Croatian Journal of Philosophy
Vol. XVII, No. 52, 2018

Moral Thought-Experiments, Intuitions, and Heuristics

FRIDERIK KLAMPFER
University of Maribor, Maribor, Slovenia

Philosophical thought-experimentation has a long and influential history. In recent years, however, both the traditionally secure place of the method of thought experimentation in philosophy and its presumed epistemic credentials have been increasingly and repeatedly questioned. In the paper, I join the choir of the discontents. I present and discuss two types of evidence that in my opinion undermine our close-to-blind trust in moral thought experiments and the intuitions that these elicit: the disappointing record of thought-experimentation in contemporary moral philosophy, and the more general considerations explaining why this failure is not accidental. The diagnosis is not optimistic. The past record of moral TEs is far from impressive. Most, if not all, moral TEs fail to corroborate their target moral hypotheses (provided one can determine what results they produced and what moral proposition these results were supposed to verify or falsify). Moral intuitions appear to be produced by moral heuristics which we have every reason to suspect will systematically misfire in typical moral TEs. Rather than keep relying on moral TEs, we should therefore begin to explore other, more sound alternatives to thought-experimentation in moral philosophy.

Keywords: Thought-experiments, moral intuitions, evidence, the Ticking Bomb, moral heuristics.

0. *Introduction*

Philosophical thought-experimentation has a long and influential history. While philosophers may not wear this as a badge of honour, as far as public opinion goes, thought-experiments (TEs for short) are a trade mark, or one of the trade marks, of philosophy. The proper place of the method of thought experimentation in philosophy and its epistemic credentials are more controversial, however. TEs appear to abound in epistemology, philosophy of mind and language, and metaphysics, and

they are certainly no less popular in moral and political philosophy as well as in philosophy of arts.

In the last two decades, however, philosophical thought experimentation has increasingly come under fire. Some of the discontent with the method was motivated by a growing metaphilosophical scepticism regarding the traditional (self-)conception of philosophy as an apriori, armchair intellectual activity. The other stemmed from the insights of empirical sciences studying psychological processes that underlie ordinary moral judgment, which seem to suggest, in effect, if not in intention, that our trust in TE-generated epistemic, modal, metaphysical and moral intuitions is unwarranted. In the paper, I will present and discuss two types of evidence that in my opinion undermine such blind trust in moral thought-experiments and the moral intuitions that these elicit: the discouraging record of thought-experimentation in contemporary moral philosophy, and the more general considerations explaining why this failure is not accidental.

Here is a sketch of the paper. In chapter one, I explicate what I mean by 'thought-experiment(ation)' and try to delineate the use of thought-experiments for the purpose of gathering evidence and/or providing justification for tested moral propositions (particular and general judgments, norms and principles, and theories) from other, less problematic uses of hypothetical reasoning in moral philosophy. In chapter two, I show the limitations of the TE-method by way of discussing a well-known moral thought experiment, the so-called Ticking Bomb scenario. I then proceed to arguing, in chapter three, that the limitations of the method as revealed in this particular moral TE are due neither to its poor experimental design nor to its misapplication, but are built into the method itself. In chapter four, I provide a rather sketchy account of psychological mechanisms that typically underlie the production of TE-generated intuitions and argue that we can best understand both the strengths and the weaknesses of this method by construing those intuitions as outcomes, or deliverances, of (generally social or specifically moral) heuristics. In the concluding chapter, I show what room is still left for the use of hypothetical examples and counterfactual reasoning in moral philosophy once we've given them up as sources of justification.

1. *Hypothetical reasoning and thought experimentation*

Hypothetical reasoning is ubiquitous and indispensable in moral philosophy. Regularly, and without much thought, we use it for moral guidance, judgment or as a helpful heuristic. So in evaluating our own and other people's decisions and/or actions we ask questions such as: "What if everyone did that?", "Would I want to see X done to me if I were at the other, receiving end of the action?" (the Golden Rule), "Can

I conceive, or will, without contradiction a world in which everyone acted on the given maxim, i.e. a world in which this maxim became a universal law?" (the Universal Law version of Kant's Categorical Imperative), "Would A have consented to X, had she been competent to judge?" (the substitute-judgment test for (proxy) consent or authentic will), and many more. Some or other form of idealization, i.e. counterfactual thinking, is also at work in various non-reductive accounts of normative properties: from the Whole-Life-Satisfaction theory of happiness, Full-Information accounts of the good, Desire-Based accounts of (normative, or justifying) reasons for action, Ideal Observer theories of right action, accounts of personal value or good, hypothetical consent-based accounts of legitimate political authority, to justice-as-fairness and contractualist accounts of right and wrong.

Whether these non-reductive accounts of various normative properties are correct or not, they serve as a helpful reminder of how heavily we rely on hypothetical reasoning as either a definitional tool or an instrument of discovery with respect to a whole range of normative properties. In this paper, I'm not suggesting we should abandon counterfactual reasoning in moral philosophy as utterly useless. Neither is my aim to launch a frontal attack on intuitions as such. My specific target is what I will call 'TE-evidentialism', i.e. a popular view that treats TE-generated moral intuitions as (at least prima facie) reliable pieces of evidence for or against moral propositions, i.e. accords them at least some (initial, even though defeasible) credibility, justifiability, epistemic value, and the like.

But first, some preliminary clarifications. What makes an exercise in imagination a thought-experiment, what sets it apart from other occurrences of hypothetical reasoning in (moral) philosophy? In order for a piece of imaginative, or counterfactual, thinking to qualify as a moral TE, we need to engage in it for a specific reason—namely to test a moral hypothesis that cannot be reliably tested in any other way. Or, as Tamar Gendler elegantly put it: "To perform a thought experiment is to reason about an imaginary scenario *with the aim of confirming or disconfirming some hypothesis or theory*" (Gendler 2007; my emphasis).

The idea, then, of experiments conducted in pure thought, is simple.[1] A controversial philosophical, or, in our case, moral proposition needs to be put to the test; so why not construct a thought-experiment, i.e. describe some hypothetical situation (kids pouring gasoline over a cat and setting it on fire; the world being populated by twice as many people as in the actual world but with lives barely worth living; having your brain removed and transplanted into someone else's body; seeing/experiencing colours for the first time; being lied to by someone you trust; not having, in your conceptual repertoire, the concept of a right; seeing, on your way to work, a kid drowning in a pond; finding a magical ring that renders you invisible and, by extension, grants you

[1] Deceptively so, as we'll see later.

impunity, and so on), ask people to think, and form a judgment, about it (would it be permissible, right, morally good, or better than some alternative, just, legitimate, and so on) and, finally, collect the 'raw data', the spontaneous, intuitive judgments elicited in them by that thought-experiment and see if they confirm or disconfirm the original hypothesis.

When does a moral judgment formed in response to such a hypothetical scenario qualify as intuitive? Here, again, I'm simply going to follow the tradition.[2] Intuitive moral judgments are characterized by their (i) distinct genealogy; (ii) characteristic phenomenology; (iii) modality; and (iv) epistemic status. Let me briefly elaborate: moral intuitions (i) spring into one's mind effortlessly; even when formed after careful observation, consideration, contemplation, or thinking about the subject matter at hand, they are not consciously inferred from other beliefs or believed propositions as their justifying grounds; (ii) they strike us as vivid, clear, inescapable, forced upon us; (iii) they present things as being necessarily the way they appear before our mind; and, finally, (iv) they strike us as self-evident, beyond doubt, as inconceivably at odds with moral reality, or truth.[3]

2. *TEs in moral philosophy*

On the standard view, philosophical TEs are used to access the non-empirical, i.e. abstract, normative and/or modal realm. More specifically, moral TEs are seen as the window into the moral realm. Here are some typical questions that moral philosophers aim to answer by means of moral TEs: Is it ever permissible to lie? May we kill, or torture, one to save five? Is it ever permissible to go to war? Can you do wrong blamelessly? Is harming always worse than merely allowing harm? Should we punish the most heinous crimes by death? What is just(ice) and how is it related to equality? When, if ever, is the rule of some people over others legitimate? What form of government is morally best? Is political violence, i.e. violence in the service of political goals, ever permissible? Can you be morally obliged to do that which you cannot possibly do? Can you be blameworthy for that which you only did out of ignorance and/or with no evil intention?

Having earlier delineated TEs from other (perfectly legitimate) forms and uses of hypothetical reasoning in moral philosophy which, however, don't qualify as moral TEs, since we don't engage in it with the aim of confirming or disconfirming some moral hypothesis, there are still plenty examples left that meet the above criteria. Below is a

[2] See, for instance, Miščević (2004) and Cappelen (2012).

[3] Of the aforementioned defining features, I consider the one that Herman Cappelen calls epistemic 'Rock status' most important one—for a judgment, or a belief, or a mere inclination to believe, to count as intuitive, it need not be seen as indefeasible, but it should at least be treated—in effect, if not in thought—as fairly evidence-recalcitrant.

random selection of such hypothetical scenarios and corresponding hypotheses that the former are designed to confirm or disconfirm:

(i) The Ring of Gyges → no one would act justly, if everyone were in possession of a magic ring that granted them absolute impunity. (Morality/justice is rightly appreciated merely for its positive consequences, i.e. instrumentally, but not (primarily, or also) for its own sake, i.e. intrinsically.) (Plato 1993)

(ii) The Ticking Bomb → torture is not absolutely prohibited (McMahan 2008a and 2008b)

(iii) Feinberg's Nowheresville → rights are necessary for self- and other-respect, as well as our sense of human dignity (Feinberg 1970)

(iv) Singer's Pond → assistance to the poor and destitute is morally obligatory, not just morally commendable (Singer 1993)

(v) Singer's Shelter/Fairhaven → hermetically closed borders and restrictive laws on (im)migration cannot be morally justified (Singer 1993)

(vi) Feinberg's 31 variations on the Ride on the Bus story → the offence principle (there are (crudely six types of) human experiences that don't constitute harm, yet are so unpleasant that we can rightly demand legal protection from them even at the cost of other persons' liberty (Feinberg 1985)

(vii) Nozick's Experience Machine → pleasure is not the only kind of thing that is valuable in and of itself, irrespective of its consequences, and everything else of value in our lives is not valuable only insofar as, and to the extent that, it promotes pleasure (Nozick 1974)

(viii) Thomson's Violinist → the right to life does not entail the right to a non-consensual use of someone else's body for one's own survival (Thomson 1971)

(ix) Rachels' Smith and Jones → killing is not intrinsically morally worse than letting die (Rachels 1975)

The above list is far from exhaustive, of course. Still, given the frequency and relative popularity of the method, the results of thought experimentation in moral philosophy are discouraging, to say the least. Hardly any controversial issue in moral philosophy (I'd even risk to say 'none') has been settled, or brought a bit closer to resolution, by means of moral thought experimentation, however ingenious. How come? My aim in this paper is to offer a preliminary, still rather crude diagnosis of this failure.

3. *mTE-evidentialism*

But let me first clarify the scope of my argument in order to prevent potential misunderstandings. As already said, the main target of this paper is not counterfactual thinking or reasoning as such, but rather the view that for want of a better name I will call mTE-evidentialism:

Intuitive moral judgments formed in response to moral TEs, provide some initial, prima facie credible evidence for or against moral propositions (particular and general moral judgments, principles, norms, distinctions and theories)[4]

A brief clarification of why I chose this particular formulation is due before we can proceed to critical evaluation. First, the view that I'd like to criticize is formulated in terms of evidence, not justification. I take evidence, in contrast to justification, to be if not itself a primitive notion, then at least one that can be fairly simply explicated in terms of reasons for believing—E provides evidence for mp (i.e. certain moral proposition), if, as a consequence of me coming to know or believe about E I now have a prima facie reason to believe that mp. According to this (admittedly, simplified) account, when someone treats an intuition elicited by a typical moral TE as evidence for or against a certain moral proposition, he or she is committed to the view, at a minimum, that the fact that we intuit, i.e. spontaneously judge an (fictional) agent's particular (fictional) decision and/or action in a given (once again fictional) situation as right or wrong, provide us with some reason for believing that this very decision and/or action (as well as all those that share all the morally relevant features with it) is indeed such, a reason that was not available to us before we engaged in judgment, or contemplation, of this hypothetical, fictional situation.

Secondly, what I try to advance here is an argument for scepticism about the evidential value or role of, in particular, moral TEs, not philosophical TEs in general. I want to suspend, as far as I can, my judgment on thought-experimentation in other areas of philosophy, such as metaphysics, epistemology, philosophy of mind, philosophy of language. It does seem to me that fairly little progress has been made

[4] The kind of view that I have in mind with 'mTE-evidentialism' is nicely laid out in the following paragraph by one of its most outspoken advocates, Jeff McMahan: "Suppose that one is curious about whether a certain factor is morally significant in a certain specific way—for example, whether the intention with which a person acts can affect the permissibility of her action. It may happen that reflection on intention in the abstract proves inconclusive. One might then devise a pair of hypothetical examples in each of which an agent goes through the same series of physical movements and in which consequences of those movements are identical. The *only* difference is that in one case the consequences are intended as a means whereas in the other they are unintended but foreseen side effects. Suppose that a large majority of people from a variety of cultures judge that the agent who intends the bad consequences acts impermissibly while the agent who merely foresees them acts permissibly. *That is at least prima facie evidence for the view that an agent's intentions can affect the permissibility of her action.* Yet if one had sought to elicit people's intuitions about a pair of actual historical examples, it would have been inevitable that people would have been influenced by irrelevant historical associations, distracted by irrelevant details, or guided in their evaluations by morally relevant differences between the two cases having nothing to do with the agents' intentions. The value of hypothetical examples is that they can exclude all such features that are irrelevant to the purpose of the example." (McMahan 2008b, my emphasis)

thanks to Gettier- or Frankfurt- or Lehrer- or Chalmers-types of examples in those areas of philosophical inquiry as well. Nevertheless, I'd like to limit my conclusions to the alleged evidential role of moral thought experiments alone, if for no other reason than to avoid inviting further, unnecessarily provoked criticism.

Thirdly, my critique is primarily directed against a small subset of moral intuitions, namely those generated by moral TEs, not against moral intuitions as such. Personally, I find claims about appeals to moral intuitions being constitutive of any moral inquiry, grossly exaggerated. No doubt, there is a rich and lively tradition of moral philosophizing that makes appeals to what we clearly intuit about this or that described moral setup central to moral inquiry (McMahan 2002, Kamm 2008, Parfit 1984 and Unger 1995 naturally spring to mind). That said, however, many books in moral philosophy (certainly the three moral philosophy classics, Aristotle's Nicomahean Ethics, Kant's Groundwork and Mill's Utilitarianism) make little or no use of moral TEs or even explicitly refuse to credit moral intuition with any evidential import. Opinions on whether appeals to intuitions are central or marginal to the practice of contemporary analytic philosophy are divided. (For three antagonistic views, see Cappelen 2011, Weatherson 2014 and Deutsch 2015) But even if most appeals to intuitions in philosophical literature are merely colloquial and thus not really indicative of deep methodological commitments, it is hard to deny both the existence and the influence of a vocal tradition in contemporary moral philosophy which makes the so-called method of cases central to moral inquiry and is insofar committed to taking the evidential value of our (in fact, mostly author's own) intuitions at face value.[5]

Finally, I tried to make mTE-evidentialism as undemanding as possible. No one really holds that TE-generated moral intuitions can establish the truth or falsity of any moral proposition *on their own*. (Well, at least declaratively they don't, the existing philosophical practice is a different story.) To claim otherwise (as Deutsch 2015 occasionally does) is to build a straw man. Still, many philosophers seem to treat TE-generated moral intuitions as an independent source of at least some, prima facie and defeasible evidence for the truth or falsity of moral propositions under consideration. In this paper, I want to deny them even that much epistemic significance.

Let me express my principled worry, then. When we try to solve some moral quandary by means of a moral TE, we are invited first to contemplate and then to judge some poorly described hypothetical situation. But why acknowledge pretty much any answer to the question

[5] Whether practiced frequently or not, as Kuntz and Kuntz (2011) show, there is a fairly strong support, among professional philosophers, for the justificatory or evidential role of appeals to intuitions. With the following proviso: most of them find intuitions useful but not also essential to the justification process; and they typically assign a more important role to intuitions in the process of the discovery of philosophical theories than for the purpose of their justification.

"Imagine/consider such and such a situation? Would it instantiate such and such moral property or not?" as epistemicaly authoritative and truth-conducive? Why treat our swift, spontaneous, automatic moral judgments, whether particular or general, instant or delayed, as revealing anything else but how *our mind* works; how *we feel and think about the world*? Psychologically, we find transitions from 'A's ϕ-ing in C appears wrong to me, strikes me as such' to 'ϕ-ing is sometimes/often/always wrong' fairly easy and natural to make, but what, if anything, warrants them? What are the epistemicaly relevant features of TE-generated moral intuitions? Admittedly, they share most of their phenomenal properties with other TE-generated philosophical intuitions, but do they so clearly share their putative epistemic credentials as well?[6]

Let me strengthen the above challenge with another analogy. When in opinion polls we ask people "Do you think the use of torture against suspected terrorists in order to gain important information can often be justified, sometimes be justified, rarely be justified, or never be justified?", i.e. about the (im)permissibility of torturing a terrorist in what is basically a Ticking Bomb type of scenario, we treat their replies *as evidencing their subjective opinion* on this contentious moral issue; when, on the other hand, we ask them to form a moral judgment in response to a Ticking Bomb thought experiment with exactly the same informational content, we are expected to treat their judgments as *a prima facie evidence for the moral truth about torture*. The proponents of moral thought experimentation need to provide an explanation for what, if anything, warrants such different treatment.

[6] I'd also like to remain agnostic on the issue of epistemic credentials of intuitions about more general moral principles, since these will typically avoid some of the pitfalls of, or won't necessarily display the same shortcomings as, our intuitions about particular cases described in moral TEs. So, as far as I am concerned, the following may be instances of prima facie credible intuitions: that harming is worse than merely allowing harm which, in turn, is worse than failing to benefit; that in order for something to be better or worse, it must be better or worse for someone; that we ought to do that which will make the world a better place; that, other things being equal, promises ought to be kept; that killing civilians is worse than killing soldiers; that killing a (human) person is normally more seriously wrong than killing a (non-human) animal (the infamous speciesist intuition); that adding new person to the world is morally neutral, and the like. Perhaps there is such an epistemicaly noble thing as 'rational intuition' after all and professional philosophers are particularly apt in using this special faculty to access the realm of noble philosophical truths. I don't have much patience with any sort of intuitionism, but since this is no place for opening up the Pandora box of intuitionism debate, what I would simply deny in this case, then, is that philosophers actually make any use of this formidable faculty when, as part of their arguments for or against contentious moral propositions, they advance moral TEs and make appeals to intuitions thereby elicited. For a more systematic and detailed attack on the idea of a rational (philosophical) intuition and its alleged epistemic credentials, see Mizrahi 2014.

4. *The Ticking Bomb*

Let me illustrate the limitations of the case method, or thought experimentation in moral philosophy, by way of a well-known example, the so-called Ticking Bomb scenario. In fact, there is no one Ticking Bomb scenario, but many.[7] Hence, I will take the following description as paradigmatic of this particular kind of moral TE:

A terrorist has planted a nuclear bomb in New York City. It will go off in a couple of hours. A million people will die. Secret agents capture the terrorist. He knows where it is. He's not talking. But they can break his silence by torturing him. In fact, torture is the only way to extract the information about the location of the bomb from him in time to successfully deactivate the bomb and save those million innocent lives. Given that, would it be morally permissible for the agents to torture the terrorist?

Now, the Ticking Bomb scenario (or TBS, for short) has been subjected to a lot of fierce criticism since its inception, probably more than any other philosophical thought experiment with the due exception of Trolley cases. David Luban gives voice to most common concerns when he writes:

> The first thing to notice about the TBS is that it rests on a large number of assumptions, each of which is somewhat improbable, and which taken together are vanishingly unlikely. It assumes that an attack is about to take place, and that 'the authorities' somehow know this; that the attack is imminent; that it will kill a large number of innocent people; that the authorities have captured a perpetrator of the attack who knows where the time-bomb is planted; that the authorities know that they have the right man, and know that he knows; that means other than torture will not suffice to make him talk; that torture will make him talk—he will be unable to resist or mislead long enough for the attack to succeed, even though it is mere hours away; that alternative sources of information are unavailable; that no other means (such as evacuation) will work to save lives; that the sole motive for the torture is intelligence-gathering (as opposed to revenge, punishment, extracting confessions, or the sheer victor's pleasure in torturing the defeated enemy); and that the torture is an exceptional expedient rather than a routinized practice. Some of these assumptions can be dropped or modified, of course. But in its pure form, the TBS assumes them all. That makes the TBS highly unlikely. (Luban 2008)

Hence, as the first objection goes, a typical TBS rests on a number of improbable assumptions which combined render it highly unlikely that anyone would ever have to face such an agonizing choice. How damaging is this objection? It is certainly a legitimate worry, for it shows the TBS to be practically useless for moral guidance in those more realistic,

[7] The Ticking Bomb scenario seems to have made its inaugural appearance in Michael Walzer's seminal article "Political action: the problem of dirty hands". In it, Walzer describes "a political leader who is asked to authorize the torture of a captured rebel leader who knows or probably knows the location of a number of bombs hidden in apartment buildings around the city, set to go off" (Walzer 1973).

everyday contexts that have (re)ignited the moral debate on torture after 7/11 attacks in the first place. Admittedly, low likelihood is not the same as impossibility—for all we know, such circumstances could occur, however miniscule their likelihood, and when they did, the Ticking Bomb thought experiment appears to suggests, agents would be morally permitted or even obliged to resort to torture. But what good is this true insight, if it is one at all, if either these conditions will never apply or even when they do, we won't be able to tell that anyway? So even on the assumption that we all (or a fair majority of us) clearly intuit that torturing the terrorist in order to prevent the massive loss of innocent people's lives is permissible under described circumstances,[8] this would only justify torture in those extremely rare circumstances where the terrorist's guilt/liability is established with hundred-percent certainty and torture cannot possibly fail to work. Practically never, then.

The unrealistic epistemic assumptions are only part of the problem with TBSs. What other critics found equally problematic is their lack of wider social context. For torture to work, but not kill the terrorist in TBS, it would have to be applied competently and with highest precision. But such know-how is not simply given, it must be learned. Effective, yet not life-threatening torture thus requires expert torturers, which in turn presuppose systematic training in torture. So the ultimate price of having a secret agent competent enough in torture to extract the life-saving information from the terrorist in a TBS without rendering him unconscious or even killing him, is the institutionalization and, inevitably, normalization of torture. By being silent on this and other morally relevant conditions for effective defensive or preventive torture, TBSs fail to give proper weight to real moral costs involved in rescuing a million.

The list of objections to TBS is hereby not exhausted. Many authors, for example, use TBS as a building stone in their moral case for the legalization of torture. Suppose, then, for the sake of the argument that the TBS (or, more precisely, people's overwhelming moral approval of the use of torture under those circumstances) does manage to provide some new evidence that could tip the evidential balance in the initial dispute over whether torture is absolutely morally prohibited, i.e. morally wrong without exception, or not. Even on this fairly generous assumption, however, it would be pretty naive to expect the TBS to validate further inferences about the proper legal status of torture. In other words, the fact that the secret agents' torturing of the terrorist in the TBS wins our intuitive moral approval, whether it provides us with some reason for believing that, indeed, torture sometimes is morally permissible or not, does not constitute a reason, however weak this reason may be, for a further belief that torture ought to be legalized. So those who do treat it as a piece of evidence for the latter, more

[8] Which, given the results of the opinion polls, we have strong reasons to doubt. More on that later.

ambitious, but also more controversial claim, are simply overstating its logical implications. We can add, then, to TBS's so-far recorded sins, namely practical irrelevance and normative misrepresentation, the third one, misapplication.

Given the unpopularity of TBS and the multitude of objections raised against it, a proponent of moral thought experimentation might at this point protest that its limitations are in no way indicative of, or representative for, moral thought experimenting as such. I'd like to insist, however, that there is nothing special about this particular type of moral TE, meaning that there are no features of its design or implementation that are both (a) unique and (b) such that they clearly disqualify it as a test of moral propositions. In this, I concur with the following observation by Jeff McMahan:

> When one understands what hypothetical examples are designed to do (namely filter out irrelevant details that can distract or confuse our intuitions, thereby allowing us to focus on precisely those considerations that we wish to test for moral significance, op. FK), one can see that the ticking bomb case is an entirely respectable philosophical tool. It is relevantly similar to thousands of other hypothetical examples that have appeared in the work of moral philosophers in recent decades and that most philosophers regard as legitimate components of philosophical arguments. It has no features that are not characteristic of the majority of hypothetical examples in moral philosophy. It is no different in relevant respects from the familiar trolley cases, transplant cases, examples comparing and contrasting terror bombers and tactical bombers, and so on. It is, if anything, more realistic than most. (McMahan 2008b: 3)

I agree. There is nothing peculiar about TBSs, at least nothing that would a priori disqualify them as, to quote McMahan, 'respectable philosophical tools'. Provided, of course, that you consider moral TEs 'respectable philosophical tools' (which I don't). The choice situation may be less likely to occur in the real world than those described in other, less disputed moral TEs, those who appeal to them as a way of justifying torture may not be entirely honest about what it takes for those options to be truly viable, and sometimes people overstate their evidential potential, but let's face it, it is a typical moral TE. The problem with TBSs does not lie in the details of its design or their misapplication—even though the design is often flawed and the TE misapplied—, it is more fundamental and as such shared by (most) other moral TEs.[9] It resides, above all, in the unquestioned transition from appearance to reality, from moral feeling and emotion to its (corresponding) object, but also in its debilitating under-description and impoverished context. And that's why no amount of redesigning the initial setting in order to

[9] All but one, to be fair: since TBS is typically advanced as a counter-example to a universal moral claim ("Torture is never morally permitted."), it lacks the generalization stage characteristic of many famous moral TEs. Given that generalizations in TEs are even less justified than initial particular intuitive judgments, TBS turns out to be, somewhat paradoxically and at least in this one respect, less problematic than most moral TEs.

make it more socially, epistemologically and psychologically realistic, will help.[10] All it might do instead is undermine whatever little initial moral consensus there was about it.[11]

5. *General scepticism about moral TEs*

Showing an instance of a moral TE flawed is not the same as discrediting the method of moral thought experimenting as such, of course.[12] In what follows, I will present and briefly discuss some more general considerations that should, when properly acknowledged, significantly reduce our level of confidence in the capacity of moral TEs—and the moral intuitions thereby generated—to resolve substantive moral disputes, or, at a minimum, (dis)confirm competing moral hypotheses.[13] These include, but are not limited to, the following: (i) unresolved disputes over experimental design, (ii) indeterminate outcomes of moral TEs, (iii) confusion over the correct level of generality, (iv) mistaken moral arithmetic, (v) vicious circularity, (vi) sensitivity, or responsiveness, to morally irrelevant features (framing effects, order of presentation,…), (vii) reliance on dubious moral heuristics, and, last but not

[10] See Walsh (2011) for an interesting, but eventually failed, attempt to provide a set of reasonable criteria for a legitimate use of TEs in moral inquiry.

[11] This comes to surface in McMahan's own clever redesigning of the original TBS where instead of agents torturing the terrorist in order to prevent nuclear explosion and the resulting death of one million innocent people, we are asked to imagine agents torturing the same terrorist in order to prevent his accomplice from torturing an innocent hostage at some hidden location. While this scenario is no doubt better suited for the job of determining what valid moral consideration or principle could possibly justify torture in the paradigmatic TBS, the lesser evil or the preventative justice, it would be unreasonable to expect the 'Is it permissible to torture one culpable person to prevent the torture of one innocent person?' to generate the same degree of agreement as the 'Is it allowed to torture one culpable person to prevent the violent deaths of one million of innocent persons'. McMahan need not be bothered by this prospect, of course, since he only ever consults his own intuitions about his ingenious TEs anyway. Frances Kamm is another famous advocate and practitioner of the TE method in moral philosophy who never seem to have any doubts about her own TE-generated intuitions, however at odds they might be with everyone else's.

[12] In Klampfer (2017), I argued for the evidential irrelevance, or impotence, of Feinberg's 31 variants of the Ride on the Bus stories and in its longer, unpublished version I made a similar point about Plato's famous Ring of Gyges thought experiment.

[13] What level of confidence in the TE-generated moral intuitions will be reasonable to preserve after said adjustment? Not enough, in my opinion, to justify their further use, as long as at least some viable alternatives are available. Some authors (for instance, Liao et al 2012) believe the evidence of unreliability supports a more qualified form of scepticism—if it has been demonstrated of some moral TE that people's intuitive responses to that TE can be influenced by manipulating what we all agree are morally irrelevant features of the experimental situation, then—and only then—can this particular moral TE no longer be used as a source of evidence for or against any moral proposition. Everything else we are free to use, until and unless it is similarly discredited.

least, (viii) mostly undetected and uncorrected (even incorrigible) effects of bias and prejudice.

Our moral intuitions, a growing body of research seems to suggest, are quick, snap, unreflective, spontaneous, almost automatic judgments; they are influenced by mood, affection, emotion, fatigue, and as such easily swayed one way or the other by simple rephrasing of the story, a change in the order of presentation, emotional and social priming, or simply by tampering with our physiological needs; they escape conscious control and seem to rely, for their formation, on similar cognitive shortcuts, heuristics, that we use in our judgments in other domains (such as availability and representativeness); and yet, despite their contingent origin and shape, they are mostly dogmatic, i.e. resistant to contrary evidence; when our intuitive judgments are challenged or questioned, we are seldom able to provide good reasons or compelling evidence in their support (or if we are, the reasons we adduce are often not those that were operative in the production of our judgment); even more, we fail to see any need for that and, consequently, don't consider this to be a problem (what is called 'moral dumbfounding'). The most recent psychological research suggests that even professional philosophers' moral intuitions are not immune to systematic and distorting effects of framing, ordering, prejudice, affect and bias. (Schwitzgebel and Cushman 2015, Liao et al 2012) The upshot: our intuitive responses to moral TEs, however carefully we may design the latter, will always track a host of morally irrelevant features of the hypothetical situation (such as novelty, excitement, disgust, surprise or arbitrary convention) and will hence serve as rather poor guides to moral truths.

These and similar shortcomings of TE-generated moral intuitions have been observed over and over again and are fairly well-documented by now. In what follows, I want to focus on (ii), (iii) and (vii) instead, since even though these problems with moral TEs are no less serious than the shortcoming of moral intuitions listed above, they tend to be both overlooked by the critics and underestimated by the advocates of moral thought experimenting.

5.1. *What evidence?*

Ideally, an experiment, whether conducted in a lab or in one's mind, would yield results that, whether quantifiable or not, measurable or not, are unequivocal. Most moral TEs fall embarrassingly short of this ideal, however.[14] It is no surprise that the more controversial and divisive some moral issue, the more widely distributed along a spectre intuitive moral judgments will be that the supposedly crucial moral TE elicits. The size of disagreement can be somewhat reduced by turning away from what looks like a fairly random distribution in the responses

[14] Jeff McMahan clearly underestimates the depth of intuitive disagreements or else he wouldn't have assumed that "large majority of people from a variety of cultures" will often converge in their judgments about particular moral TEs.

of lay people and considering only the more ordered 'considered moral judgments' of professional philosophers instead, but even the latter are seldom homogenous enough to admit of a unanimous verdict.

Let me illustrate this by way of what is probably the best known, and by far the most overexploited, moral TE, the Standard Trolley case. In the path of a runaway trolley car are five people who will definitely be killed unless you, a bystander, flip a switch which will divert it on to another track, where it will kill one person. In a huge BBC online survey, 77 percent of the total 65.000 respondents answered the question of whether they would flip the switch with 'yes' and 23 percent with 'no' (Sokol 2006). We can make the distribution of answers to the above question more uneven by turning to professional philosophers, but the prospects of getting anywhere near a unanimous decision will nevertheless remain bleak. A survey of 1,972 contemporary philosophers, conducted via PhilPapers (Bourget and Chalmers 2014), brought the following results: 68.2% 'yes, flip the switch' votes, 7.6% 'no, don't flip the switch' votes and the remaining 24.2% either agnostic or undecided or something else.[15] So while over two thirds of philosophers agree that it is permissible (or even obligatory) to flip the switch in the Standard Trolley case and only a tiny minority departs from that, still more than one in four philosophers refuse to share the predominant intuition. Has the Trolley moral TE delivered a clear result in this case, then, or failed to do so? And if the latter, what ratio of 'yes' to 'no' answers would be enough to validate such an affirmative answer?[16]

No similar data has been so far collected on the Ticking Bomb scenario(s), so we can only guess how much agreement in moral judgment it would generate among lay people and how those numbers would compare to the judgments of professional philosophers. What is available, however, is some relevant statistical data gathered over the years in many nation-wide opinion polls in the USA. And these leave a lot to be desired. A 2005 public opinion poll, for instance, asked, "Do you think the use of torture against suspected terrorists in order to gain important information can often be justified, sometimes be justified, rarely be justified, or never be justified?" Forty-six percent of Americans surveyed answered 'often' or 'sometimes', but 32%, on the other hand, answered 'never'. Another poll from June 2006 found 36% of Americans agreeing that "Terrorists now pose such an extreme

[15] I've lumped all other categories under 'other' to arrive at this figure. In the original questionnaire, the rest of the options are fairly diverse, ranging from 'agnostic' over 'not familiar enough' to 'unclear question'. Some of those that not many, but still some, respondents have chosen, such as 'accept both', 'reject both', 'intermediate', 'find another alternative', may raise doubts about the benefits of philosophical training.

[16] The more complicated the variations on the default thought experiment get (Fat man or Bridge, Loophole, and so on), the faster we can expect the last group, the 'other' or the 'undecided', to grow/expand and, correspondingly, the initial wide agreement, if there was any, to quickly dissolve.

threat that governments should now be allowed to use some degree of torture if it may gain information that saves innocent lives." (Luban 2008: 3) Given the history of heated disputes over the legitimacy of the use of Ticking Bomb scenarios in the moral debates on torture, there is little hope that the judgments of professional philosophers on this very issue would display a significantly higher agreement rate than that.

Now one may want to object that the above requirement of homogeneity of the experimental results is too strong, since very few, if any, laboratory experiments or field trials yield outcomes that come anywhere near this ideal. Suppose you are investigating the efficiency of a new drug, call it Perosan, with respect to some chronic condition and so to do that you divide 20 patients diagnosed with this condition into two groups of ten people. Over the course of three months, those in the control group receive placebo, while those in the experimental group are given exactly the same dosage of Perosan. After three months, you measure and compare the most common symptoms along three dimensions: variety, duration and intensity. Now even if Perosan turns out to be an efficient drug, it would be close to a miracle if it had exactly the same measurable beneficial effect on everyone. What is more realistic to expect with respect to results is a certain degree of variation, with some people's condition improving more, other's less and still others perhaps showing no improvement at all. Overall, drug efficiency may be 20 percent, ranging from zero to forty. The researchers will then typically go on to investigate what factors could have facilitated the effects of the drug where it worked better and what other factors could have blocked them where it worked less well or not at all. It's usual business in science, so why insist that thought-experimental results must exhibit a much stricter uniformity?

Note, however, that this line of argumentation is not really available to the advocates of moral thought experimentation. Unlike lab experiments or field trials, the lack of uniformity in thought experimental results cannot be accounted for in terms of patterns of distribution characteristic of statistical rather than deterministic connections between two or more observed variables. Where people's intuitive moral judgments diverge, as they always do to some extent, we cannot simply convert the resulting variation into, say, degrees of confidence in a tested moral proposition, so that in the above Standard Trolley case, where 77-percent of respondents opted for the flip-the-switch option and 23-percent were opposed to it, the epistemicaly rational thing would be to either lower your level of confidence in the moral proposition 'flipping the switch is the morally right thing to do in those circumstances' (if prior to these results you had no doubts about that) or increase it (if prior to this vote you were fully convinced that you ought not intervene). Given that you clearly intuit the former to be the case (and necessarily so), your corresponding confidence level should be maximal. But then those 23-percent just as clearly intuit exactly the opposite, so unless you have good reasons to doubt their moral competence, maybe you should reduce

your confidence level to reflect that fact?[17] This, however, cannot really be done without questioning your moral intuitions' credential in this (and all the other) case(s) of conflicting intuitions.

5.2. *Evidence for what?*

Legitimate doubts about what counts as the single outcome of a moral thought experiment and when it is correct to say that the latter has actually delivered a clear-cut, unambiguous result are amplified by yet another quandary—what moral proposition or hypothesis was actually confirmed or disconfirmed by a particular moral TE?

The problem is that contested moral propositions can rarely, if ever, be put to test in pure thought directly. Consider James Rachels' (Rachels 1975) famous Smith and Jones TE, where the reader is invited to contemplate and morally evaluate the following two hypothetical scenarios: in the first, Smith, wanting to secure huge inheritance for himself, sneaks in the bathroom and drowns his young nephew in a bath; in the second, Jones, driven by the same motive, merely lets his nephew drown after the latter has hit his head against the edge of the bath and lost consciousness. The moral issue that Rachels is trying to resolve by means of this TE is rather different, however: "Is killing intrinsically worse than letting die?". And he takes our shared intuitions that Smith and Jones are equally culpable, or blameworthy, for their respective (in)actions (which, it needs to be said, is presumed rather than demonstrated) as evidence that at least in this one pair of cases letting someone die is just as bad, or wrong, as killing him. But surely equal culpability for X and Y respectively, even if it were unambiguously established by the responses of an overwhelming majority of people to this moral TE, does not by itself imply moral equivalence between X and Y—all it means is that people consider Smith and Jones both fully responsible for the wrongful harm (of premature death) that befell their nephew, and not that it doesn't matter, in their opinion, whether this harm was directly caused or merely not prevented.[18] The evidence that people's intuitions about moral TEs are meant to provide for or against moral propositions, can thus at best be indirect, and the link between the evidence provided by people's responses to a given moral TE and the tested claim is often established only retrospectively, via abductive reasoning—intuitive moral judgments elicited by any given moral TE are taken to provide evidence for the truth of that one among many candidate moral propositions which best explains their occurrence on this particular occasion. The problem is that this 'evidence',

[17] This does look like a textbook example of moral peer disagreement—not only should we treat each other as moral peers, given that basic moral competence is normally not considered something one needs to acquire through formal learning, my disagreeing counterpart and I use exactly the same source of justification, i.e. our own intuion, for the moral belief that we formed in response to the given moral TE.

[18] Levy (2004) offers a devastating critique of this 'the-one-difference-that-makes-all-the-difference, or none' approach.

even when sufficiently unambiguous not to raise the 'what-evidence?' question, will always be consistent with more than just one hypothesis, and often with several of them. And not just consistent with, but also equally well explained by, several of them, I'd like to add. So even on the assumption of phenomenal conservatism which takes moral appearances or seemings at their face value, as more or less veridical,[19] there will always be room for asking which particular moral proposition was confirmed or disconfirmed by people's intuitive responses to any given moral TE, however homogenous and unified these may be.

That this is a principled worry, another famous moral TE, Singer's Pond, nicely illustrates. You are on our way to work, and as you pass through the park, you see a small child drowning in the nearby pond. You can jump in the water and pull the child out, thereby ruining your expensive clothes and shoes, or you can proceed to work, minding your own business, and let the child drown. Hardly anyone finds the latter option morally justifiable, but what exactly is it that we clearly intuit with respect to the described situation: (a) that I ought to save the child drowning in front of me; (b) that, in general, everyone in a position to do so ought to save children from drowning; or (c), the option that Singer himself prefers, that one ought to prevent something bad from happening, as long as he or she can do so without sacrificing anything of comparable value? Whether we understand the role of the Pond TE as providing evidential support for the principle stated in (c), or merely as reminding the reader that he or she already tacitly subscribes to a version of this moral principle, one can fairly easily come up with a counter-example to the principle[20] and this will set the inquiry back to the beginning. All that we clearly intuit in Pond is that we ought to pull that particular drowning child out of that particular pond, since nobody else is around to help and we can rescue the child at an insignificant cost. Everything else is extrapolation and generalization beyond what is prima facie evident and consequently questionable.[21]

The problem of determining the exact scope of TE-generated moral evidence is epidemical. Recall the Ticking Bomb scenario and its relatively brief, yet tumultuous history. Originally, the TB scenario served as a remainder that political necessity may force leaders to violate the constraints of ordinary morality (say, by ordering the torture of a suspect rebel to extract the life-saving information about the location of a planted bomb). Later, it was redesigned to better serve the needs of a

[19] Phenomenal Conservatism is a theory in epistemology that seeks, roughly, to ground justified beliefs in the way things "appear" or "seem" to the subject who holds a belief. The intuitive idea is that it makes sense to assume that things are the way they seem, unless and until one has reasons for doubting this (Huemer 2013).

[20] As Peter Unger has done with another moral TE, called Envelope. See Unger 1995.

[21] This problem is often underestimated by friends of moral thought experimenting. See, for instance, rather casual remarks about the generalization stage in Plato's Ring of Gyges (and elsewhere) in Miščević (2013b).

newly sparked debate on the morality and/or legality of torturing terrorist suspects and many of its original features were either dropped or replaced for that reason (rebel became terrorist, bomb became nuclear device, political leader's choice was substituted by that of the secret agents' and epistemic uncertainty, implicit in the word 'suspect', was replaced by full confidence both about the terrorist's culpability/liability and the outcomes of alternative courses of action). Those who vigorously opposed appeals to Ticking Bomb scenarios in recent heated debates on morality and/or legality of torture, mostly understand them to show, if successful, that torture ought to be legalized and/or institutionalized. Jeff McMahan, on the other hand, emphatically denies such an implication. What he believes the Ticking Bomb in its role as a moral TE convincingly shows is that torture cannot be absolutely wrong (and obviously so). This clear moral insight, he insists, has no direct implications for a related, but separate morally issue, how we ought to regulate torture by legal and political means. But even if one accepts his arguments that the proper place of the Ticking Bomb thought experiment is within debates on morality, not legality, of torture, it is still surprising and somewhat inexplicable that so many philosophers could have been so mistaken about its proper place and scope. Furthermore, things become even more complicated when we try to specify what exact moral proposition this particular moral TE is meant to test—what prima facie justification for torture does it provide, if any—and, consequently, what types of torture does it legitimize, a necessity or lesser-evil one or a liability-based one? Unless and until we can answer this question—and it takes McMahan himself pages of sophisticated reasoning to accomplish this goal—we don't know what TB-generated moral intuitions are supposed to establish, the moral permissibility of consequential (i.e. overall beneficial) torture or the same moral status for defensive (i.e. wrongful-harm-preventing) torture.

5.3. *Whence evidence?*

In order to correctly assess the reliability of intuitive moral judgments elicited by moral TEs, we would need to know more than we currently do about the mechanisms that typically produce them. As well as the mechanisms which typically distort them, when they go astray. Several competing psychological accounts are currently on the table, from a somewhat outdated and increasingly unpopular view that we form our moral judgments after careful deliberation, consciously weighing evidence for and against a given moral proposition (Kohlberg), to Jonathan Haidt's social intuitionist model (Haidt 2001 and 2012) and Joshua Green's dual (and later upgraded multi-) process theory (Green 2013) to Daniel Kahneman's two system theory (Kahneman 2011), as well as several recent attempts to identify, as the underlying psychological mechanism, moral, domain-specific heuristics (Sunstein 2005 and 2008, Gigerenzer 2008a, 2008b and 2008c).

Let me say a few words about moral heuristics, the explanatory account that I myself find most promising, and how these kinds of psychological mechanisms can explain both successes and failures of our moral intuitions. What is common to all heuristics? According to a prevalent view, heuristics include any mental short-cuts or rules of thumb that generally work well in common circumstances but may, and do, lead to systematic errors in untypical situations. This definition includes explicit rules of thumb, such as "Invest only in blue-chip stocks" and "Believe what scientists rather than priests tell you about the natural world." Unfortunately, this broad definition includes so many diverse methods that it is hard to say anything very useful about the class as a whole (Sunstein 2005). A narrower definition captures the features of the above heuristics that make them a suitable model for moral intuitions. On this narrow account, which I shall adopt here, all heuristics work by means of *unconscious attribute substitution* (Kahneman and Frederick 2005). A person wants to determine whether an object, X, has a target attribute, T. This target attribute is difficult to detect directly, often due to the believer's lack of information or time pressure. Hence, instead of directly investigating whether the object has the target attribute, the believer uses information about a different attribute, the heuristic attribute, H, which is easier to detect. The believer usually does not consciously notice that he is answering a different question: "Does object, X, have heuristic attribute, H?" instead of "Does object, X, have target attribute, T?" The believer simply forms the belief that the object has the target attribute, T, if he detects the heuristic attribute, H.

Assuming that this is how heuristics, the moral ones included, typically work, can we rely on them to deliver at least prima facie reliable judgments about hypothetical scenarios that moral philosophers devise with the aim of testing moral propositions? I'm afraid not. True, heuristics are mostly reliable tools of cognition. (Even Sunstein 2005 grants that.) And yet moral TEs are specific in respects that make misfiring more likely and render the deliverances of such heuristics less credible. Or so I'd like to claim in the remainder of this chapter.

First of all, examples of misfiring should alert us against carelessly using proxies for target moral properties. In Haidt's famous Incest Case, respondents seemed to have jumped automatically from the heuristic attribute, 'incestuousness' to a target attribute, 'impermissibility', flatly ignoring that the features that typically render incest wrong were all carefully removed from the story. The other case at hand is our wrought and fairly confused responsibility judgments.[22] Since the

[22] See Knobe and Doris (2010) for a frustratingly long list of inconsistencies, incoherencies, arbitrary asymmetries and confusions exhibited in the ordinary people's judgments of moral responsibility. Instead of taking all this compelling evidence as undermining any evidential value of the intuitive attributions of moral responsibility once and for all, however, the authors make a surprising u-turn and choose to treat this hodgepodge of conflicting criteria as evidence clearly falsifying

exact degree of the agent's responsibility is difficult enough to assess in real life cases, and is even more concealed in often tricky moral TEs, it is a fair bet that judgments of responsibility will be routinely formed by means of subconscious attribute substitution. The prevalence of this mechanism in their formation can partly explain why judgments of responsibility display such little stability and coherence overall. Whenever the target attribute is undetectable—and let's assume that Pizzaro and Tannenbaum (2011) are correct and responsibility judgments really are just covert character assessments or a shorthand to them—we resort to those contextual cues that are more readily available: the moral status of the action (is it harmful or not? does it violate any deontological constraints?), its likely consequences (overall positive or negative?), the intentions we ascribe to the agent based on those two (good or bad? selfish or unselfish?), and so on. The problem is that these proxies are only loosely correlated with the agent's character, and the latter is only vaguely connected to the degree of responsibility in any particular case under consideration. Moral TEs only amplify the problem. For we are trying to assess the relevance of different features for the moral status of action, or the degree of the agent's responsibility for it, and in order to do that we vary those very features—even to the point where all plausible candidates for morally relevant features are removed from the picture. And yet in these cases the rigid moral heuristic ("incest forbidden!") will, as Haidt's Incest Case shows, still deliver its verdict no matter what. The same applies to harmful actions, another common proxy—in reality, they may (or may not) be relatively strongly correlated with bad character and via bad character with blameworthiness, our target attribute. But not only is this connection clearly defeasible even in reality, the two features, the wrongness of actions and blameworthiness, will typically come apart in all sorts of ways in moral TEs. For in those, we are trying to determine the moral impact of various features and correspondingly hold some of them fixed while varying others regardless of how unlikely, or even impossible, such disassociations are in the real world. Accordingly, the harmfulness of an agent's actions may serve as a relatively reliable indicator (via badness of her character) of her blameworthiness in real life, but to keep using it as a proxy in moral TEs where all usual dependency relations are turned upside down,[23] strikes me as a rather short-sighted strategy.

Another characteristics of moral TEs amplifies the aforementioned effect. Moral TEs force us to resort to unreliable shortcuts, heuristics, even on those occasions when we are given enough time to consider various aspects of a hypothetical situation. This is so because the scenarios that are commonly used in vignettes, but to no less extent those

uniform, 'invariantist' (in fact merely internally coherent) philosophical accounts of moral responsibility.

[23] As in Glaucon's morally inverted world (MIW) where good people suffer bad reputation and bad people enjoy good reputation and excellent social standing (Plato 1993).

commonly discussed in philosophical literature, are commonly underdescribed and often devoid of both relevant information and wider context. It is plausible to assume, then, that when we are faced with the task of morally evaluating the agent's conduct in such informationally poor situations, the most optimal strategy is to resort to economical, informationally undemanding rules of thumb. For instance, when in Rachels' TE we judge Smith's and Jones' conduct morally equivalent, this judgment of equivalence can be best explained by the fact that we form an action judgment on the basis of prior character evaluation. In other words, we treat 'Smith and Jones are equally evil' as a proxy to 'what Smith and Jones did was equally wrong'. Other examples of such shortcuts that are simply convenient in normal contexts, but can become a matter of necessity in more philosophical ones where supplying extra information means changing the situation, shouldn't be difficult to find.

In moral (and even more so political) philosophy, the ease with which we assign blame to people for their destiny is disconcerting. On the one hand, judgments of moral responsibility or, more specifically, attributions of blame do play a crucial role in our moral and political judgment (where 'desert' is often a proxy for 'just' and 'fair' and 'desert' is a direct function of the agent's degree of 'responsibility'), on the other, however, they seem to be extremely responsive to morally irrelevant features of our natural and social world. As said before, our judgments of moral responsibility are hopelessly confused and incoherent. Alicke summarizes these depressing findings thus:

> it often seems that blame waxes and wanes imperfectly in relation to the evidence that implicates an individual in a harmful or offensive act. Even with all the usual criteria held constant (e.g., causation, intent, foresight, foreseeability, mitigating circumstances), personal values, unfortunate outcomes, emotional reactions, feelings of betrayal, antipathy for the harmdoer or sympathy for the victim, beliefs about the efficacy of forgiveness, and projections about future wrongdoings have an enormous impact on whether any blame occurs, how much of it is meted out, and how it evolves over time. (Alicke 2014)

People are stubborn moralists, inclined to blame other people for their actions ahead, and even in spite, of the evidence of the absence of intention and/or control, ascribe agency and goal-directed behaviour even to inanimate objects, and even readily accommodate judgments of causality and intentionality to reflect their antecedent moral judgments. (Pizarro and Helzer 2010) Furthermore, we tend to personalize social judgment and we tend to moralize personal judgment—when we ask of some hypothetical arrangement whether it would be just or not, people subconsciously understand this as asking "do people who would benefit from this arrangement, really deserve the (extra) benefits?" and in order to answer the latter question, resort to their character assessment. Which, in turn, is often heavily influenced by implicit bias and prejudice. And so a vicious circle is closed.

6. *Three preliminary qualifications*

In the previous chapter, I have presented some compelling evidence for the claim that our TE-generated moral intuitions are not to be trusted. Let me now qualify the scope of my criticism.

First, my disillusionment with mTE-evidentialism rests primarily on empirical findings which discredit one particular (albeit central) type of moral judgments and may fail to generalize to others. For all we know, judgments of responsibility (or blame) may be simply the most difficult type of moral judgments, and a-typically so.[24] The empirical findings presented could therefore leave other types of intuitive moral judgments (of action's rightness and wrongness, of agent's character, of virtues and vices, and the like) intact. The problem with this solution is that on some very influential moral theories judgments of moral responsibility are not just closely related to, but even constitutive of, these other types of moral judgments. So to say, for example, that what A did was wrong is to say that A is blameworthy, i.e. deserves blame for what he did. Personally, I find these accounts of moral wrongness mistaken, but if true, the damage of cutting corners in moral judgment and treating correlations and co-instantiations as indicative of some stronger dependency relations will be difficult to contain locally.

Alternatively, one could try to neutralize my attacks on TE-generated moral intuitions by separating lay intuitions from professional ones.[25] Not all philosophical intuitions count the same, or bear the same evidential weight, only professional philosophers' intuitions do. So, according to this, so-called expertise-defence, we should acknowledge that not all intuitions are created equal. Physical intuitions of professional scientists, for instance, are much more trustworthy than those of undergraduates or random persons in a bus station" (Hales 2006: 171) The mathematical intuitions of professional mathematicians are similarly more trustworthy than those of the folk. So it might seem reasonable to expect philosophical intuitions of professional philosophers to be more trustworthy than the intuitions of typical subjects of experimental philosophy. In the light of this, the practice of appealing to *philosophical* intuitions about hypothetical cases, properly construed, should be the practice of appealing to *philosophers'* intuitions about hypothetical cases. Correspondingly, we should dismiss studies conducted on the intuitions of untutored folk as providing no evidence at all against the evidentiary role of TE-generated moral intuitions. For reasons I cannot go into here, I don't find this line of argumentation particularly promising, but it would be unwise and unfair to disqualify it outright and without a compelling argument.[26]

[24] I tried to offer an alternative, more unifying (but also admittedly more counterintuitive) account of moral responsibility in Klampfer (2014).

[25] As Bengson 2013 and Wong 2018 try to do, among others.

[26] See Weinberg et al (2010) and Schwitzgebel and Cushman (2015) for serious doubts that the epistemic credentials of professional philosophers' intuitions surpass those of lay people.

Thirdly, deep divisions over the correct normative moral theory make it difficult, if not impossible, to find a noncontroversial set of criteria for classifying moral cognizers' performance as success or deriding it as failure. As Robert Shaver correctly remarked about our practice of responsibility attributions long ago:

> In a perfectly fair and rational attributional world, according to the precepts of Anglo American jurisprudence and rational decision theory, blame attributions would be derived by assessing whether (i) the action violated some valid moral or legal norm (i.e. was either harmful or wrongful or illegal); (ii) a perpetrator's action were intentional, reckless, or negligent; (iii) the consequences were foreseen or foreseeable; (iv) to what extent the perpetrator's behavior caused the harmful consequences or could potentially have done so; and (v) any mitigating circumstances prevailed. In the attributional world in which we live, however, a host of biasing factors influences blame and responsibility judgments. (Shaver 1985, quoted in Alicke and Zell 2009: 2101)

In fact, assuming even this much shared agreement on the criteria of success is somewhat naïve and prejudicial, at least when our focus are attributions of moral, as opposed to legal, responsibility. The truth is that no such widely shared agreement on the features that are individually necessary and jointly sufficient for determining the agent's degree of blame (let alone appropriate punishment) is currently at hand. And this is not accidental—it is in principle much easier to measure the performance of a non-moral heuristic, which is measured against demonstrable facts and the laws of logic and probability, all relatively undisputed;[27] determining whether a moral heuristic misfired in delivering a particular moral judgment or not is much harder, since there is often very little agreement on what the correct moral assessment of the case at hand should be.

Finally, the jury assessing the merits of competing psychological accounts of intuitive moral judgment is still out; and, as we've seen, some of the candidates for what was traditionally called 'the faculty of moral intuition' fare better than others. Nevertheless, none of the proposed accounts of what goes on in one's mind when one spontaneously judges some action right or wrong, or someone culpable or innocent of some moral offence, has so far managed to win the undivided support of the majority of psychologists. But as long as the jury assessing the merits of competing psychological accounts of intuitive moral judgment is still in session, we cannot but for the time being suspend our final verdict on the credibility of TE-generated moral intuitions.

[27] Here I am simplifying a bit. In fact, as we learn from a long stand-off between the most vocal critic and proponent of heuristics, Kahneman and Gigerenzer, criteria of success are not so uncontroversial even when it comes to people's apparently objective probability and risk assessments and human decisions grounded on them. For a brief, yet instructive overview of the dividing issues see Gigerenzer 2008c.

7. *Hypothetical reasoning in moral philosophy*

Once we abandon the idea of moral TEs as a potential source of evidence, or justification, of moral propositions, is there any room left in moral philosophy at all for reasoning about hypothetical, counterfactual situations? Plenty. By renouncing mTE-evidentialism, we don't need to deprive ourselves of the many benefits of hypothetical reasoning. We can still use it to improve our understanding and deepen our knowledge of various moral and political issues: in the form of abstractions, idealizations, as well as for illustration, implication and exemplification (O'Neill 1987). Furthermore, there is room in moral (and political) philosophy for what I'd like to call 'normative forecasting'—assessments of whether a given political, social, legal, and so on change in the world would constitute moral progress or regress (see Feinberg 1970 and Nussbaum 1997). We don't even need to give up thought-experimenting altogether. We can continue to use moral TEs for diagnostic purposes—to help us identify psychological mechanisms that are operative in the formation of our intuitive moral judgments (Knobe 2007). And we can keep using moral TEs as a valuable source of *hypotheses for further testing*.[28]

That's not all. Even if hypothetical scenarios cannot resolve any disputes in moral and political philosophy, they can be instrumental in alerting us to the inconsistencies in our belief system, thus prompting further thinking and discussion.[29] In other words, the point of hypothetical scenarios such as Judith Thomson's Violinist is not so much to prove the proposition that abortion is permissible (at least in cases where conception results from rape), but rather to alert those who find it impermissible, but also happen to deny the existence of duties of assistance to people in need, of potential inconsistency in their belief-set. So apart from helping us better understand the workings of our minds and providing hypotheses for further investigation, contemplating such scenarios can also prompt us to reconsider our moral and political values—not because a single moral TE has proven any of them wrong but rather because our particular response to them gives rise to suspicion that we may subscribe to two or more conflicting principles. In and

[28] The difference between using TE-generated intuitions as pieces of evidence and using them as hypotheses for further testing is not the easiest to spell out. I find the following criterion offered by Herman Cappelen helpful: Are we using a particular TE-generated intuition (a) as a datum which confirms, or lends support, by way of abductive reasoning, to some contested principle or theory, and at the same time disconfirms other, rival ones; or are we using it (b) to generate, or suggest, possible explanations (or justifications) of the observed moral phenomenon which only further, independent investigation can either confirm or disconfirm? That is, are we treating this intuition as (a)an established fact that calls for an explanation (but no further confirmation), or as (b) a mere hypothesis in need of further testing and (dis)confirmation?

[29] This was suggested in a post by Harry Brighouse on the online forum Crooked Timber.

by themselves, the intuitions thus generated would give no advice as to which of those conflicting beliefs we should abandon; they will merely force us to critically re-examine them. I can happily accept this.

Last but not least, hypothetical (i.e. abductive) reasoning could be used in political philosophy for what Miščević (2013a) labels 'rational (as opposed to historical) reconstruction' of particular social institutions, norms and practices. Think of John Locke and his incredibly influential attempt to provide rational grounds for the institution of private property—a rational reconstruction of how you can get from the initial state of nature where, presumably, (i.e. according to biblical testimony) nobody owned anything, to the current state of affairs where most goods (land, houses, farms, woods, cars, and so on) are owned by someone, be it private individuals or companies/corporations or states (Locke 1980). Or think of Hobbes and his attempts to rationally reconstruct the path from absolute freedom, enjoyed in the state of nature, to absolute monarchy, his preferred form of government (Hobbes 1998). At least on the face of it, rational reconstruction does not presuppose the thinker's engagement in classical TEs or the use of intuitions, thereby generated, to support her claims. I suspect this use of hypothetical reasoning will be problematic, if it turns out to be such, for reasons other than the ones that make mTE-evidentialism unattractive. But that's a topic for another paper.

8. *Conclusion*

Let me conclude. In the paper, I argued against a particular use of thought-experimentation in moral philosophy, a view that I labelled 'mTE-evidentialism'. According to this view, moral TEs (or, rather, moral intuitions that they elicit in response) are a valuable source of evidence for and against moral propositions (particular and general moral judgments, principles, distinctions, theories, and so on). Such epistemic credentials, I argued, are mostly unfounded.

The past record of moral TEs is far from impressive. Most, if not all, moral TEs fail to corroborate their target moral hypotheses (provided one can determine what results they produced and what moral proposition these results were supposed to verify or falsify). Moral intuitions appear to be produced by moral heuristics with not just fairly bad general track record, but the ones that we have good reasons to suspect will regularly misfire in typical moral TEs. Rather than keep relying on moral TEs, we should begin to explore other, more sound alternatives to thought-experimentation in moral philosophy.

References

Alicke, M. D. 2014. "Evaluating Blame Hypotheses." *Psychological Inquiry* 25: 187–192.

Alicke, M. D. and Zell, E. (2009). "Social attractiveness and blame." *Journal of Applied Social Psychology* 39: 2089–2105.

Bengson, J. 2013. "Experimental attacks on intuitions and answers." *Philosophy and Phenomenological Research 86 (*3): 495–532.

Bourget, D. and Chalmers, D. J. 2014. "What do Philosophers Believe?" *Philosophical Studies* 170 (3): 465–500.

Cappelen, H. 2011. *Philosophy Without Intuitions*. New York: Oxford University Press.

De Smedt, J. and De Cruz, H. 2015. "The epistemic value of speculative fiction". *Midwest Studies in Philosophy* 39: 58–77.

Deutsch, M. 2015. *The Myth of the Intuitive. Experimental Philosophy and Philosophical Method*. Cambridge: The MIT Press.

Feinberg, J. 1970. "The nature and value of rights." *The Journal of Value Inquiry* 4: 243–257. Reprinted in Feinberg 1980. *Rights, Justice & the Bounds of Liberty*. Princeton: Princeton University Press: 143–58.

Feinberg, J. 1985. *Offence to Others. The Moral Limits of Criminal Law*. Vol. 2. Oxford: Oxford University Press.

Frederick, S. 2005. "Cognitive reflection and decision making." *Journal of Economic Perspectives* 19 (4): 25–42.

Gendler Szabo, T. 2007. "Philosophical thought-experiments, intuitions and cognitive equilibrium." *Midwest Studies in Philosophy* 31: 68–89.

Gigerenzer, G. 2008a. "Moral Intuition = Fast and Frugal Heuristics?". In W. Sinott–Armstrong (ed.). *Moral Psychology*. Vol. 2: The Cognitive Science of Morality: Intuition and Diversity. Cambridge: A Bradford Book / The MIT Press: 1–26.

Gigerenzer, G. 2008b. "Reply to Comments". In W. Sinott-Armstrong (ed.). *Moral Psychology*. Vol. 2: The Cognitive Science of Morality: Intuition and Diversity. Cambridge: A Bradford Book / The MIT Press: 41–45.

Gigerenzer, G. 2008c. "Why heuristics work." *Perspectives on Psychological Science* 3 (1): 20–29.

Greene, J. 2013. *Moral Tribes. Emotion, Reason, and the Gap Between Us and Them*. New York: The Penguin Press.

Haidt, J. 2001. "The emotional dog and its rational tail." *Psychological Review* 108 (4): 814–34.

Haidt, J. 2012. *The Righteous Mind. Why Good People Are Divided by Politics and Religion*. New York: Pantheon Books.

Hobbes, T. 1998. *Leviathan*. Ed. by J.C.A. Gaskin. Oxford: Oxford University Press.

Huemer, M. 2013. "Phenomenal conservatism." *Internet Encyclopedia of Philosophy*. URL: https://www.iep.utm.edu/phen-con/

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kamm, F. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

Klampfer, F. 2014. "Consequentializing moral responsibility." *Croatian Journal of Philosophy* 14 (1): 121–150.

Klampfer, F. 2017. "The false promise of thought-experimentation in moral and political philosophy." In B. Borstner and S. Gartner (eds.). *Thought Experiments between Nature and Society. A Festschrift for Nenad Miščević*. Newcastle upon Tyne: Cambridge Scholars: 328–348.

Knobe, J. 2007. "Experimental philosophy and philosophical significance." *Philosophical Explorations* 10 (2): 119–121.

Knobe, J. and Doris, J. 2010. "Responsibility." In Doris, J. (ed.). *The Moral Psychology Handbook*. Oxford and New York: Oxford University Press.

Kuntz J. R. and Kuntz J. R. C. 2011. "Surveying Philosophers About Philosophical Intuition." *Review of Philosophy and Psychology* 2: 643–665.

Levy, S. S. 2004. "A limit on intuitionistic methods of moral reasoning." *Journal of Value Inquiry* 37: 463–470.

Liao, M. S. et al 2012. "Putting the trolley in order: Experimental philosophy and the loop case." *Philosophical Psychology* 25 (5): 661–671.

Locke, J. 1980. *Second Treatise of Government*. Ed. by C. B. Macpherson. Indianapolis: Hackett Publishing House.

Luban, D. 2008. "Unthinking the Ticking Bomb." *Georgetown Law Faculty Working paper*. URL: http://lsr.nellco.org/georgetown/fwps/papers/68/

McMahan, J. 2002. *The Ethics of Killing*. New York: Oxford University Press.

McMahan, J. 2008a. "Torture in Principle and in Practice". *Public Affairs Quarterly* 22 (2): 91–108.

McMahan, J. 2008b. "Torture and method in moral philosophy." In S. Anderson and M. Nussbaum (eds.). *Torture, Law, and War*. Chicago: University of Chicago Press.

Miščević, N. 2004. "The explainability of intuitions." *Dialectica* 58 (1): 43–70.

Miščević, N. 2007. "Modelling intuitions and thought-experiments." *Croatian Journal of Philosophy* 7: 181–214

Miščević, N. 2013a. "In search of the reason and the right: Rousseau's social contract as a thought experiment." *Acta Analytica* 28 (4): 509–526.

Miščević, N. 2013b. "Political thought-experiments from Plato to Rawls." In M. Frappier, L. Meynell and J. R. Brown (eds.). *Thought Experiments in Philosophy, Science, and the Arts*. New York and London: Routledge: 191–206.

Miščević, N. 2013c. "The ontology of secondary and tertiary qualities." *Balkan Journal of Philosophy* 5 (1): 45–58.

Miščević, N. 2015. "Intuitions: reflective justification, holism and apriority." *Croatian Journal of Philosophy* 15 (3): 307–323.

Mizrahi, M. 2014. "Does the Method of Cases Rest on a Mistake?"*Review of Philosophy and Psychology* 5 (2): 183–197.

Norton, M. and Ariely, D. 2011. "Building a Better America. One Wealth Quintile at a Time". *Perspectives on Psychological Science* 6 (1): 9–12.

Nozick, R. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Nussbaum, M. 1997. "If Oxfam ran the world." *London Review of Books* 19 (17): 18–19.

O'Neill, O. 1987. "Abstraction, Idealization and Ideology in Ethics." *Royal Institute of Philosophy Lectures* 22: 55–69.

Pizarro, D. A. and Helzer, E. G. 2010. "Stubborn Moralism and Freedom of the Will." In Baumeister, et al. (eds.). *Free will and Consciousness: How Might They Work*?. New York: Oxford University Press: 101–120.

Pizarro, D. A. and Tannenbaum, D. 2011. "Bringing character back: How the motivation to evaluate character influences judgments of moral blame." In M. Mikulincer and P. R. Shaver (eds.). *The Social Psychology of Morality: Exploring the Causes of Good and Evil*. Washington: American Psychological Association: 91–108.

Plato 1993. *The Republic*. Transl. by Robin Waterfield. Oxford: Oxford University Press.

Rachels, J. 1975. "Active and passive euthanasia." *The New England Journal of Medicine*, 292 (9): 78–80.

Schwitzgebel, E. and Cushman, F. 2015. "Philosophers' biased judgments persist despite training, expertise and reflection." *Cognition* 141: 127–137.

Singer, P. 1993. *Practical Ethics*. Second Edition. Cambridge: Cambridge University Press.

Sokol, Daniel 2006. "What if… the results." URL: http://news.bbc.co.uk/2/hi/uk_news/magazine/4971902.stm

Sunstein, C. R. 2005. "Moral heuristics." *Behavioral and Brain Sciences* 28: 531–73.

Sunstein, C. R. 2008. "Fast, Frugal, and (Sometimes) Wrong.". W. Sinott-Armstrong (ed.). *Moral Psychology*. Vol. 2: The Cognitive Science of Morality: Intuition and Diversity. Cambridge: A Bradford Book / The MIT Press: 27–30.

Thomson, J. J. 1971. "A defense of abortion." *Philosophy and Public Affairs* 1 (1): 47–66.

Unger, P. 1995. *Living High and Letting Die. Our Illusion of Innocence*. Oxford: Oxford University Press.

Walsh, A. 2011. "A moderate defence of the use of thought experiments in applied ethics." *Ethical Theory and Moral Practice* 14: 467–481.

Walzer, M. 1973. "Political action: the problem of dirty hands." *Philosophy and Public Affairs* 2: 160–80.

Wang, T. 2018. "The experimental critique and philosophical practice." *Philosophical Psychology* 31 (1): 89–109.

Weatherson, B. 2014. "Centrality and Marginalization." *Philosophical Studies* 171 (3): 517–533.