Published in Borstner, B. & Gartner, S. (Eds.). Thought Experiments between Nature and Society. A Festschrift for Nenad Miščević. Cambridge Scholars Publishing, 2017; pp. 328-348. A published version of the paper is available upon request.

FRIDERIK KLAMPFER

THE FALSE PROMISE OF THOUGHT-EXPERIMENTATION IN MORAL AND POLITICAL PHILOSOPHY

Abstract

Prof. Miščević has long been an ardent defender of the use of thought experiments in philosophy, foremost metaphysics, epistemology and philosophy of mind. Recently he has, in his typically sophisticated manner, extended his general account of philosophical thought-experimenting to the domain of normative politics. Not only can the history of political philosophy be better understood and appreciated, according to Miščević, when seen as a more or less continuous, yet covert, practice of thought-experimenting, the very progress of the discipline may crucially depend on finding the right balance between the constraints of (biological, psychological, economic, political, and so on) reality and political-moral ideals when we set to design our basic political notions and institutions.

I have much less confidence in this project than prof. Miščević does. As a subspecies of moral TE, political TE share all their problems plus exhibit some of their own. In the paper, I present and discuss two types of evidence that threaten to undermine political philosophers' trust in thought-experiments and the ethical/political intuitions elicited by them: (i) the dismal past record of thought-experimentation in moral and political philosophy; and (ii) variety, prevalence, and stubbornness, of bias in social/political judgment.

1. Introduction

Thought-experimentation in philosophy has a long 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 and its credentials are more controversial, however. TEs seem to abound in epistemology, philosophy of mind, and metaphysics, and they are certainly also popular among the ethicists. Political philosophy, on the other hand, seems to have largely ignored their potential. To its own detriment?

Not quite. Prof. Miščević has recently argued, in a series of papers, in favour of extending his general account of philosophical thought-experimentation to the domain of normative politics. Not only can we shed new light on, as well as better understand and appreciate, the history of political philosophy by looking at it through the lens of the continuous rivalry between more empirically oriented theorizing and one based on thought-experimentation, but the very progress of the discipline may crucially depend on finding the right balance

between the constraints of (biological, psychological, economic, political, and so on) reality on the one hand and counterfactual reasoning on the other when we set out to design our basic political institutions and arrangements.

I have much less confidence in this project than prof. Miščević. As a subspecies of moral TE, political thought-experiments (PTEs) share all their shortcomings while also exhibiting some of their own. In the paper, I present and discuss two types of evidence that in my opinion undermine our trust in political thought-experiments and the intuitions elicited by them: (i) the abysmal record of thought-experimentation in the history of (moral and) political philosophy, and (ii) more general considerations aimed at explaining why this failure is not accidental.

2. Hypothetical reasoning and thought-experimentation

Hypothetical reasoning is ubiquitous and indispensable in moral & political philosophy. Here are some reminders of how often, and without much thought, we use it for guidance, judgment or as a moral heuristic:

What if everyone did that?

The Golden Rule (would I want to see X done if I were at the other, receiving end of the action?)

Kant's Universalizability Test (can you conceive, or will, without contradiction a world in which everyone acted on the given maxim, i.e. a world in which this maxim became universal law?)

The substitute-judgment test for (proxy) consent/authentic will (doing X to A is in accordance with A's will, iff A would've consented to X, had she been competent to judge)

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 (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 wrongness/rightness.

Whether these non-reductive accounts of various normative properties are correct or not, they are 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 variety of normative properties. In this paper, I'm not suggesting we should abandon such reasoning in moral and political philosophy as utterly useless. As I'll explain in the concluding section of the paper, I have no quarrel with other possible uses of M&PTEs: to illustrate particular principles (or clashes between them), to generate hypotheses for further testing, to draw implications, to generate a philosophical puzzle or riddle, or simply to motivate further reflection/thought on an issue that might otherwise look deceptively simple and straightforward. My specific

target is what I will call 'TE-evidentialism', i.e. a view which treats TE-generated intuitions as (at least prima facie) reliable pieces of evidence for or against normative theories and/or principles, i.e. accords them at least some (initial, even though defeasible) credibility, justifiability, epistemic value, and the like.

So what makes an exercise in imagination a thought-experiment? When does a piece of hypothetical reasoning qualify as a thought-experiment? Here I'm simply going to follow Jim Brown & Yiftach Fehige (Brown & Fehige 2014) and Tamar Gendler (Gendler 2007). In order for a piece of imaginative (=counterfactual?) thinking to qualify as a TE, we need to engage in it for a specific reason – namely to test a hypothesis that cannot be tested reliably in any other way. Either because we couldn't possibly monitor all the relevant variables, or because even though we could design a small or a large-scale social experiment to test the hypothesis under consideration, weighty moral considerations speak against doing so. Or, as Tamar Gendler (whom Nenad quotes approvingly) elegantly puts it: "To perform a thought experiment is to reason about an imaginary scenario *with the aim of confirming or disconfirming some hypothesis or theory*". (ibid.; my emphasis)

The idea, then, of experiments conducted in thoughts, is simple.[1] A controversial philosophical 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 the right; seeing, on your way to a college graduation, the kid drowning in a pond; having a magical ring that can render you invisible and, by extension, impunity, and so on), ask people to think, and form a judgment, about it (would it be permissible, right, morally good/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 judgment formed in response to such a hypothetical scenario qualify as intuitive? Here, again, I'm simply going to follow the tradition: intuitive judgments are characterized by a distinct genealogy (they spring into one's mind effortlessly. 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); phenomenology (they strike us as vivid, clear, inescapable, forced upon us); modality (they present things as being necessarily the way they appear before our mind); and epistemic status (they strike us as self-evident, beyond any doubt, as inconceivably at odds with reality/the truth). (Miščević 2004)[2]

---

1 Deceptively so, as we'll see later.

2 I'd like to stipulate that of the aforementioned defining features, what Herman Cappelen calls epistemic 'Rock status' is the 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.

2.1. TEs in moral and political philosophy

On the standard view, which Miščević himself endorses, M&PTEs are used to access the normative realm. They're meant to facilitate our ascent to normative (and/or conceptual) 'facts' or 'truths' as well as their interdependence, i.e. the internal structure of normative reality; perhaps less controversially, they are sometimes (Miščević 2004, Goodin 2008) taken to help uncover the underlying causal and modal structure of the social and natural world.

Here are some typical questions that philosophers aim to answer by means of TEs: Is it ever permissible to lie? May we kill one to save five? Is it permissible to go to war? Is there such a thing as blameless wrongdoing? Is harming always worse than merely allowing harm? Should we punish the most heinous crimes by death? What is just(ice)? When, if ever, is the rule of some people over others legitimate? What form of government is best? Is political violence, violence in the service of political goals, ever permissible? What makes one group's claim to a particular territory legitimate, or at least more legitimate than another's?

Having thus delineated TEs from other (perfectly legitimate) forms and uses of hypothetical reasoning in moral and political philosophy that don't qualify as TEs, there are still plenty examples left that meet the above criteria and hence qualify as MTEs or PTEs. Here is a random selection of hypothetical scenarios and corresponding hypotheses that the former are designed to confirm or disconfirm:

(i) The ring of Gyges -> morality/justice is instrumentally, but not also intrinsically, valuable (Plato 1993)

(ii) The ticking bomb -> there can be a liability-based moral justification for particular cases of defensive torture (as well as, far more controversially, its legalization) (McMahan 2008)

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

(iv) Goodin's No-Party Democracy –> there can be no functioning democracy without political parties (Goodin 2008)

(v) Rawls' Original Position –> justice as fairness (i.e. as the object of agreement that properly motivated rational individuals would have reached under fair conditions) (Rawls 1971/1999)

(vi) Singer's Pond –> positive duties to help the poor (alt. moral legitimacy of Samaritan laws, i.e. obligatory assistance) (Singer 1993)

(vii) Singer's Shelter/Fairhaven –> there is no moral justification for hermetically closed borders, or restrictive laws on (im)migration (ibid.)

(viii) 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' liberties (Feinberg 1985)

(ix) Garrett Hardin's Pasture –> tragedy of the commons hypothesis (free commons – i.e. when the benefits of using some public good are privatized, but the costs are externalized – is destined to go to ruin) (Hardin 1968)

The above list is far from complete. And yet, given the frequency and relative popularity of the method, the results of thought experimentation in moral and political philosophy are discouraging, to say the least.[3] Hardly any controversial issue in moral or political philosophy has been settled, or brought a bit closer to resolution, by means of M&PTEs. How come?

3. TE-evidentialism

As said before, my target in this paper is not counterfactual thinking or reasoning as such, but rather the view that for want of a better name I'll call TE-evidentialism. Also, for reasons that will become apparent later, I'd like to limit my pessimism concerning TE-evidentialism to the domain of moral and political philosophy. In other words, I will not take a stance on either the frequency or the merits of thought-experimentation in other areas of philosophy, from metaphysics over epistemology to philosophy of mind. Though my uninformed guess is that the method, once subject to scrutiny, won't fare much better in those other areas as well.

Here, then, is the view that, unfortunately and for too long, has dominated contemporary moral and political philosophy.

(TE-evidentialism): Intuitive judgments formed in response to M&PTEs, provide some initial, prima facie credible evidence for or against normative propositions (principles, norms, theories of distributive and retributive justice, legitimacy, political authority and/or obligation, and the like)

Nenad clearly embraces PTE-evidentialism or something close to it. (see, for example, Miščević 2007) He advocates the view of TEs as condensed arguments/reasoning. (ibid.) His favourite examples are geometrical and arithmetical as well as metaphysical intuitions, but in his more recent papers (see Miščević 2013a and Miščević 2013b), he has extended his account, and defence, of philosophical TEs to moral and political philosophy. He distinguishes seven or eight stages of the process, from visualization and understanding of the issue at hand via the formation of a particular judgment and its generalization to, finally, a narrow and wide equilibrium test. Epistemically, the most controversial step, or, shall I say, leap, is from a particular judgment ("such and such imagined/hypothetical social arrangement would (not) be fair/just", or 'such magic ring would transform me into an unjust, or immoral, person') to the conclusion that no social arrangement that shares all the relevant universal features can be just and, finally, that this is a matter of necessity, not

3 I'm not taking a stance on thought-experimentation in other areas of philosophy, such as metaphysics, epistemology, philosophy of mind, philosophy of language, though it seems to me that fairly little progress (in terms of knowledge growth and/or better understanding of the issue under consideration) has been made thanks to Gettier- or Frankfurt- or Lehrer- or Chalmers-types of cases in those areas of philosophical inquiry as well.

simply a contingent feature of our social world. Psychologically, we seem to find these transitions fairly easy and natural to make, but what, if anything, warrants them? What are the epistemically relevant features of PTE-generated intuitions? Admittedly, they share most of their phenomenal properties with other TE-generated intuitions, but do they so clearly inherit their epistemic status as well?

Here, then, is my main worry. Why treat any answer to the question "Imagine/consider such and such a situation? Would it instantiate such and such a normative property or not?" as authoritative? Why treat our swift, spontaneous judgments, whether particular or general, as revealing anything else but how **our mind** works; how **we feel and think about the world**?

4. A case study: Feinberg's Ride on the Bus

Let's illustrate the limited potential of M&PTEs by way of a telling example, Joel Feinberg's famous Ride on the Bus vignettes (Feinberg 1985). Feinberg introduces his 39 variations on the Ride on the Bus story with the following interesting methodological remarks:

"The question raised by this chapter is *whether there are any human experiences that are harmless in themselves yet so unpleasant that we can rightly demand legal protection from them even at the cost of other persons' liberties***.** The best way to deal with that question at the start is to engage our imaginations in the inquiry, consider hypothetically the most offensive experiences we can imagine, and then sort them into groups in an effort to isolate the kernel of the offense in each category. *Accordingly, this section will consist of a number of vividly sketched imaginary tales, and the reader is asked to project himself into each story and determine as best he can what his reaction would be. In each story the reader should think of himself as a passenger on a normally crowded public bus on his way to work or to some important appointment in circumstances such that if he is forced to leave the bus prematurely, he will not only have to pay another fare to get where he is going, but he will probably be late, to his own disadvantage. …In each story, another passenger, or group of passengers, gets on the bus, and proceeds to cause, by their characteristics or their conduct, great offense to you.*" (Feinberg 1985; 10-11; my emphasis)

The 39 vignettes, then, are designed to help us answer the question of whether – or, in the jargon used here, to test the hypothesis that – "there are some human experiences that are harmless in themselves yet so unpleasant that we can rightly demand legal protection from them even at the cost of other persons' liberties". Not only will imagining oneself in these situations, or, as Feinberg puts it, 'projecting oneself on the bus', provide conclusive

evidence for some version of his famous offence principle,[4] but nothing else, he insists, and certainly no amount of reasoning, could do the job:

"I have tried to make a number of different points by telling these bloodcurdling tales: that there are at least six distinguishable classes of offended states that can be caused by the blameable conduct of others; that to suffer such experiences, at least in their extreme forms, is an evil; but that to the normal person (like the reader) such experiences, unpleasant as they are, do not cause or constitute harm. *It is very important that the reader put himself on the bus and imagine his own reactions, <u>for no amount of abstract argument can convince him otherwise that the represented experiences are in principle of a kind that the state can legitimately make its business to prevent</u>*." (ibid. p. 14, my emphasis)

But I find Feinberg's remarks utterly puzzling. How could these brief vignettes, however vivid they may be, make/help us realize that 'suffering certain offences is an evil', that such evil 'may not constitute harm', but that it nevertheless 'justifies state interventions in other people's liberties'? How exactly is imagining yourself as a passenger on a crowded bus, who experiences various types of inconveniences, and trying to imagine what your respective reactions would be to a whole range of undoubtedly unpleasant sensory and cognitive stimuli, supposed to answer our initial question, specifically whether there are any human experiences that are *harmless* in themselves yet so unpleasant that we can *rightly* demand legal protection from them even at the cost of other persons' liberties?

Surely, it can prompt me to begin considering some such inconveniences bad enough to warrant a corresponding legal ban, and even convince me to join a public campaign in support of such a ban. But none of this should obscure the obvious fact that what the whole intellectual exercise comes down to, in the end, is an informed guess about how such vividly imagined experiences *would make me feel*, and whether, as a result of that, *my attitude would shift from the initial one of indifference to the resulting one of intolerance*. Needless to say, none of this either carries in itself, or confers, any epistemic credibility to Feinberg's famous offence principle.

This much, at least, becomes clear when we consider the difference between the following two questions:

(i) "Do I find any inconveniences experienced during my imaginary bus ride bad, or unpleasant, enough to *want to see myself protected* against them by law?",

and

---

4 Feinberg states his offense principle as follows: "it is always a good reason in support of a proposed criminal prohibition that it would probably be an effective way of preventing serious offense (as opposed to injury or harm) to persons other than the actor, and that it is probably a necessary means to that end (i.e., there is probably no other means that is equally effective at no greater cost to other values)." And then summarizes it thus: "The principle asserts, in effect, that the prevention of offensive conduct is properly the state's business." And he defines 'offensive conduct' as one which (a) causes some inconvenience, nuisance, or discomfort, and (b) does so wrongfully, i.e. in violation of one of the victim's rights (or at least without sufficient justification or valid excuse).

(ii) "Are any inconveniences caused to us by other passengers on the imaginary bus *bad*, *or unpleasant*, *enough to warrant* a legal ban, i.e. so bad that they would *justify* circumcising other people's liberties?".

The first question merely asks us to predict our reactions to a hypothetical situation (to various types of 'nuisance') and Feinberg's vignettes may or may not be instrumental in finding a correct answer to this question, depending on how well we know ourselves.[5] The second, on the other hand, is a normative question that no amount of imagined or veridical experiences, however vivid, can ever answer on its own.

So I simply fail to see the epistemic advantage that Feinberg claims for the TE method in the given context. It may well be that without me personally experiencing, even if only in thought or imagination, the humiliation, the embarrassment, the affront to one's senses or the shock to one's religious sensibilities or patriotic feelings that others can cause with either their behaviour or mere appearance, I cannot know how bad these things *can make one feel*. And yet, I wonder, how can even the most vividly represented subjective experience provide the required type of insight, namely that these are the kinds of experiences (that are so bad) *that the state may legitimately protect us against by curtailing other people's liberty to do things that cause such offence*?

Let me summarize, then. I've raised two sets of objections against Feinberg's proposed TE methodology. He is asking us to take (i) something that is merely imagined rather than genuinely experienced, as (ii) tracking normativity rather than, say, revealing our sensitivity, i.e. sensory and emotional responsiveness, to smell, nudity, profanity, shrewdness, abjectness and the like. But such an inference from imagined experience to the truth of a normative proposition is neither supported nor sound.

5. More general worries concerning M&PTEs

Showing an instance of M&PTE flawed is not the same as discrediting the method as such, of course. In what follows, I'll present and briefly discuss eight types of considerations that, when properly acknowledged, should significantly reduce our level of confidence in the capacity of M&PTE-generated intuitions to resolve normative disputes, or, at a minimum, test alternative normative hypotheses. Here is a tentative list:

(i) unresolved disputes over experimental design

(ii) indeterminacy of M&PTE results

(iii) confusion over the correct level of generality

(iv) mistaken moral arithmetic

(v) vicious circularity

5 Even though, given people's poor record in what psychologists call 'affective forecasting', i.e. assessments of the impacts of different types of events on our future subjective well-being, a real instead of an imaginary bus ride would be much better suited for this task.

(vi) sensitivity, or responsiveness, to morally irrelevant features (framing effects, order of presentation,…)

(vii) reliance on dubious moral heuristics

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

These are repeatedly identified and fairly well-documented shortcomings of both moral and political TE-generated intuitions. For reasons of brevity, I'll restrict my discussion to (ii), (iii), (vi), (vii) and (viii).

(iii) Indeterminacy of M&PTE results

Ideally, an experiment, whether conducted in a lab or in one's mind, would yield results that, whether measurable or not, whether quantitative or qualitative, are unequivocal. Philosophical TEs, moral and political in particular, fall far below this ideal, however. It appears that the more controversial and divisive an issue, the more evenly distributed the judgments are that we get in response to a given TE.

Take one of the probably best known MTEs, Judith Thomson's Violinist. You wake up in hospital, next to a world famous violinist connected to you with various tubes. You've been kidnapped by the Music Appreciation Society, you are told by the doctor. Aware of the maestro's impending death, they hooked you up to the violinist. If you stay connected, he will be totally cured in nine months. You are unlikely to suffer harm. No one else can save him. Do you have an obligation to stay connected? According to a relatively recent BBC online survey (Sokol 2006), one in four of the almost 58.000 participants opted for 'yes' and three in four for 'no'. Opinion was almost identically divided with respect to another famous MTE, 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. 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'. 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 choice will nevertheless remain bleak. A survey of 1,972 contemporary philosophers, conducted via PhilPapers (Bourget & Chalmers 2014), produced 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.[6] 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 TE failed to deliver the verdict in this case, then? And

6 I've lumped all other categories under 'other' to come up with 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.

if not here, what ratio of 'yes' to 'no' answers would be enough to validate such a conclusion?[7]

(iv) confusion over the correct level of generality

This is a continuous worry, as nicely illustrated by another famous MTE, Singer's Pond. What exactly is it that we 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 the option that Singer himself prefers, (c), 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 principle, we can fairly easily come up with a counter-example to the principle (as Peter Unger has shown with another MTE, called Envelope) and this will set our inquiry back to the beginning. All that we clearly intuit in Pond is that we ought to pull the drowning child out of the pond, since nobody else is around to help and we can rescue the child at an insignificant cost. Everything else is extrapolation and generalization and insofar questionable.[8]

(v) sensitivity to morally irrelevant features (framing effects, order of presentation, level of abstractness, likeness,…)

One important reason for distrusting TE-generated intuitions is their malleability, instability, and vulnerability to manipulation. Our intuitions are 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. What psychologists tell us about the mechanisms that produce them and what we know influences them doesn't exactly build confidence. Intuitions are quick, snap, unreflective, spontaneous, almost automatic judgments; if preceded by any reasoning at all, it must be subconscious; they rely, for their formation, on similar cognitive shortcuts, heuristics, that people use in their judgments in other domains (availability, representativeness); they are subject to the framing effects ('lives not saved' vs. 'lives lost') and moralizing spill-over effects; shaped by mood, affection, emotion, fatigue, affected by the level of abstraction and sensitive to the order of presentation; despite their contingent origin, 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').

---

7 The more complicated the variants on the default thought experiment become (Fat man or Bridge, Loophole, and so on), the faster we can expect the last group, the 'other' or the 'undecided', to grow and, correspondingly, the quicker that little initial agreement to dissolve.

8 Nenad, I believe, underestimates this problem. See, for example, his rather casual remarks about the generalization stage in Miščević (2013b).

In short, our intuitive responses to TEs seem to track a host of morally irrelevant features of the world (such as novelty, excitement, disgust, surprise or arbitrary convention) and hence are poor guides to moral truths. The challenge facing advocates of TEs, then, seems to be to show that, to borrow Hallvard Lillehammer's words, "intuitions do not simply express some form of irrational prejudice". In moral and political philosophy, this is not at all a trivial concern. For one type of moral judgment, judgments of moral responsibility or, more specifically, attributions of blame which 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), seems to be particularly sensitive to these and other irrelevant features of our natural and social world. Our judgments of moral responsibility, empirical research has repeatedly shown, are hopelessly confused and incoherent. Alicke (2014) 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."

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 & 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 character assessment. This, in turn, is heavily influenced by implicit bias and prejudice.

(vi) reliance on dubious moral heuristics

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

Let me say a few words about moral heuristics and how this model is supposed to explain both the successes and failures of moral and political intuition. 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 also lead to systematic errors in untypical situations. (Sunstein 2005) 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. 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 & 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. 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 and political ones included, typically work, can we rely on them to deliver at least prima facie reliable judgments about hypothetical scenarios that moral and political philosophers devise with the aim of testing normative propositions? I'm afraid not. True, heuristics are mostly reliable. (Even Sunstein 2005 grants that.) But philosophical 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.

First of all, examples of misfiring should alert us against carelessly using proxies for target normative 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 impermissible were all carefully removed from the story. The other case at hand is our wrought responsibility judgments. Since the exact degree of the agent's responsibility is difficult enough to assess in real life cases, and is even more concealed in the often tricky philosophical 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 Pizarro & Tannenbaum (2011) are correct and responsibility judgments really are just covert character assessments – 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 the particular case under consideration, and that's why our moral heuristics regularly misfire. M&PTEs 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 so we

vary those very features – even to the point where all the plausible candidates for morally relevant features are removed from the picture. And yet in these cases the rigid heuristic 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 philosophical TEs. For in those, we try to determine the moral impact of various features and correspondingly hold some of them fixed while varying others regardless of how unlikely such disassociation may be in reality. Hence, the harmfulness of the 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 TEs where all the dependency relations are turned on their head,[9] strikes me as a rather short-sighted strategy.

(vii) mostly undetected and uncorrected (or even incorrigible) effects of bias and prejudice

None of us is completely free from bias and prejudice. We are biased towards ourselves (what Pronin 2006 calls 'the first-person bias'), our children, friends, football clubs, fellow citizens, and so on. We blame others and the misfortune for our own failures and at the same time appropriate, as our own accomplishments, things that either others did for us or helped us achieve or that we largely owe to good luck. We have a sense of justice, which, however, is often drowned out by a false sense of entitlement. (Mendelberg & Kapovitz 2007) And we normally prefer the social and political status quo over change.

There is more bad news. Evaluation, social and political including, often precedes explanation (or understanding) of both individual and collective action. Agent-evaluations are fundamental, whereas evaluations of actions and their consequences, i.e. states-of-affairs, are merely derivative and bear a strong imprint of prior agent-evaluations, particularly when the latter are negative. We are prone to apportion blame to others no matter what – even when some of the necessary conditions for moral responsibility are clearly not met. In evaluating agents, and potential collaborators, we rely on binary moral vocabulary: we classify people as either good or bad (with little room in between), and our initial assessment determines how we believe they ought, or deserve, to be treated. Furthermore, we tend to evaluate people-like-us as good/better and people-unlike-us as bad/worse – both based on the thinnest possible evidence. In matters of morals and politics, we are deeply partisan – whatever supports our joint cause, we embrace emphatically, thus displaying our loyalty to the team. No detail of this unflattering picture provides even comfort, let alone confidence in our spontaneous social and political judgment.

9 Recall Glaucon's morally inverted world (MIW).

6. What room, if any, is then left for hypothetical reasoning in moral and political philosophy?

Plenty, I'd say. By renouncing TE-evidentialism, we don't 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. (see 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 M&PTEs for diagnostic purposes – to help us identify psychological mechanisms that are operative in the formation of our intuitive judgments. (Knobe 2007) And we can keep using M&PTEs as a valuable source of *hypotheses for further testing*.[10]

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.[11] 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 the 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 views – not because a single MTE or PTE has proven any of them wrong but rather because our particular response to them gives rise to the suspicion that we may subscribe to two or more conflicting principles. The intuitions thus generated themselves would give no advice as to which of those conflicting beliefs we should abandon; it 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 Nenad calls 'rational (as opposed to historical) reconstruction' of particular social institutions, norms and practices. (Miščević 2013a) Think of John Locke and his incredibly influential attempt to provide rational grounds for the institution of private

10 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?

11 I owe this suggestion to a post by Harry Brighouse on the online forum Crooked Timber.

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 M&PTE-evidentialism unattractive. But that's a topic for another paper.

7. Conclusion

Let me conclude. In the paper, I argued against a particular use of thought-experimentation in moral and political philosophy, a view that I labelled 'TE-evidentialism'. According to this view, which Nenad endorses, M&PTEs (or, rather, intuitions that they elicit in response) are a valuable source of evidence for and against normative propositions (principles, distinctions, theories, and so on). Not sharing Nenad's optimism about the role of TEs in political philosophy, I have tried to argue that it's unfounded.

The past record of M&PTEs is far from impressive. Most, if not all, M&PTEs fail to corroborate their target hypotheses. PTEs inherit most of their shortcomings from their cousins, the MTEs. Social and political judgments are subject to a plethora of biases, which are difficult to both detect and neutralize. They are produced by heuristics with not just a fairly bad general track record, but the ones we have specific reasons to expect will systematically misfire in (M&)PTEs. As such, PTE-generated intuitions are even less trustworthy than their discredited cousins, the MTE-generated intuitions. Rather than keep relying on TEs, we should begin to explore other, more sound alternatives to thought-experimentation in moral and political philosophy.[12]

8. References

Alicke, M.D. (2014). "Evaluating Blame Hypotheses". Psychological Inquiry, 25; 187-192.

Bourget, David & Chalmers, David J. (2014). "What do Philosophers Believe?". Philosophical Studies, 170/3; 465-500.

12 Here are three such alternatives: (i) testing normative (i.e. moral and political) propositions empirically, that is, more or less directly (Prinz 2007); (ii) putting speculations about what perfectly rational individuals would've preferred for themselves, or consented to, or chosen as the principles for the regulation of social interaction, in unrealistic, idealized circumstances, to empirical test (Frohlich & Oppenheimer 1992); or (iii) consulting speculative fiction instead of increasingly bizarre TEs (de Smedt & de Cruz 2015). Given Nenad's sympathies for a response-dependence account of some other normative (Miščević 2013c) and non-normative concepts and/or properties (Miščević 2013d), my guess is that option (i) will appeal to him most.

Brown, James Robert & Fehige, Yiftach (2014). “Thought experiments”. Stanford Encyclopaedia of Philosophy, URL: https://plato.stanford.edu/entries/thought-experiment/.

Cappelen, Herman (2011). Philosophy without Intuitions. Oxford & New York: Oxford University Press.

Feinberg, Joel (1970). “The nature and value of rights”. The Journal of Value Inquiry, 4; 243-257.

Feinberg, Joel (1985). Offence to Others. The Moral Limits of Criminal Law. Vol. 2, Oxford: Oxford University Press.

Frederick, Shane (2005). “Cognitive reflection and decision making”. Journal of Economic Perspectives, 19/4; 25-42.

Frohlich & Oppenheimer (1992). Choosing Justice. An Experimental Approach to Ethical Theory. Berkeley: University of California Press.

Gendler Szabo, Tamar (2007). “Philosophical thought-experiments, intuitions and cognitive equilibrium”. Midwest Studies in Philosophy, XXXI; 68-89.

Goodin, Robert (2008). Innovating Democracy. Oxford University Press.

Haidt, Jonathan (2001). “The emotional dog and its rational tail”. Psychological Review, Vol. 108, No. 4; 814-34.

Hardin, Garrett (1968). “The tragedy of the commons”. Science, Vol. 162, No. 3859; 1243-1248.

Hobbes, Thomas (1998). Leviathan. Ed. by J.C.A. Gaskin, Oxford & New York: Oxford University Press.

Knobe, Joshua (2007). “Experimental philosophy and philosophical significance”. Philosophical Explorations, 10 (2); 119 – 121.

Locke, John (1980). Second Treatise of Government. Ed. by C.B. Macpherson, Indianapolis, CA: Hackett Publishing House.

McMahan, Jeff (2008). “Torture in Principle and in Practice”. *Public Affairs Quarterly*, 22/2; 91-108.

Mendelberg, Tali & Kapovitz, Christopher (2007). “How people deliberate about justice: groups, gender and decision rules”. In: Rosenberg, Shawn W. (Ed.). Deliberation, Participation, and Democracy. Can the People Govern?. New York: Palgrave Macmilan; 101-129.

Miščević, Nenad (2015). “Intuitions: reflective justification, holism and apriority”. *Croatian Journal of Philosophy*, Vol. XV, No. 45; 307-323.

Miščević, Nenad (2013a). “In search of the reason and the right: Rousseau's social contract as a thought experiment”. *Acta Analytica*, vol. 28, no. 4; 509-526.

Miščević, Nenad (2013b). “Political thought-experiments from Plato to Rawls”. In: Frappier, Mélanie, Meynell, Letitia, Brown, James Robert (eds.). *Thought Experiments in Philosophy, Science, and the Arts*. New York & London: Routledge; 191-206.

Miščević, Nenad (2013c). “The colors of life : Boran Berčić on the meaning of life”. *Prolegomena*, Vol. 12, Iss. 2; 199-221.

Miščević, Nenad (2013d). “The ontology of secondary and tertiary qualities”. *Balkan Journal of Philosophy*, Vol. 5, Iss. 1; 45-58.

Miščević, Nenad (2007). “Modelling intuitions and thought-experiments”. *Croatian Journal of Philosophy*, Vol. 7, No. 20; 181-214

Miščević, Nenad (2004). “The explainability of intuitions”. *Dialectica*, Vol. 58, No. 1; 43-70.

Nussbaum, Martha (1997). “If Oxfam ran the world”. London Review of Books, Vol. 19, No. 17; 18-19.

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

Plato (1993). The Republic. Transl. by Robin Waterfield, Oxford & New York: Oxford University Press.

Prinz, Jesse (2007). “Can moral obligations be empirically discovered?”. Midwest Studies of Philosophy, 31; 271-291.

Pronin, Emily (2006). “Perception and misperception of bias in human judgment”. TRENDS in Cognitive Sciences, Vol. 11, No. 1; 37-43.

Rawls, John (1971/1999). A Theory of Justice. Revised Edition, Oxford: Oxford University Press.

Schwitzgebel, E., & Cushman, F. (2015). “Philosophers’ biased judgments persist despite training, expertise and reflection”. *Cognition*, 141, 127-137.

Singer, Peter (1993). Practical Ethics. 2nd Ed., Cambridge: Cambridge University Press.

De Smedt, Johan & De Cruz, Helen (2015). "The epistemic value of speculative fiction". *Midwest Studies in Philosophy*, vol. 39; 58-77.

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

Sunstein, Cass (2005). "Moral heuristics". Behavioral and Brain Sciences, Vol. 28; 531-73.

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