paper we have presented a specific guidebook to gain such insight.[27]

[25] That is, according to Thomson. In later papers (1985, 2008), she considers a few further factors (e.g. the factor of whether or not one is prepared to deflect the threat onto oneself), yet it is not clear that these weaken her earlier conclusions (cf. FitzPatrick 2009).

[26] Also, to show that people are inconsistent, you'd have to show that people have different intuitions about two structurally similar cases, and *not* that people have different intuitions about two cases that comprise different factors (such as Passenger and Fat Man).

[27] This paper is written by Jan Willem; the material from section 4 derives from Matthijs' research. We thank Phil Robichaud, Naomi Kloosterboer, Rik Peels, the referees and editor of the journal, and several audiences for stimulating discussion and feedback.

References

Black, M. 1952. The Identity of Indiscernibles. *Mind* 61: 153-64.

Brennan, A. 2003. Necessary and Sufficient Conditions. In *Stanford Encyclopedia of Philosophy*.

Cappelen, H. 2012. *Philosophy without Intuitions*. OUP.

Clarke, M. 1963. Knowledge and Grounds. A Comment on Mr. Gettier's Paper. *Analysis* 24: 46-8.

Fischer, J. M. 1995. Stories. *Midwest Studies in Philosophy* 20: 1-14.

FitzPatrick, W. J. 2009. Thomson's Turnabout on the Trolley. *Analysis* 69: 636-43.

Foot, P. 1967. The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review* 5: 5-15.

Frankfurt, H. G. 1969. Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66: 829-39.

Gendler, T. S. 2000. *Thought Experiment. On the Powers and Limits of Imaginary Cases*. Garland.

Gettier, E. 1963. Is Justified True Belief Knowledge? *Analysis* 23: 121-3.

Gomes, G. 2009. Are Necessary and Sufficient Conditions Converse Relations? *Australasian Journal of Philosophy* 87: 375-87.

Grundmann, T. & J. Horvath 2014. Thought Experiments and the Problem of Deviant Realizations. *Philosophical Studies* 170: 525-33.

Häggqvist, S. 2009. A Model for Thought Experiments. *Canadian Journal of Philosophy* 39: 55-76.

Kripke, S. 1970/80. *Naming and Necessity*. HUP.

Nichols, S. & J. Knobe 2007. Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs* 41: 663-85.

Praëm, S. K. & A. Steglich-Petersen 2015. Philosophical Thought Experiments as Heuristics for Theory Discovery. *Synthese* 192: 2827-42.

Sorensen, R. A. 1992. *Thought Experiments*. OUP.

Sosa, E. 2009. A Defense of the Use of Intuitions in Philosophy. In D. Murphy & M. A. Bishop eds., *Stich and His Critics*, ch. 6. Wiley-Blackwell.

Thomson, J. J. 1976. Killing, Letting Die, and the Trolley Problem. *Monist* 59: 204-17.

— 1985. The Trolley Problem. *Yale Law Journal* 94: 1395-415.

— 2008. Turning the Trolley. *Philosophy and Public Affairs* 36: 359-74.

Williamson, T. 2007. *The Philosophy of Philosophy*. Blackwell.