

CROATIAN JOURNAL OF PHILOSOPHY

On Thought Experiments

MARGHERITA ARCANGELI

MAJDA TROBOK

RAWAD EL SKAF

FRANÇOIS KAMMERER

ERHAN DEMIRCIOĞLU

DANIEL DOHRN

MIOMIR MATULović

HOSSEIN DABBAGH

FRIDERIK KLAMPFER

NENAD MIŠČEVIĆ

Book Discussions

ANA BUTKOVIĆ

DANILO ŠUSTER

Book Reviews

MIA BITURAJAC

TOMISLAV MILETIĆ

NENAD MIŠČEVIĆ

DAVOR PEĆNJAK

MARKO DELIĆ

Croatian Journal of Philosophy

1333-1108 (Print)

1847-6139 (Online)

Editor:

Nenad Mišcević (University of Maribor)

Advisory Editor:

Dunja Jutronić (University of Maribor)

Managing Editor:

Tvrtko Jolić (Institute of Philosophy, Zagreb)

Editorial board:

Stipe Kutleša (Institute of Philosophy, Zagreb),

Davor Pećnjak (Institute of Philosophy, Zagreb)

Joško Žanić (University of Zadar)

Advisory Board:

Elvio Baccarini (University of Rijeka), Carla Bagnoli (University

of Modena), Boran Berčić (University of Rijeka), István M. Bod-

nár (Central European University), Vanda Božičević (Bergen

Community College), Sergio Cremaschi (Milano), Michael Devitt

(The City University of New York), Peter Gärdenfors (Lund

University), János Kis (Central European University), Friderik

Klampfer (University of Maribor), Željko Loparić (Sao Paolo),

Miomir Matulović (University of Rijeka), Snježana Prijic-

Samaržija (University of Rijeka), Igor Primorac (Melbourne),

Howard Robinson (Central European University), Nenad

Smokrović (University of Rijeka), Danilo Šuster (University

of Maribor)

Co-published by

“Kruzak d.o.o.”

Naserov trg 6, 10020 Zagreb, Croatia

fax: + 385 1 65 90 416, e-mail: kruzak@kruzak.hr

www.kruzak.hr

and

Institute of Philosophy

Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia

fax: + 385 1 61 50 338, e-mail: filozof@ifzg.hr

www.ifzg.hr

Available online at <http://www.ceeol.com> and www.pdcnet.org

CROATIAN
JOURNAL
OF PHILOSOPHY

Vol. XVIII · No. 52 · 2018

On Thought Experiments

Introduction NENAD MIŠČEVIĆ	1
The Hidden Links between Real, Thought and Numerical Experiments MARGHERITA ARCANGELI	3
The Mathematics-Natural Sciences Analogy and the Underlying Logic. The Road through Thought Experiments and Related Methods MAJDA TROBOK	23
The Function and Limit of Galileo's Falling Bodies Thought Experiment: Absolute Weight, Specific Weight and the Medium's Resistance RAWAD EL SKAF	37
Is the Antipathetic Fallacy Responsible for the Intuition that Consciousness is Distinct from the Physical? FRANÇOIS KAMMERER	59
On Understanding a Theory on Conscious Experiences ERHAN DEMIRCIOĞLU	75
'Mais la fantaisie est-elle un privilège des seuls poètes?' Schlick on a 'Sinnkriterium' for Thought Experiments DANIEL DOHRN	87
Thought Experiments in the Theory of Law: The Imaginary Scenarios in Hart's <i>The Concept of Law</i> MIOMIR MATULOVIĆ	101

Intuiting Intuition: The Seeming Account of Moral Intuition HOSSEIN DABBAGH	117
Moral Thought-Experiments, Intuitions, and Heuristics FRIDERIK KLAMPFER	133
Simulation and Thought Experiments. The Example of Contractualism NENAD MIŠČEVIĆ	161

Book Discussions

The ‘Arguments Instead of Intuitions’ Account of Thought Experiments: Discussion of <i>The Myth of the Intuitive</i> by Max Deutsch ANA BUTKOVIĆ	191
“The Brain in Vat” at the Intersection DANILO ŠUSTER	205

Book Reviews

Michael Stuart, Yiftach Fehige and James Robert Brown (eds.), <i>The Routledge Companion to Thought Experiments</i> MIA BITURAJAC	219
Harris Wiseman, <i>The Myth of the Moral Brain.</i> <i>The Limits of Moral Enhancement</i> TOMISLAV MILETIĆ	230
Amy Kind and Peter Kung (eds.), <i>Knowledge Through Imagination</i> NENAD MIŠČEVIĆ	237
Bojan Borstner and Smiljana Gartner (eds.), <i>Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević</i> DAVOR PEĆNJAK	241
Boran Berčić (ed.), <i>Perspectives on the Self</i> MARKO DELIĆ	244

On Thought Experiments

Introduction

Thought experiments are a hot issue in methodology of philosophy. They have been a topic of constant interest at both Rijeka and Maribor philosophy departments (see Davor Pečnjak's review in this issues). This issue brings together some of the results, and also some papers from the conference on Simulation and thought experiment, held in Geneva in June 2017, and some papers that were spontaneously submitted by their authors.

The work on the issue was financially supported by the Slovenian Research Agency: research core funding No. P6-0144, The Thought Experiments from Nature to Society.

NENAD MIŠČEVIĆ

The Hidden Links between Real, Thought and Numerical Experiments

MARGHERITA ARCANGELI*
Humboldt-Universität zu Berlin, Germany

The scientist's toolkit counts at least three practices: real, thought and numerical experiments. Although a deep investigation of the relationships between these types of experiments should shed light on the nature of scientific enquiry, I argue that it has been compromised by at least four factors: (i) a bias for the epistemological superiority of real experiments; (ii) an almost exclusive focus on the links between either thought or numerical experiments, and real experiments; (iii) a tendency to try and reduce one kind to another; and (iv) an excessive attention to the outputs of these types of experiments, more than to their processes. In this paper I support an unbiased triangular comparative analysis that focuses on the processes involved in real, thought and numerical experiments, and claim that all three types of experimentation are fundamental to scientific research. I do so by clarifying different notions of experimental processes, and by introducing a distinction between two varieties of mental simulation that play a role in them (i.e., mental models and imaginings). I then compare real, thought and numerical experiments in light of this distinction, showing their similarities, but also fundamental differences, which suggest that none of them is dispensable.

Keywords: Real experiments, thought experiments, computer simulations, mental models, imagination.

Introduction

The paradigmatic scientific practice is to conduct, usually in a laboratory setting, experiments by using equipment made up of measurement

* I am grateful to Jérôme Dokic for helpful comments on earlier versions of this paper. I am also indebted to audiences at the SPS 2016 in Lausanne and the Geneva "Simulation and Thought Experiment" Conference (2017) for their valuable feedbacks. I was supported by an Alexander von Humboldt Foundation Fellowship during my work on this paper.

devices, substances, etc. Albert Michelson and Edward Morley, for instance, used an interferometric setup, mounted on a cast stone floating in a trough of liquid mercury, aiming to detect aether wind effects on the speed of light. These attempts to capture the real world in a laboratory setting have been called “ordinary” or “real” experiments. This terminological clarification is crucial when comparing real experimentation with another scientific practice, namely thought experimentation.

Thought experiments are not conducted in laboratories, but rather in the “laboratory of the mind” (Brown 1991a; 1991b). Erwin Schrödinger, for example, did not pen a cat in a steel chamber with a device comprising a radioactivity detector, a hammer and a poison flask. He merely imagined, and asks us to imagine, a situation showing that an unclear divide between the quantum or microscopic level and the macroscopic level leads to paradoxical cases in which macroscopic objects like cats are at the same time dead and alive.¹

The scientist’s toolkit counts another scientific practice, namely computer simulations. We should distinguish between two senses of the term “computer simulation”. In a narrow sense a computer simulation is a computer program which models and explores via algorithmic procedures the dynamics of a system. “More broadly, we can think of computer simulation as a comprehensive method for studying systems. In this broader sense of the term, it refers to an entire process. This process includes *choosing a model*; finding a way of *implementing* that model in a form that can be run on a computer; *calculating the output* of the algorithm; and *visualizing* and *studying the resultant data*” (Winsberg 2015, italics added). Real and thought experiments can also be described as procedures consisting of similar steps. Thus, computer simulations, broadly understood, can be labelled “numerical experiments”.

How are real, thought and numerical experiments related? Rawad El Skaf and Cyrille Imbert nicely capture the state of the art on the matter in the following quote:

Overall, while studies comparing [NE], TE and [R]E have kept developing, the literature offers more a battlefield than a steadily developing domain; (...). More annoyingly, many of these disagreeing accounts heavily lean onto incompatible conceptions of these activities, which does not help disentangling these issues. (El Skaf and Imbert 2013: 3453)

Although getting clear about the relationship between real, thought and numerical experiments seems to be an important issue in the debate, a deep comparative analysis of these three scientific tools has

¹ The Copenhagen interpretation of Quantum Mechanics postulates that a physical system can be in a very special state which is a simultaneous superimposition of different states. Any observation or measurement causes a collapse of the physical system into one of the superimposed states. This physical phenomenon occurs only at the quantum level, but we can imagine cases in which the superimposition of a microscopic state is causally tied to the superimposition of a macroscopic state: imagine a particle in a superimposed state $A + B$ and a device that kills a cat in a steel chamber, if the particle is in A , and, does nothing, if the particle is in B .

been compromised by a bias for the epistemological superiority of real experiments. The aim of this paper is to deepen our understanding of the relationships between real, thought and numerical experiments by fostering an unbiased triangular comparative analysis.

In §1 I shall offer a diagnosis of the poor situation described by El Skaf and Imbert. I shall identify four sources of limitation for the debate on real, thought and numerical experiments: (i) the aforementioned bias for the superiority of real experiments; (ii) an almost exclusive focus on the links between either thought or numerical experiments, and real experiments; (iii) a tendency to be concerned mainly by identity issues (e.g., Are numerical experiments real experiments?); and (iv) an excessive attention to the outputs of these scientific tools, more than to their processes. Thus, my proposal is that the best treatment is to pursue an unbiased triangular comparative analysis focusing on the processes involved in real, thought and numerical experiments.

I shall set the stage for such an analysis by clarifying two points. First, in order to get an unbiased triangular comparison we need to put real, thought and numerical experiments on an equal footing. I shall argue that this can be done by acknowledging that these scientific practices share a common functional structure. This allows to study them as if they were all experiments, regardless of ontological and/or epistemological differences. Second, I shall clarify different processes and dimensions along which a triangular comparative analysis can be made. More precisely, I shall distinguish between two levels of procedure: (a) a level of *production*, in which an experiment is elaborated and executed, and (b) a level of *presentation*, in which the results of an experiment are related to a scientific hypothesis and published. I shall, then, proceed to compare real, thought and numerical experiments along these two dimensions (in §2 and §3, respectively).

My analysis shall pivot on the role played by mental simulation in real, thought and numerical experimentation. Indeed, mental simulation has been identified as a key process that would show that between thought and numerical experiments there is an essential similarity, which paves the way to a replacement of the former by the latter. More should be said about the relevant notion of mental simulation, however. The literature in philosophy of mind recognises at least two different types of mental simulation, namely mental models and imaginings. As far as the production dimension is concerned, my claim is that in the performance of thought experiments imaginings are more important than mental models. Moreover, thought experiments call for mental simulation, in both its guises, in a way that neither numerical, nor real experiments do, although their elaboration can involve mental simulations. With respect to the presentation dimension, I shall consider the narrative aspect common to real, thought and numerical experiments and how it connects to the pivotal role of imagination in thought experiments. The upshot of the overall analysis will be that all three kinds of experimentation are fundamental to scientific research, in opposition

to the provocative view that numerical experimentation will replace thought experimentation.

1. *The State of the Art*

1.1. *Diagnosis*

A triangular comparison between real, thought and numerical experiments can be mutually illuminating. As stressed by El Skaf and Imbert, however, the current debate on these scientific tools seems to be stuck. Such a situation could be unblocked by diagnosing what is limiting the analysis. I believe that there are at least four factors at stake.

Firstly, I contend that quite often real, thought and numerical experiments are not put on the same level, although this is a preliminary step to a genuine triangular comparison. One way to see whether these scientific tools are equally treated is to ask: How experimental are thought and numerical experiments? Let us start with thought experiments.

This is a much-discussed question in the literature on thought experimentation, which has received answers that range from the denial of the experimental character of thought experiments (e.g., Humphreys 1993; Norton 2004), to claims that thought experiments are a kind of experimentation (e.g., Sorensen 1992; Buzzoni 2004—for a detailed review of both positions, see Arcangeli 2017). The problem is that a bias for the intrinsic epistemological superiority of real experiments very often guides analyses of the experimental character of thought experiments. For instance, calling thought experiments “imaginary experiments” can reveal such a bias: just as imaginary friends are not genuine friends, thought experiments would not be genuine experiments, but mere imaginary visualisations of experiments. If we comply with the aforementioned bias, we run the risk of paying attention mainly (if not only) to the features proper to real experiments that thought experiments lack. For instance, a typical plea for real experiments would stress that thought experiments are less reliable and lack justificatory power, since they do not directly examine nature. Reasoning along this line leads to consider thought experimentation a dispensable tool: thought experiments are useful only in preparation for real experiments or when the latter are not available.²

It is interesting to notice that if we replace the expression “thought experiment” with “numerical experiment” in the above, we can retrieve an analogous set of considerations which actually features in the literature on computer simulations. Asking whether numerical experiments have an experimental character has generated a heated debate, which is still open. For instance, some (e.g., Gilbert and Troitzsch 1999; Beisbart and Norton 2012) argue that real experiments and numerical experiments could not possibly differ from each other more; while others (e.g., Dowling 1999; Barberousse et al. 2009) regard numerical

² A good illustration of a plea for real experiments can be found in Hull 1997 (see also Hull 1989). Many of the sceptical worries raised by David Hull against thought experiments can also be found in Paul Thagard’s work (see Thagard 2010 and 2014).

experiments as a genuine experimental practice. Moreover, it has been pointed out that the analysis of the relationship between numerical and real experiments is often influenced by the widespread bias about the intrinsic epistemological superiority of real experiments. Eric Winsberg writes:

Whether we want to contrast simulations with “experiments” or with “ordinary experiments” (...) seems to be to [sic] an issue of whether or not to award them an honorific title. And that motivation (...) is grounded in the misguided intuition that “experiments” are intrinsically epistemologically superior. (Winsberg 2009: 583, fn 11—see also Parker 2009)

Very often both thought and numerical experiments have not been evaluated *per se*: they have been judged from the standards of real experiments, rather than on the basis of a broad analysis treating all types of experimentation on a par. The upshot of this line of reasoning is to rank thought and numerical experimentation below real experimentation: in the train of scientific practice, the latter is the first class and the former just the tourist car, so to speak. This view seems also to suggest that the three tools are incompatible: real experimentation is always to be preferred and, when it is available, both thought and numerical experimentation become less useful, if not useless. The debate should avoid relying on such a portrait of the relationship between real, thought and numerical experiments, which, I think, is doomed to underestimate the similarities between them and overlook their specificities.³

A second source of limitation for the debate is to have privileged a binary analysis. Although some authors have commented in passing on the parallelism between thought and numerical experimentation (e.g., Sorensen 1992; Nersessian 1993; Stöltzner 2003; Buzzoni 2004; Cooper 2005), and others have suggested that numerical experiments can be seen as a type of thought experiments (e.g., Di Paolo *et al.* 2000; Swan 2000) and will even replace the latter (Chandrasekharan *et al.* 2013), the “trading zone” between thought experiments and numerical experiments has been sparsely considered (e.g., Staüdner 1998; Velasco 2002; El Skaf and Imbert 2013; Lenhard 2011 and 2018). Most works have primarily focused their attention on how either thought or numerical experiments relate to real experiments. This seems also due to a greater importance given to real experiments.

In a thought-provoking paper, Sanjay Chandrasekharan, Nancy Nersessian and Vrishali Subramanian highlight two further limitations of the current debate. They are concerned with the comparison between numerical and thought experiments only, but what they say can be extended to real experiments as well. Thus, their diagnosis can be added to mine.

³ I agree with Marco Buzzoni when he observes that “the one-sidedness of the attempt to establish the concept of experiment without a simultaneous clarification of the concept of thought experiment is not less serious than the one-sidedness of fixing the attention on the latter, by taking for granted the meaning of the former” (Buzzoni 2004: 10—mine translation). The same holds, I think, for numerical experiments.

Thirdly, the leading issues in the debate concern “identity relations” between the three scientific tools (Chandrasekharan *et al.* 2013: 242). What Chandrasekharan and colleagues wish to emphasise is the excessive attention paid to questions such as: Are numerical experiments a kind of thought experiments? Are numerical/thought experiments a kind of real experiments? The idea is that the analysis is mainly taxonomical and aimed at identifying necessary and sufficient conditions, which can establish what belongs to the categories “real experiment”, “thought experiment” and “numerical experiment”. Chandrasekharan and colleagues point out that taxonomical analysis is not negative in itself, but we should avoid studying real, thought and numerical experiments exclusively from this point of view.⁴

Moreover, they claim that existing taxonomical comparisons analyse (real, thought and numerical) experiments with respect to their outputs, more than the processes underlying them. A “process-oriented analysis” would, however, be also worth pursuing, because we could consider not only the result of an experiment, but also *how* we get it (cf. Arcangeli 2010 and 2017). This connects to the other worry raised by Chandrasekharan and colleagues about the current debate. A fourth factor limiting comparisons of real, thought and numerical experiments is precisely such lack of attention for the processes yielding results.

Now that the current debate has been diagnosed with these four problems, we need a treatment.

1.2. *Treatment*

The most promising treatment seems to be to favour a process-oriented comparison, which analyses real, thought and numerical experiments from an unbiased perspective. Two clarifications are in order: to make sure that these scientific tools are on the same level as experiments, and to specify the relevant processes they involve.

A good way to establish the equality of experiments is precisely to start from their definition as procedures involving different steps. Recall Winsberg’s words about numerical experiments quoted in the Introduction. A numerical experiment can be defined as a process involving the following steps: choosing a model, finding a way of running it on a computer, calculating the output, visualising and studying the data. We can offer definitions for real and thought experiments describing them as procedures including steps functionally similar to those highlighted for numerical experiments. The first detailed description of both real and thought experiments in these terms is due to Ernst Mach.

Mach (1896) claimed that both real and thought experimentation are based on the “method of variation”. The method of variation can be seen as a four-step procedure describing what in general an experimenter has to do:

⁴ This point also connects to my first point: we should avoid taxonomical analysis benchmarking thought and numerical experimentation against real experimentation.

- (i) selecting and isolating the features which act as variables;
- (ii) “manipulating” these variables, i.e., making them interact;
- (iii) observing what consequently happens;
- (iv) interpreting the results in the light of a theory.

Many philosophers have focused their attention on at least one of these steps and recognised the similarity between thought and real experimentation (e.g., Gooding 1990 and 1993; Humphreys 1993; Nersessian 1993; Häggqvist 1996; Bishop 1998; Wilkes 1988; Reiner and Gilbert 2000; Brendel 2004; Buzzoni 2004 and 2008; Lenhard 2018). Mach’s analysis can be applied to numerical experiments too. It is easy to see that Winsberg’s description of numerical experiments fits quite well the given description of the experimental practice. While choosing a model and finding a way of running it on a computer belong to (i), calculating the output, visualising and studying the data roughly coincide with (ii), (iii) and (iv), respectively. Nevertheless, the fact that the method of variation is common to experimental practice *in toto*, numerical experimentation included, has rarely been pointed out (see Stäudner 1998). El Skaf and Imbert seem to have this point in mind in their analysis of real, thought and numerical experiments.

El Skaf and Imbert claim that real, thought and numerical experiment are “composed of functionally similar parts” (El Skaf and Imbert 2013: 3455). They call the set of these parts a “CUI pattern of inquiry” where C stands for “construction of a scenario in the context of an inquiry”, U for “unfolding of the scenario” and I for “interpretation of the result”.⁵ This description also matches the four steps I suggested following Mach: C-step roughly is (i), U-step summarises (ii) and (iii), and I-step coincides with (iv). This functional structure similar to real, thought and numerical experiment is enough, according to El Skaf and Imbert, “to provide a framework to analyze them together and draw non trivial consequences about their use in science” (El Skaf and Imbert 2013: 3455). I agree with them: thinking of real, thought and numerical experiments as functioning in a similar way is an excellent step to establish the equality of the experiments.

The equality that I am pleading for should not be read too strongly. Drawing parallelisms between real, thought and numerical experimentations does not necessarily commit to claim that these practices belong to the same natural kind. The methodological point is that it is worthwhile to study them as if they were all (structurally) experiments, even if one believes that they are not (on this point see also Sorensen 1992 and Bishop 1998). A triangular comparison made on this basis is mutually illuminating, despite ontological and/or epistemological differences.⁶

⁵ More specifically, they list five parts: question-oriented activities, scenarios, unfolding of scenarios, result(s) of scenarios and scientific conclusion(s).

⁶ These differences start to emerge once we look more closely at the specific issues raised by each of the four steps (see Arcangeli 2017, as far as real and thought experiments are concerned). The first step, for instance, raises the question about

I suggested that the debate should pay more attention to the processes involved in real, thought and numerical experimentation. Emphasising the functional structure common to these practices does help to identify relevant processes and dimensions, upon which a triangular comparative analysis should focus. The performance of an experiment occurs at two levels: (a) a level of *production* and (b) a level of *presentation*. While (a) involves elaborating and executing the experiment—i.e., steps (i) to (iii), (b) concerns the interpretation phase in which the results of an experiment are related to a scientific hypothesis and published—i.e., step (iv).

These two dimensions are roughly equivalent to the dimensions stressed by Chandrasekharan and colleagues in their comparison between thought and numerical experiments. Indeed, they distinguish between two types of processes, namely building and interpretation processes. With respect to thought experiments they say:

When a thought experiment is presented, what we get is the final, polished product, usually a narrative, which is the end-point of a long building process. (...) The interpretation phase involves relating the final results of the TE to a theory or phenomena (Chandrasekharan *et al.* 2013: 241).

The authors maintain that the same distinction holds for numerical experiments. In their analysis, however, they seem to ignore the difference between merely elaboration processes and execution processes. Arguably, the “building process” is not a unitary process: it is one thing to design an experiment, and another to carry it out. In what follows I will take building processes to refer to the exploratory phase (cf. Lenhard 2018) in which experimenters design and develop the experiment, and distinguish them from execution processes, in which experimenters “concretely” perform the experiment. This distinction will prove to be extremely important in order to see to what extent thought experiments differ from both real and numerical experiments.

In the following sections I will compare real, thought and numerical experiments along both the production and presentation dimension. Actually, I will be more concerned with the production dimension. This choice answers Chandrasekharan and colleagues’ complaint about an excessive focus “on interpretation, rather than building” (Chandrasekharan *et al.* 2013: 241). In general, my analysis will be driven by some of their hypotheses with the aim of undermining their view that thought experiments will be replaced by numerical experiments.

the kind of variables that are involved in the different practices. Answers to this question seem to be hostage to the held view about the nature of real, thought and numerical experimentation. On one view, only real experiments examine directly nature, whereas both thought and numerical experiments merely explore theoretical models (e.g., Gilbert and Troitzsch 1999; Humphreys 1993). On an alternative view, it is simplistic to say that real experiments always examine directly nature: very often in any kind of experiment we learn something about a target, dealing with a (concrete or abstract) model that is intended to stand for it (see Winsberg 2009, though he is only concerned with real and numerical experiments, and El Skaf and Imbert 2013).

2. *The Production Dimension*

As suggested beforehand, the production dimension involves different processes which can be attributed to two different phases: the elaboration phase and the execution phase.

These two phases are clearly identifiable as far as real and numerical experimentation are concerned. Take two astrophysicists, Sam and Maria, interested in supersonic gas jets that are formed when gasses flow into the gravity well of a black hole. Sam starts thinking how to study this phenomenon by using a fluid-dynamical setup, consisting of gas bubbles, a tank of fluid, simple cylindrical and spherical objects and a physical mechanism for causing shock waves and propagating them through the tank. By contrast, Maria wants to create a software based on the theory of fluid dynamics that, once run on a digital computer, can simulate the relevant flow dynamics.⁷ Sam and Maria employ different strategies (real and numerical experimentation, respectively), but they are both going through a first preliminary phase in which they are elaborating their experiments, before carrying them out. The genuine execution of their experiments comes at a second stage, after a careful planning, setting up and, probably, a series of trials.

A first difference between thought experimentation, on the one hand, and both real and numerical experimentation, on the other hand, emerges. In the case of thought experiments it seems that we cannot easily distinguish between elaboration and execution processes. Plausibly a thought experimenter has to consider different scenarios before finding the right one, but, when she does so, it seems that she is also performing a thought experiment. Elaborating and executing a thought experiment are so much intertwined processes that they can even be thought as two sides of a single process.⁸

No matter whether there is only one process, another distinctive aspect of thought experimentation is that the production dimension is not publicly accessible. More precisely, the subjective process of production of a thought experiment seems to be accessible only via the introspection of the thought experiment's author (Nersessian 1993; Chandrasekharan *et al.* 2013). For example, we might have asked Henri Poincaré to tell us how he conceived his famous thought experiment of a sphere-world whose inhabitants believe to live an infinite world.⁹ This way of gaining

⁷ The example has been suggested by an example in Winsberg 2009.

⁸ Buzzoni (2004), for instance, maintains that thought experiments cannot be designed.

⁹ In *Science and Hypothesis*, Poincaré writes: "Suppose, for example, a world enclosed in a large sphere and subject to the following laws:—The temperature is not uniform; it is greatest at the centre, and gradually decreases as we move towards the circumference of the sphere, where it is absolute zero. (...) — If R be the radius of the sphere, and r the distance of the point considered from the centre, the absolute temperature will be proportional to $R^2 - r^2$. (...) all bodies have the same coefficient of dilatation, so that the linear dilatation of any body is proportional to its absolute temperature. Finally, I shall assume that a body transported from one point to another of different temperature is put instantaneously in thermal equilibrium with

information about the production of a thought experiment contrasts with the method we might use to access both the elaboration and the execution phases of real and numerical experiments. As far as the elaboration phase is concerned, we can also ask experimenters how they came up with their experimental procedure and they can partially rely on introspection in their answers, but other information is publicly available (e.g., reports, calculations, preliminary data, software, laboratory equipment). Moreover, as a rule the execution phase is publicly accessible to facilitate the reproducibility of real and numerical experiments.¹⁰

Despite these differences Chandrasekharan and colleagues suggest a common key to the production dimension of both thought and numerical experiments, namely mental simulation. They think of these practices as tests for counterfactual situations which are difficult to implement in real settings. Mental simulation is the underlying mechanism that makes such explorations possible. Chandrasekharan and colleagues go further and claim that numerical experimentation extends our capacity for mentally simulating in a way similar to how telescopes have extended our visual capacities. The point is that numerical experiments support more sophisticated mental simulations, which can deal with complex phenomena showing non-linear dynamic. For this reason, according to the authors, numerical experimentation is bound to replace thought experimentation. Chandrasekharan and colleagues' view is open to criticism, once we acknowledge that "mental simulation" is not a univocal notion.

In philosophy of mind the notion of mental simulation can refer at least to two types of mental simulation. This distinction is clearly described by Alvin Goldman:

[W]hen I speak of "mental simulation," I shall mean a replication or duplication of another mental state. A mental simulation is a simulation of a mental state by a mental state. By contrast, the mental models approach regards mental simulations as simulations of (...) external or physical states of affairs. (Goldman 2006: 51, fn 10)

In other words, on the one hand, we would have what we might call "objectual" mental simulation, which mentally simulates states of af-

its new environment" (Poincaré 1902/1905: 65). What is the geometry experienced by the inhabitants of this hypothetical world? Assuming, for simplicity of calculation, that the world is a disc, rather than a sphere, and that chartered surveyors are measuring a circle, whose centre coincides with that of their world and whose circumference lies at the distance in which bodies undergo a halving of their maximum length. They will calculate a circumference/diameter ratio not equal to π , as predicted by Euclidean geometry, but $>\pi$, in line with the predictions of hyperbolic geometry. Moreover, if the inhabitants wanted to try to understand whether their world is finite, it is clear that they would have no doubt about its infinity. In fact, their measuring instruments would tend to shorten in proportion to their moving towards the edge of the sphere.

¹⁰ I am not suggesting a purely internalist conception of the production of thought experiments. In devising a thought experiment one can use external supports (e.g., sketches, diagrams, notes). However, they are not sufficient to give exhaustive access to the production dimension.

fairs by encoding spatial configurations and other manifest properties (e.g., shapes and colours). On the other hand, we would have “mental” mental simulation, which mentally simulates mental states. To give an example, an objectual mental simulation of a cat captures most of its manifest properties, whereas a mental mental simulation of a cat is phenomenologically and/or functionally similar to an experience of a cat (e.g., a visual experience).

As suggested by Goldman himself, the distinction between these two types of mental simulation can be traced back to the distinction between two different capacities, namely mental modelling and imagination. While mental models mentally simulate in the objectual sense, imaginings mentally simulate in the mental sense.¹¹ Let me say something more about both mental models and imagination.

The most influential account of mental models has been proposed by Philip Johnson-Laird (1983, 2004). According to him, a mental model represents a real-world or imaginary situation and can be seen as a structure stored in short or long-term memory. He defines mental models as a third type of mental representation, half way between propositional and pictorial mental representations. It means that there is always a structural analogy between mental models and what they represent, but not all mental models can be visualised.

An account of imagination in terms of mental simulation has been strongly advocated by Gregory Currie and Ian Ravenscroft.¹² In their view imagination is defined as a complex activity that simulates non-imaginative mental states, producing mental states phenomenologically and/or functionally similar to their non-imaginative, simulated “counterparts”. When imagination simulates perception, a similarity at the psychological level is postulated between the subject who imagines and the one who genuinely perceives. If, for example, Emma imagines seeing a cat, her mental state is phenomenologically very similar to the visual perception of a cat, she feels like seeing a cat. Moreover, in her mental economy her imagining can play roles similar to those played by a percept (e.g., Emma can imagine interacting with the cat and think whether to adopt one). This is not to underestimate the differences between imaginings and their counterparts: plausibly there is also a phenomenological difference

¹¹ I have already suggested such a distinction (Arcangeli 2010, 2013, 2018), although with sometimes a different terminology (in Arcangeli 2010 I speak of iconic versus recreative imagination). See also: Abell and Currie 1999; Currie and Ravenscroft 2002 (41–43; where they claim that imagination would not be involved in thought experimentation—see Arcangeli 2010 for a critique) and Zeimbekis 2011. John Zeimbekis suggests that both types of mental simulation are involved in thought experiments, but that mental simulation is a source of epistemic bias, at least for thought experiments in the moral domain. It is not clear, however, whether Zeimbekis is really concerned with the concept of mental simulation as imagination (Arcangeli 2017).

¹² Goldman (2006) is another advocate of this view, which, in fact, can be seen as a common ground between very different approaches to the imagination (see Arcangeli 2013).

between Emma's imagining and a percept with the same content, and she can always imagine a cat in bed even if there is none, but she can hardly perceive one—unless she is hallucinating or confusing a pillow with a cat. Beyond perception, whatever type of non-imaginative mental state is simulated by imagination, it should bear phenomenological and functional similarities to its imaginative homologue.¹³ These phenomenological and functional similarities are grasped by the notion of imagination (i.e., mental mental simulation), but not by the notion of mental model (i.e., objectual mental simulation).¹⁴

With the distinction between imaginings as mental mental simulations and mental models as objectual mental simulations, we can go back to Chandrasekharan and colleagues' suggestion that mental simulation is a crucial mechanism in the production of both thought and numerical experiments, and ask: Which type of mental simulation is crucial?

Chandrasekharan and colleagues are only concerned with mental models. They do not distinguish the two types of mental simulation and take mental simulation to coincide with mental modelling. This is clear in their drawing on the simulation-based account of thought experimentation developed by Nancy Nersessian (one of the co-authors).¹⁵ According to her, in thought experiments we reason through manipulating a mental model, that is, a "structural analogue of the situation described" in the

¹³ Determining how many types of imagination there are is a very controversial issue in the literature. According to many, imagination is not reduced to the simulation of perception and can simulate other non-imaginative mental states, such as beliefs (e.g., Currie and Ravenscroft 2002), desires (e.g., Doggett and Egan 2007), and emotions (e.g., Goldman 2006). In collaborative work (Dokic and Arcangeli 2015a), I take as a plausible starting point the idea that imagination simulates, at least, perception and belief and put forward a taxonomy neutral with respect to the exact numbers of the varieties of imagination.

¹⁴ One might object that at least sensory imagination (i.e., the type of imagination similar to perception) can be reduced to mental models. The idea is that mental models, by encoding spatial configurations and other manifest properties, also encode information concerning the viewpoint from which a given situation is taken, which would exhaust what sensory imagination conveys. However, it seems that we cannot reduce sensory imagination, like perception, to mere perspectival information. Sensory imaginings, like percepts, convey self-relative information: they specify the type of self who "occupies" the given viewpoint. For example, Emma can imagine seeing the Panthéon from the other end of rue Soufflot from her point of view, but also from the point of view of her twin sister and even, going further beyond her perceptual capacities, from a virtual or counterfactual first-person perspective—"in the sense that she is imagining a situation from a spatial perspective that a normally-sighted subject *would* have if she were suitably oriented in the imaginary world" (Dokic and Arcangeli 2015b: 4). There are many ways of visually imagining one and the same spatial perspective, depending on what self is at stake, and the relevant self-relative information cannot be reduced to information about manifest properties. Thus, sensory imagination cannot be reduced to mental modelling.

¹⁵ Nenad Mišević (1992 and 2007) has also developed an approach to thought experimentation based on the notion of mental model. For a more in-depth analysis of the positions held by Mišević and Nersessian, as well as model-based views relying on other notions of model, see Häggqvist 1996, Arcangeli 2010 and 2017.

thought experimental narrative (Nersessian 1993: 297; see also Nersessian 2007). Chandrasekharan and colleagues extend this approach to numerical experiments. Elsewhere (Arcangeli 2010 and 2018) I have argued that Nersessian overlooks the complexity of the notion of mental simulation and, in focusing on mental models only, she fails to recognise the key role played by imagination in thought experimentation.

Thus, what about imagination in the production of thought and numerical experiments, as well as in the production of real experiments? I will start with thought experiments.

Imagination is crucial, and more important than mental models, for thought experimentation (Arcangeli 2010 and 2018). The reason is to be found in the fact that imagination captures something that mental modelling does not. Recall that mental simulation is better suited than objectual mental simulation to capture phenomenological and functional aspects of experiences. Imagination (viz. mental simulation) enables us “to project ourselves into another situation and to *see*, or *think* about, the world from another perspective” (Currie and Ravenscroft 2002: 11—italics added). Thought experimenters need precisely this: to be put in the position to perceive and believe from perspectives that are not directly present to their senses. Thought experimenters are like observers, like real experimenters. Thus, imagination is what gives thought experiments an experimental character.¹⁶ It does not follow that mental models have no role to play in thought experiments. Building a mental model of the scenario presented by the thought experiment can help our imaginative reasoning within this scenario.

Both imagination and mental models can have a role in the elaboration of numerical experiments. Plausibly, a numerical experimenter, like Maria, who is figuring out the good model to be translated into a software that can be run on a computer, can exploit her capacity to mentally model different scenarios and to project herself into them in order to examine closely how she should proceed. Similar words can also be said about real experimenters. Also Sam, in planning his real experiments, can rely on mental models of the fluid-dynamical setup is building and imagine how to interact with it in a successful way.

The question is whether mental models and imagination are exploited when numerical and real experiments are carried out. It is not clear the role that mental models can play when Sam is manipulating the physical mechanism that causes shock waves in the fluid tank and observing what happens; or when Maria runs her software on a computer and visualises the output. In the case of thought experiments, mentally modelling is a way of building the given scenario and, I suggested, an aid to trigger or improve the imaginative projection into such a scenario. In the case of real and numerical experiments concrete objects or displayed patterns

¹⁶ Now we are in the position to see that calling thought experiments “imaginary experiments” does not always convey a negative view of what they are (see 1.1.): thought experiments can be positively seen as imagination-based experiments, rather than mere imaginary visualisations of experiments.

seem to play, at least partially, such a role.¹⁷ When, for instance, Maria visualises the output, her imagination can be triggered: what she visualises can help her reason from the virtual world created by the numerical experiment. Real experiments are meant to deal with the real world, rather than with virtual worlds. However, even real experimenters are very often dealing with proxies for their real targets (see fn 7)—e.g., Sam is using a fluid-dynamical setup as representing gasses flowing into the gravity well of a black hole. These proxies can act as props inducing an imaginative change in perspective—i.e., Sam imagines dealing with gasses flowing into the gravity well of a black hole.

Thus, I agree with Chandrasekharan and colleagues in thinking that mental simulation is a relevant mechanism in the building (i.e., elaboration and execution) of thought and numerical experiments, and I add real experiments, but imagination (i.e., mental simulation) seems to be more important than mental models (i.e., objectual mental simulation). Moreover, once the production dimension is divided into genuine building processes (i.e., elaboration processes) and execution processes, we can realise that thought experiments call for mental simulation (especially imagination) in a way that real and numerical experiments do not. We should not underestimate the different nature of the laboratories in which these experiments take place: thought experimentation is implemented in the mind, numerical experimentation in the computer, real experimentation in real laboratory settings.¹⁸ Thought experiments do not simply involve imagination, they are grounded in it, and this is due to the fact that their hardware component is the mind. Both real and numerical experiments can involve imagination. Their execution, however, does not rely on it, but rather on the relevant instrumental apparatus.

3. *The Presentation Dimension*

Chandrasekharan and colleagues suggest that mental simulation can also be a key to the presentation dimension of thought and numerical experiments. They do not go into details of this idea, since they complain that comparisons have mostly taken into account such a dimension. However, drawing on Nersessian's previous work, they suggest that an important aspect of the presentation dimension is the narrative

¹⁷ I say partially, because there might be a role for mental models in the execution of real and numerical experiments, especially when experimenters need to strengthen the link between their target and the (concrete or abstract) model they are dealing with. For instance, Sam might mentally model how the behaviour of his fluid-dynamical setup can stand for the supersonic gas jets formed when gasses flow into the gravity well of a black hole. Likewise, Maria might complete, so to speak, with mental models the outputs of her software run on a computer in order to better see how they can tell her something about the supersonic gas jets formed when gasses flow into the gravity well of a black hole.

¹⁸ Minimising this aspect can lead to consider thought and numerical experiments in mere abstract terms, as "abstract entities that can be implemented on different kinds of systems" (Humphreys 1993: 220).

aspect, which is common to real, thought and numerical experiment.¹⁹ Mental simulation would be the key to this narrative aspect.

Narratives are the means via which the results of an experiment (be it real, thought or numerical) are publicly presented. In these narratives the results are not merely presented, of course, but interpreted, that is, they are analysed in light of a theory or a scientific hypothesis. The narratives can be more or less detailed and rhetorically colourful, and include pictures. Nersessian (1993) points out that research in cognitive science has underlined the importance of mental simulation (i.e., mental modelling) in our comprehension of narratives. Thus, the idea is that the narrative aspect of experiments involves mental models. More specifically, the construction of a mental model can be helpful in making the narrative. Objectually mentally simulating what has been found, and how, can facilitate the interpretation processes and the writing of the public report. Mentally modelling can also be helpful in receiving the narratives. A reader of the description of the experiment can objectually mentally modelling what is described and better grasp the conclusions.

We have seen that the notion of mental simulation has two senses: objectual mental simulation (i.e., mental modelling) and mental mental simulation (i.e., imagination). What about imagination with respect to the narrative dimension of real, thought and numerical experiments? It is interesting to notice that also imagination has been identified as pivotally involved in our engagement with narratives. Imagination has been seen as what enables us to come up with alternative perspectives and to access them. Thus, imagination can underlie not only the production of narratives, but also their reception. All kinds of narratives are props that activate our imagination and enable us to reason in the described world, that is, to project ourselves into the described situation and to simulate what one would perceive or think from this perspective.

Stressing the role that mental simulation (especially imagination) has in the reception of the narrative describing an experiment reveals a feature proper to thought experimentation. Thought experiments show a re-performing aspect that both real and numerical experiments lack. In thought experiments the narrative component acts as a prop that induces the recipient to *re-perform* the given thought experiment, and thus to grasp its conclusion. Contrary to real and numerical experiments, thought experiments are executed twice: at the level of production, but also at the level of presentation. At the first level a thought experiment is privately executed by its author (for example by Schrödinger or Poincaré), and at the second level it is presented in a narrative form to a public, who is capable to re-execute it. We need more than our imagination to re-perform a real or numerical experiment. Here lies the power of thought experimentation. Thought experi-

¹⁹ The narrative aspect of thought experimentation is rather neglected in the literature (see Gooding 1993; Souder 2003; Swirski 2007), but it has led some authors to make a parallelism between thought experiments and literary fiction, which are narrations par excellence (see Arcangeli 2018).

ments are extremely important because they are intentional products related to the gaining, sharing and the spreading of knowledge. They do this “in human-friendly and graspable forms” (El Skaf and Imbert 2013: 3472). For this reason it is hard to think that numerical experimentation outperforms, and will finally replace, thought experimentation. This is not the end of thought experiments.

Conclusion

In conclusion, I agree with (the late) David Gooding:

How do scientists go from the actual to the possible, on the impossible, and return to an actual world altered by that journey? The short answer is that thought experimentation and real experimentation [and numerical experimentation] have much in common. (Gooding 1990: 70).

What they have in common is the structure of the procedure. After all, from a mere philological point of view, an experiment is an operation that fills our efforts of peering into things, a coming to knowledge through a series of trials.²⁰ Real, thought and numerical experimentations proceed by involving similar steps and processes. I claimed that we can identify two main dimensions: the performance and the presentation dimension. While the former involves elaboration and execution processes, the latter involves interpretation processes.

There is another hidden similarity between these three scientific tools, namely mental simulation. Drawing on the literature in philosophy of mind, I pointed out that “mental simulation” can refer to two different capacities, namely mental modelling and imagination. Mental models objectually mentally simulate, that is, they simulate states of affairs. By contrast, imaginings mentally mentally simulate, that is, they simulate mental states. Both types of mental simulation can have a role to play in the performance and presentation of real, thought and numerical experiments.

However, I stressed that real, thought and numerical experiments are differently implemented, and such a difference shows up when we focus on execution processes. Thought experimentation depends on mental simulation and cannot be performed without it. This is not the case for real or numerical experimentation, which exploit instrumental apparatus. Moreover, I claimed that what is really needed to thought experimentation is imagination, more than mental modelling, because only the former simulates phenomenological and functional aspects of experiences, which are crucial for the thought experimenter to project herself into a situation not present to her senses. The fact that thought experiments rely on imagination can explain another important aspect of them, that is, their capacity to be re-performed. Considering the presentation

²⁰ This is what *expèrior* (*ex* plus *pèrior*) means, whereas the suffix *mèntum* simply indicates the act. These two Latin words taken together form “*expèrimèntum*”, the Latin word from which “experiment” derives.

dimension, I stressed that who engages with the narrative presenting a thought experiment re-executes it in her mind thanks to her imagination (possibly helped by mental models of the scenario described by the narrative). This makes thought experimentation extremely helpful as a means to disseminate ideas.

It seems to me that three laboratories and three types of experimentation are at our disposal for studying the facets of the phenomena. Thought experimentation is fundamental to scientific research, just like real and numerical experimentation, and is not bound to be replaced by the latter, given the human-friendly way in which it conveys knowledge.

The analysis I offered here is, of course, just a starting point. The approach to thought experiments based on the notion of imagination and on a solid taxonomy of varieties of imagination should be strengthened. Moreover, a bridge is needed between this approach and epistemological issues concerning real and numerical experimentation as well.²¹ I hope to have shown, however, that it is worth treating equally, at a methodological level, real, thought and numerical experiments, and comparing them on the basis of a “process-oriented analysis”.

References

- Abell, C., Currie, G. 1999. “Internal and External Pictures.” *Philosophical Psychology* 12 (4): 429–445.
- Arcangeli, M. 2013. “Immaginare è simulare: cosa e come?” *Rivista di Estetica*, 53 (2): 135–154.
- Arcangeli, M. 2010. “Imagination in Thought Experimentation: Sketching a Cognitive Approach to Thought Experiments.” In L. Magnani, W. Carnielli, C. Pizzi (eds.). *Model-Based Reasoning in Science and Technology*. Berlin: Springer: 571–587.
- Arcangeli, M. 2018. “L’expérience de pensée comme récit. Le rôle crucial de l’imagination.” C. Bouriau, G. Schuppert (eds.) *Perspectives philosophiques sur les fictions*. Paris: Kimé: 111–129.
- Arcangeli, M. 2017. “Thought Experiments in Model-Based Reasoning.” In L. Magnani, T. Bortolotti (eds.). *Springer Handbook of Model-Based Science*. Berlin: Springer: 465–495.
- Barberousse, A., Franceschelli, S., Imbert, C. 2009. “Computer simulations as experiments.” *Synthese* 169 (3): 557–574.

²¹ One issue, for instance, deals with how these tools face the phenomena’s complexity. Following Winsberg (2009), it would be interesting to look at the differences between the three kinds of experimentation in the argument given for the legitimacy of the move from model to target, and in the background knowledge that grounds that argument. Moreover, it would be worth exploring the similarities between thought and numerical experiments in this respect. Indeed, they seem to explore the actual world and its physical possibilities via excursions on metaphysically possible worlds, for they can deal with scenarios contradicting physical laws of the actual world. However, a numerical experiment is more useful when the target’s dynamics is well known, and when we are interested in exploring it on multiple worlds. By contrast, a thought experiment is not committed to the dynamics of the target, and explores worlds one at a time. I leave the exploration of such differences to another occasion.

- Beisbart, C., Norton J. 2012. "Why Monte Carlo Simulations are Inferences and not Experiments." *International Studies in Philosophy of Science* 26: 403–422.
- Bishop, M. 1998. "An Epistemological Role for Thought Experiments." In N. Shanks (ed.). *Idealization in Contemporary Physics*. Amsterdam/Atlanta: Rodopi: 19–33.
- Brendel, E. 2004. "Intuition pumps and the proper use of thought experiments." *Dialectica* 58: 88–108.
- Brown, J. R. 1991a. "Thought Experiments: a Platonic account." In T. Horowitz, G. Massey (eds.). *Thought Experiments in Science and Philosophy*. Lanham: Rowman and Littlefield: 119–128.
- Brown, J. R. 1991b. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. London: Routledge.
- Buzzoni, M. 2004. *Esperimento ed esperimento mentale*. Milano: FrancoAngeli.
- Buzzoni, M. 2008. *Thought experiment in the natural sciences: An operational and reflexive-transcendental conception*. Würzburg: Königshausen and Neumann.
- Chandrasekharan, S. et al. 2013. "Computational Modeling: Is This the End of Thought Experiments in Science?" In Frappier, M., L. Meynell, J. R. Brown (eds.). *Thought Experiments in Philosophy, Science, and the Arts*. London and New York: Routledge: 239–260.
- Cooper, R. 2005. "Thought Experiments." *Metaphilosophy* 36: 328–347.
- Currie, G., Ravenscroft, I. 2002. *Recreative minds: Imagination in philosophy and psychology* Oxford: Oxford University Press.
- Di Paolo, E. A., Noble, J., Bullock, S. 2000. "Simulation Models as Opaque Thought Experiments." In M. A. McCaskill et al. (eds.). *Proceedings of the Seventh International Conference on Artificial Life*. Cambridge: MIT Press: 497–506.
- Doggett, T., Egan, A. 2007. "Wanting Things You Don't Want: The Case for an Imaginative Analogue of Desire." *Philosophers' Imprint* 7 (9): 1–17.
- Dokic, J., Arcangeli, M. 2015a. "The Heterogeneity of Experiential Imagination." In Metzinger, T. K., Wind J. (eds). *Open MIND: 11(T)*, <http://open-mind.net/papers/the-heterogeneity-of-experiential-imagination>. Frankfurt am Main: MIND Group [published in *Open MIND. Philosophy and the Mind Sciences in the 21st Century*. Cambridge: MIT Press: 431–450].
- Dokic, J., Arcangeli, M. 2015b. "The Importance of Being Neutral: More on the Phenomenology and Metaphysics of Imagination. A Reply to Anne-Sophie Brüggem." In Metzinger, T. K., Wind J. (eds). *Open MIND: 11(T)*, <http://open-mind.net/papers/the-importance-of-being-neutral-more-on-the-phenomenology-and-metaphysics-of-imagination2014a-reply-to-anne-sophie-brueggen>. Frankfurt am Main: MIND Group [published in *Open MIND. Philosophy and the Mind Sciences in the 21st Century*. Cambridge: MIT Press: 461–465].
- Dowling, D. 1999. "Experimenting on theories." *Science in Context* 12 (2): 261–273.
- El Skaf, R., Imbert, C. 2013. "Unfolding in the empirical sciences: experiments, thought experiments and computer simulations." *Synthese* 190: 3451–3474.
- Gilbert, N., Troitzsch, K. 1999. *Simulation for the Social Scientist*. Philadelphia: Open University Press.

- Goldman, A. 2006. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.
- Gooding, D. 1990. *Experiment and the making of meaning: Human agency in scientific observation and experiment*. Dordrecht and Boston: Kluwer Academic Publishers.
- Gooding, D. 1993. "What is Experimental About Thought Experiments?" In D. Hull, M. Forbes, K. Okruhlik (eds.). *PSA 1992*. East Lansing: Philosophy of Science Association: 280–290.
- Häggqvist, S. 1996. *Thought experiments in philosophy*. Stockholm: Almqvist and Wiksel.
- Hull, D. 1997. "That Just Don't Sound Right: A Plea for Real Examples." In J. Earman, J.D. Norton (eds.). *The Cosmos of Science: Essays of Exploration*. Pittsburgh: University of Pittsburgh Press: 430–457.
- Hull, D. 1999. "A Function for Actual Examples in Philosophy of Science." In M. Ruse (ed.). *What the philosophy of Biology is: essays dedicated to David Hull*. Dordrecht: Kluwer: 309–321.
- Humphreys, P. 1993. "Seven theses on thought experiments." In J. Earman, A. Janis, J. Massey, N. Rescher (eds.). *Philosophical Problems of the Internal and External World: Essays on the Philosophy of Adolf Grunbaum*. Pittsburgh/Konstanz: University of Pittsburgh Press/Universitätsverlag Konstanz.: 205–227.
- Johnson-Laird, P. N. 1983. *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness*. Cambridge: Harvard University Press.
- Johnson-Laird, P. N. 2004. "The History of Mental Models." In K. Manktelow, M.C. Chung (Eds.), *Psychology of Reasoning: Theoretical and Historical Perspectives*, pp. 179–212, (Psychology Press, New York 2004).
- Lenhard, J. 2011. "Epistemologie der Iteration: Gedankenexperimente und Simulationsexperimente." *Deutsche Zeitschrift für Philosophie* 59: 131–154.
- Lenhard, J. 2018. "Thought experiments and simulation experiments: exploring hypothetical worlds." In M. T. Stuart, Y. Fehige, J. R. Brown (eds.). *The Routledge Companion to Thought Experiments*. London: Routledge.
- Mach, E. 1896. "Über Gedankenexperimente" *Zeitschrift für den Phys. und Chem. Unterr.* 10: 1–5. Translated by W. O. Price, S. Krinsky 1973. "On thought experiments." *Philosophical Forum* 4 (3): 446–457.
- Miščević, N. 1992. "Mental Models and Thought Experiments." *Int. Stud. Philos. Sci.* 6: 215–226.
- Miščević, N. 2007. "Modelling Intuitions and Thought Experiments." *Croatian Journal of Philosophy* 7: 181–214.
- Nersessian, N. 1993. "In the Theoretician's Laboratory: Thought Experimenting as Mental Modelling." In D. Hull, M. Forbes, K. Okruhlik (eds.). *PSA 1992*. East Lansing: Philosophy of Science Association: 291–301.
- Nersessian, N. J. 2007. "Thought Experiments as Mental Modelling: Empiricism without Logic." *Croatian Journal of Philosophy* 7: 125–161.
- Norton, J. 2004. "Why Thought Experiments Do Not Transcend Empiricism." In C. Hitchcock (ed.). *Contemporary Debates in the Philosophy of Science*. Oxford: Blackwell: 44–66.
- Parker, W. 2009. "Does Matter Really Matter? Computer Simulations, Experiments and Materiality." *Synthese* 169 (3): 483–96.

- Poincaré, H. 1902. *La Science et l'hypothèse*. Paris: E. Flammarion. Translated by W. J. Greenstreet: *Science and hypothesis*. New York: The Walter Scott publishing, 1905).
- Reiner, M., Gilbert, J. 2000. "Epistemological resources for thought experimentation in science learning." *International Journal of Science Education* 22 (5): 489–506.
- Sorensen, R. 1992. *Thought Experiments*. Oxford: Oxford University Press.
- Souder, L. 2003. "What are we to think about Thought Experiments?" *Argumentation* 17: 203–217.
- Stäudner, F. 1998. *Virtuelle Erfahrung. Eine Untersuchung über den Erkenntniswert von Gedankenexperimenten und Computersimulationen in den Naturwissenschaften*. Ph.D. Diss., Friedrich Schiller Universität.
- Stöltzner, M. 2003. "The Dynamics of Thought Experiments—Comment to Atkinson." In M. Galavotti (ed.). *Observation and Experiment in the Natural and Social Sciences*. Dordrecht: Kluwer: 243–258.
- Swan, L. S. 2009. "Synthesizing insight: artificial life as thought experimentation in biology." *Biol. Philos.* 24: 687–701.
- Swirski, P. 2007. *Of literature and knowledge: explorations in narrative thought experiments, evolution and game theory*. London and New York: Routledge.
- Thagard, P. 2010. *The Brain and the Meaning of Life*. Princeton: Princeton University Press.
- Thagard, P. 2014. "Thought Experiments Considered Harmful." *Perspectives on Science* 22: 288–305.
- Velasco, M. 2002. "Experimentación y técnicas computacionales." *Theoria* 17: 317–331.
- Wilkes, K. 1988. *Real people: Personal identity without thought experiments*. Oxford: Clarendon Press.
- Winsberg, E. 2009. "A Tale of Two Methods." *Synthese* 169 (3): 575–592.
- Winsberg, E. 2015. "Computer Simulations in Science." In E. N. Zalta (ed.). *The Stanford Enc. of Phil.* (Summer 2015 Edition). <https://plato.stanford.edu/archives/sum2015/entries/simulations-science/>.
- Zeimbekis, J. 2011. "Thought Experiments and Mental Simulations." In K. Ierodiakonou, S. Roux (eds.). *Thought Experiments in Methodological and Historical Contexts*. Leiden-Boston: Brill: 193–215.

The Mathematics-Natural Sciences Analogy and the Underlying Logic. The Road through Thought Experi- ments and Related Methods

MAJDA TROBOK*
University of Rijeka, Rijeka, Croatia

The aim of this paper is to point to the analogy between mathematical and physical thought experiments, and even more widely between the epistemic paths in both domains. Having accepted platonism as the underlying ontology as long as the platonistic path in asserting the possibility of gaining knowledge of abstract, mind-independent and causally inert objects, my widely taken goal is to show that there is no need to insist on the uniformity of picture and monopoly of certain epistemic paths in the epistemic descriptive context. And secondly, to show the analogy with the ways we come to know the truths of (natural) sciences.

Keywords: Thought experiment, epistemology, philosophy of mathematics, natural sciences, descriptive epistemic context.

1. Introduction

To endorse standard platonism in the philosophy of mathematics is not to be confined to platonic perception, as usually thought. In the same way, to defend other, non-standard, versions of platonism is not to be limited to some specific epistemic paths either. The aim of this paper is to show why this is the case and in which sense the plurality of epistemic paths in the domain of mathematics is analogous to the epistemic routes in the descriptive epistemic context given the domain of the natural sciences.

* Research for this paper was carried out under the project “Rationality: between Logically Ideal and Commonsensical in Everyday Reasoning”. The project is funded by the Croatian Science Foundation. IP-2016-06-2408.

Standardly, platonism in epistemology is “the view that mathematics is about objects of which we have a priori knowledge”, where by “object” is meant a mind-independent entity. According to standard or Gödelean Platonism we gain mathematical knowledge via platonic perception.

Usually when the background philosophy, i.e. ontology is taken to be platonism, the epistemology is concentrated on the discussion about the existence of the platonic perception and those who endorse Platonism but not the Platonist perception, offer alternative epistemic routes, in order to avoid platonic intuition, taken to be a mysterious and unclear procedure of direct grasp of abstract objects. Given Benacerraf’s argument against Platonism in epistemology, that is against the possibility of grasping truths concerning objective, abstract, non spatio-temporally located objects, other versions of platonism take other routes to have the epistemic monopoly (such as e.g. recarving in neo-Fregean platonism) or offer a variety of alternative routes to the platonic intuition (e.g. ante rem structuralism).

In this paper the goal is to redirect the attention on two different points: the plurality of platonic epistemic paths and the analogy with the epistemic paths in the natural sciences.

I shall try firstly to show that there are more than one possible epistemic routes of gaining mathematical knowledge compatible with platonism (in the sense of grasping mathematical objects and their relations). Secondly, my aim is to show how it is possible to dig up other epistemic routes without giving up on platonism and without giving up the mathematics-natural sciences analogy.

Within the domain of standard platonism, the idea is hence to both dethrone platonic intuition and situate it amongst other epistemic routes, and to show the analogy with the way we gain knowledge in the natural sciences. The focus will be on experiments, in particular on thought experiments.

2. *The historically oriented research*

The development of mathematical knowledge as well as the process of discovery in the (natural) sciences could be standardly analysed from different perspective: if could be analysed within the cognitive science oriented research, or within the historical oriented research, or a computationally oriented research, and so on.

In this paper I shall focus on the critical analysis of the mathematical descriptive epistemic paths as well as the epistemic mathematics-natural sciences analogy through the prism of the historically oriented research.

Even though it might come as a surprise, given that platonism is here taken to be the underlying ontology, I take the accepted methodology for epistemology of science and mathematics to be (Kitcher’s) pragmatic naturalism, in particular his view that we ought to look at the history in order to determine the epistemology since “history is the

teacher of epistemology” (italics mine). The underlying idea is hence that the epistemological route follows the historical one, and that “the epistemological order of mathematics broadly recapitulates the historical order.” Even though one of the tenets of pragmatic naturalism is the denial of a priori knowledge (in the domain of mathematics as much as elsewhere), here the idea is rather merely to dethrone platonistic intuition and situate it amongst other epistemic routes. And the justification for doing so comes from history itself, which offers reasons for endorsing platonist intuition as much as other epistemic modes and that thus ironically ends up as a turn-the-table for Platonism. The importance of the historical analysis is threefold: it firstly justifies the endorsement of the platonic perception, secondly it justifies the plurality of epistemic modes in gaining mathematical knowledge, and finally it justifies the endorsement of the mathematics-natural sciences analogy.

When talking about different epistemic modes in grasping mathematical objects, the mathematics-science analogy it’s imposing itself to us and turns out to be particularly strong in such descriptive epistemic context. Let us have a look at such mathematics-science epistemic analogy in more details.

3. Three modes of epistemic access and the mathematics-(natural) sciences link.

I shall propose three main modes of initial epistemic access to both mathematical and scientific reality (objects and properties): (1) *Perception: Visual and Platonic*, (2) *Experiment* and (3) *Introduction (or hypothetical positing) and positing (or categorical positing) of objects*.

The epistemological science-mathematics analogy turns out to be overall, each epistemic path in science having its counterpart in mathematics. The plenitude of such paths is (to be) determined and classified by looking at the history of science, that is mathematics.

Let us hence have a look at the mentioned epistemic paths in the given order.

(1) Perception: Visual and Platonic

In scientific research, one epistemic way is sensory, primarily visual, direct perception of objects and phenomena. The analogy in mathematics would be the platonic “pi in the sky”, direct access to the mathematical objects and statements, often called platonic perception/intuition. When talking about it, J. R. Brown points out:

The main idea is that we have a kind of access to the mathematical realm that is something like our perceptual access to the physical realm. This doesn’t mean that we have direct access to everything: the mathematical realm may be like the physical where we see some things, such as white streaks in bubble chambers, but we don’t see others, such as positrons. (Brown 1999: 13)

The platonic intuition—visual perception analogy is something standard platonists traditionally heavily insist on. Brown says:

Just as the mathematical mind can grasp (some) abstract sets, so the scientific mind can grasp (some of) the abstract entities which are the laws of nature. (Brown 1991)

Gödel in particular famously insists on the analogy while saying that the assumption of such objects is quite as legitimate as the assumption of physical bodies and there is quite as much reason to believe in their existence. They are in the same sense necessary to obtain a satisfactory system of mathematics as physical bodies are necessary for a satisfactory theory of our sense perceptions. (Gödel 1944: 456f)

and, in one of the most famous quotations in the philosophy of mathematics, that

despite their remoteness from sense experience, we do have something like a perception also of the objects of set theory, as is seen from the fact that the axioms force themselves upon us as being true. I don't see any reason why we should have any less confidence in this kind of perception, i.e. mathematical intuition, than in sense perception. (Gödel 1947: 484)

Such an analogy has been criticised by many, and for several reasons. Apart from being unnatural and forced, the analogy seems to take the existence of the platonic perception for granted while in effect it is most contentious. It hence has been heavily criticised by both platonism's friend and foes. Kitcher remarks that

...what some mathematicians call "intuition" or even (in the case of Ramanujan) the visitation of the goddess (Namakiri), *can* be explained as 'fine-tuned abilities [...] rooted in extant mathematical practice. (Kitcher 2011)

While Shapiro, who is endorsing *ante rem* structuralism (a version of non-standard platonism) follows the same line as Kitcher's when underlying that

...the axioms do not force themselves on a first (or second, or third) reading. For virtually any branch of mathematics, the psychological necessity of the axioms and inferences, and the feeling that the axioms are natural and inevitable, comes only at the end of a process of training in which the student acquires considerable practice working within the given system, under the guidance of teachers. (Shapiro 1997: 212)

How to respond to such criticisms?

The main point to be underlined is that mathematical intuitions are not just theoretical presuppositions of the philosophers of mathematics but are being asserted by working mathematicians themselves (Cantor, Gödel, Ramanujan, Hardy etc.). We hence have good reasons to transpose this fact from the history of mathematics, right into the core of our epistemology. Kitcher's main point being that what is explained as platonic perception could easily be explained by invoking the mathematicians' "fine-tuned abilities" that are "rooted in extant mathematical practice", instead of being some mysterious faculty of the mind. Wanting to have a closer look at Kitcher's remark, and to

see what such “fine-tuned abilities” would amount to, we are to face a dilemma. Namely, if what is meant by “fine-tuned” is just the perfect conformity to the extant practice in the profession, it is difficult to see how Gödel’s non-conformistic example, not to speak about Ramanujan, would fit in the picture. If, on the other hand, fine-tuning refers to an impressive ability to reach the truth beyond the extant research paradigm, then it certainly is compatible with the platonistic account on which such a fine-tuning (to the mathematical reality) might culminate in intuitive insights.

(2) *(Thought) Experiment*

Experiments have been usually perceived as the lynchpin of empirical sciences, a method for discovering the facts of nature and hence as belonging to the sphere of practical research. Mathematics being an armchair activity—how can the mathematical domain be related to any experimental epistemic route?

I am using the entry from Stanford Encyclopedia to provide a mainstream characterisation of the role of experiment:

Physics, and natural science in general, is a reasonable enterprise based on valid experimental evidence, criticism, and rational discussion. It provides us with knowledge of the physical world, and it is experiment that provides the evidence that grounds this knowledge. [...] It can also call for a new theory, either by showing that an accepted theory is incorrect, or by exhibiting a new phenomenon that is in need of explanation. Experiment can provide [...] evidence for the existence of the entities involved in our theories. Finally, it may also have a life of its own, independent of theory. Scientists may investigate a phenomenon just because it looks interesting. Such experiments may provide evidence for a future theory to explain. [...] a single experiment may play several of these roles at once. (Franklin and Perovic 2016)

The standard taxonomy, when talking about experiments, includes the distinction between confirmatory (or demonstrative) on one hand and the exploratory experiments on the other. The former having the goal of testing theories, while the latter has as the primary goal the experimentation that is not guided by hypotheses but it rather a process or searching.

My aim, at this point, is to show that, no matter which of the two main sub-species of the experiment we prefer to concentrate on, either the confirmatory or the exploratory (non-demonstrative) one, the analogy with the mathematical case holds throughout.

If talking about the confirmatory experiments, there are examples from the mathematical practice that could be treated as examples of such experiments. Let us here mention the proof that number π is irrational. The number π has been studied for centuries (since ancient time) and so was the notion of irrational numbers. Aryabhata apparently hinted at number π being irrational in 500 CE. Such an outcome was accepted as a new mathematical result not prior to the 18th cen-

tury when Lambert (in 1761) proved it to be irrational. And then again, in the same paper in which he proved π 's being irrational, Lambert conjectured that number π is transcendental too, which was accepted in mathematics in 1882, when proved by Lindemann.

Other mathematical results and proofs are analogous to the exploratory, non-demonstrative experiments. In such experiments the experimentation is not guided by hypotheses. An example in mathematics could be the problem of trisection of an arbitrary angle. The attempts to solve the problem, can be seen as an exploratory experiment that had been going on for centuries.

Notwithstanding the mentioned mathematical examples, when comparing the experiments in science with those in mathematics, we might still find the proposed analogy implausible. And basically for two reasons: (a) experiments in science are practical procedures, done in laboratories, unlike in the mathematical domain, and (b) if anything, given the possibility to directly intervene on the objects in scientific experiments, which is not possible in the mathematical domain given the abstract nature of mathematical objects (and hence their being causally inert). Let us have a closed look at the possible replies at the two just-mentioned reasons.

(a) Experiments in science are practical procedures, done in laboratories, unlike those in the mathematical domain. Well, ought experiments to be practical in the first place? When thinking about, e.g. high school experimentation, than we all have in mind the paradigmatic example of the laboratory and the practical procedures that we were performing there during the natural-sciences classes. The taxonomy however includes three types of experiments: the real, the imaginary and the thought experiments. The real ones are those that have been performed, the imaginary are those that haven't been formed but could have been, while the thought experiments are those that those that could not be performed due to the lack of technology or because impossible in principle.

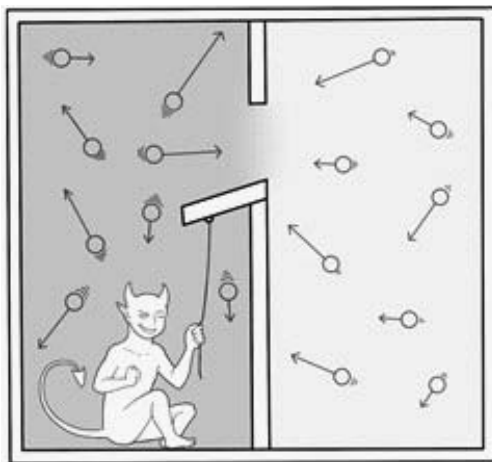
And when we look at the way experiments have been perceived by scientists through history, there is no uniformity of picture; not even a general agreement on experiments being real-world, practical methods for acquiring knowledge. Even in Galileo's writings the distinction between real and imaginary experiments is not a sharp one and it is, for some experiments, a contentious issue. Newton's bucket experiment is another example of an experiment that was originally an imaginary experiment but needn't be. What about thought experiments? Such experiments apparently played a major role in the development of scientific theories in the work of Galileo, Newton, Einstein, Heisenberg. Examples are legion: Galileo's experiment with the result that all bodies fall at the same speed, Schrödinger's cat, Maxwell's demon, Einstein chasing a light beam, Twin paradox and others.

Still—and here we shall focus on the above remark (b)—what happens in experimental science might seem at first sight to be remote from the standard mathematical practice.

If anything given the abstract nature of mathematical objects, i.e. the fact that they are not spatio-temporally located and are causally inert. Any intervention/manipulation on the objects of the domain fails to be possible in the case of mathematical experiments, and on abstract objects involved. So, while we can, to take the morally controversial example of tissue engineering—the Vacanti mouse, implant under the skin of a mouse a cartilage structure (and then the cartilage naturally grows by itself), is not clear what the counterpart of such a direct manipulation of objects in the case of mathematical experiment would amount to.

Could we, figuratively speaking, have a Vacanti number or a Vacanti geometric figure? Certainly not! Should we (again) infer from that that the analogy is, to put it blandly, farfetched and artificially imposed? Not so fast.

Namely, it is difficult to guess in which way we literally manipulate (concrete) objects in thought experiments. We actually do not. We can hence talk about experiments without presupposing any kind of direct manipulation of concrete objects. Let take the example of one of the most famous thought experiments: “Maxwell’s demon”. According to the Second law of thermodynamics, in any change of state entropy must remain the same or increase; it cannot decrease. In laymen’s terms, heat cannot pass from a cold to a hot body. Maxwell’s goal in the experiment is to show that the Law is to be read in probabilistic terms, which means that, in principle, it could be possible for the heat to pass from a hot body to a colder one. In order to show how this could be possible, Maxwell imagine to have two connected boxes, which the Devil at the door that connect those two boxes (see picture). The two boxes contain some gas, and in particular the gas in the left box is hot, while the gas in the right box is cold. What is to be expected, according to the Second law of thermodynamic, is for the heat to pass from the hot gas to the cold one. But, during the experiment, the Demon decides to let the fast molecules from the cold box into the hot box, and the slow molecules from the hot box into the cold one. By letting the fast molecules from the cold box into the hot box, and the slow molecules from the hot box into the cold one, there will be an increase in the average speed in the hot box and a decrease in the average speed of molecules in the cold. Since, on Maxwell’s theory, heat is just an average speed of the molecules, there has been a flow of heat from the cold box to the hot one—contrary to what is expected according to the Second law of thermodynamic. Hence, the Law—and that is precisely Maxwell’s point—has to be interpreted probabilistically.



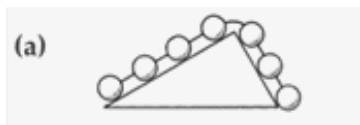
Maxwell's demon experiment.
Illustrated by Maja Grčki.

Let us further focus on some aspects of the analogy and the meeting points between the thought experiments in the natural sciences and some of the basic mathematical procedures. We shall analyse in more details the process of representations of abstract objects and the sameness of structure of some thought experiments in science with the *reduction ad absurdum* structure in mathematical proofs.

Let us start with some properties that the representations of abstract objects share both in the natural sciences and in mathematics.

The non-concrete objects which we (mentally) “manipulate” during imaginary or thought experiments are often related to their spatio-temporally counterparts. And the way these two kinds of objects are related might be analogous to the way in which representations of abstract objects—the subject of manipulations in mathematical experiments—are related to the abstract (mathematical) objects, i.e. their abstract counterparts. In the case of the trisection of an arbitrary angle, we do manipulate the representation of an abstract geometrical entity.

Thought experiments in the natural sciences can also share the same structure of standard proof methods in logic and mathematics. Let us take the example of Stevin's thought experiment. As well known, there are three possible planes: the horizontal, the vertical and the inclined plane. If we put a weight on each of these planes than we already know that on the horizontal plane the weight remains at rest, while on the vertical plane the wight freely falls. What about the inclined plane? What happens with a weight if put on an inclined plane?



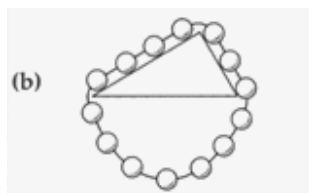
Suppose we have a chain with weights and we put it on an inclined plane (picture (a)). How does the chain move? Well, there are obviously three possible answers:

- (1) it remains at rest (in so called static equilibrium)
- (2) it moves to the left
- (3) it moves to the right.

The right answer is (1), it remains at rest. The next step is to prove it!

Let us suppose not-(1) (notice the *reductio ad absurdum* structure of the proof!)

If not-(1), it means that the force of the left is not balanced by the force in the right. Let us now add the links at the bottom so to get a closed loop (picture (b))



If not-(1) were the case, the loop would rotate and hence, we would get a *perpetuum mobile*, which is impossible. Hence, the chain remains at rest.

The analogy is better presented in the following table:


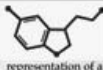
Thought experiments in science as (quasi) RAA (<i>reductio ad absurdum</i>) proofs	
example: Stevin's thought experiment	
RAA proving method	structure of Stevin's thought experiment
A —the statement that we want to prove	A —the chain remains at rest
suppose not-A	suppose the chain does not remain at rest (not-A)
not-A leads to contradiction	if the chain moved (if not-A), we would then have a <i>perpetuum mobile</i> — impossible!
hence, A	hence , the chain remains at rest (A)

Table showing Stevin's thought experiment having the structure of the *reductio ad absurdum* deduction rule

Why the “quasi” in the “Thought experiments in science as (quasi) RAA (*reductio ad absurdum*) proofs” title? Well, the *reductio ad absurdum* method of deduction require the initial hypnos to lead to *absurdum*,

which means to (logical) contradiction, i.e. to a statement of the form A and not-A. Technically, a *perpetuum mobile* is not a logical impossibility but a nomic one. Hence it is not a (logical) *absurdum*. However, the idea is here that Stevin's proof and any *reductio ad absurdum* one (in logic or mathematics) do share the same structure (see the above table).

The natural science-mathematics analogy, rather surprisingly, holds even if we decide to concentrate on *real* experiments in science instead of imaginary and thought ones. Such experiments are, in fact, in many aspects similar to some examples of manipulative procedures/proofs in mathematics. Let us have a look at two examples.

	Mathematics	Natural sciences
Abstract object	✓ e.g. an angle	
Concrete object (spatio-temporally located)		✓ e.g. a molecule
Representation	 representation of an angle	 representation of a (serotonin) molecule

The mathematical one is Lakatos' historical proof case, while the one in the natural sciences is Hooke's observations made with microscope. We easily notice the same dynamic language used in both cases, that is in both domains. (See the table below.)

Mathematics domain	Natural sciences domain
Lakatos' historical proof case $V-E+F=2$, for any polyhedron	Hooke's observations made with microscope discovery of cells
Dynamic language	Dynamic language
– take an arbitrary polyhedron	– take a small piece of e.g. onion
– remove one of the faces	– remove (peel off) the membrane
– stretch the remaining figure out on flat surface	– stretch the part you want to analyse on a glass
– remove the lines one at the time, etc.	– add a few drops of water or solution, etc.

My conclusion at this point would be that there is a strong analogy between the experiment in science and some of the central procedures in mathematics. However, experiments in science and mathematics represent just one possible epistemic path in gaining knowledge. Another one is the positing of objects, which can be either hypothetical or categorical.

(3) *Introduction (or hypothetical positing) and positing (or categorical positing) of objects*

In science and mathematics there are objects that we introduce in our theories in order to make the overall theory complete and/or to explain the appearance of discrepancies. And there again, the mathematics-science analogy enters the picture.

An example in science might be the discovery of the existence of the planet Neptune. After the discovery of Uranus, it was noticed that its orbit was not as it should be in accordance with Newton's laws, which led the astronomers to the assumptions that a (still to be discovered) planet might be the cause of the discrepancy. Astronomers hence predicted the position of an unobserved planet perturbing the orbit of Uranus.

In mathematics a nice example of introducing objects is the one concerning the methods of solution of the quadratic and cubic equations. Negative square roots appeared in Cardano's *Ars Magna* (1545) that contains the first occurrence of complex numbers. Cardano introduced negative square roots as solutions of the quadratic equation $x^2 - 10x + 40 = 0$. Since it was evident that solving the equation was impossible (in \mathbb{R}), Cardano decided to formally introduce the negative square roots. As the existence of the planet Neptune was first predicted by mathematical calculations and then empirically detected, i.e. seen by a telescope which gave the prediction the ontological force, analogously the complex numbers were first introduced formally and then, 300 years later, a more specifying meaning was attached to them. The strategy of introducing planets in the astronomy case corresponds to the formal introduction of complex numbers in the mathematical case. Similarly, in the following step, the astronomers pass from the hypothetical assumption to the categorical claim concerning the existence of the planet, while in the mathematical case, the analogous step goes from the formal introduction of negative square roots to the full-blown positing of new, complex numbers.

4. *The mathematics-natural sciences analogy.*

More aspects concerning the underlying logic

One possible reaction at this point might be: when talking about the mathematics—natural sciences analogy, is there an insurmountable difference in methodology? Possible complaint: the reasons for asserting the analogy between mathematics and the natural sciences in the descriptive epistemic context are marginal. Namely, there is a (much more) essential parameter that should be taken into consideration and that might make the difference between the epistemic paths in maths and the natural sciences come to surface: the underlying methodology.

When referring to the underlying methodology, i.e. logic, the standard view is that in maths, unlike the natural sciences, the basic methodology is the axiomatic-deductive method (of the geometric tradition). Contrary to that, in the natural sciences, the logic underlying

the research is primarily inductive/abductive. It implies that the two domains are profoundly methodologically different given the difference at the core, that is at their underlying logic. To that remarks, I find the most plausible reply to be the following one.

Proofs/theorem/theories at the final stage (textbook) do not coincide with the heuristics (in the sense of the epistemic paths within the context of discovery). The structure of the polished theory and the underlying deductive system do not however correspond to the research process in the epistemic descriptive context.

Lakatos nicely underlines the difference between the historical development of mathematical results and the procedures we find in mathematics textbooks. Such a difference amounts to the difference between the *preformal* development (correlates to the context of discovery, i.e. the epistemic descriptive context we've been focused) on and the *formal* articulation (corresponding to the context of justification) of a branch of mathematics by offering reasons for asserting that preformal proofs are not simply drafts of the formal ones but rather heuristic explanatory and exploratory tools having a development on their own.

A very simple yet illuminating example is the one Pólya presents in his *How to Solve It* (Pólya 1945: 114–117). Let us have a closer look at it. And let us start by supposing that a mathematician is helping their child to write the homework in mathematics, and at some point the child is supposed to calculate $1+8+27+64$ and solve it rightly by writing the result: 100. While waiting for the child to solve the exercise, the parent/mathematician notices that all four of the numbers/addends are cubs while the result (100) is a square. So that it is possible to write the mentioned equation in the form: $1^3+2^3+3^3+4^3=10^2$. He also notices that the mentioned sum is the sum of the cubes of the first four natural numbers. And then ask himself if it is a coincidence or it is not an isolated case to have the sum of the cubes of the first n natural numbers to be equal to a square. Pólya comments such a situation by comparing the parent/mathematician with the naturalist:

In asking this,¹ we are like the naturalist who, impressed by a curious plant or a curious geological formation, conceives a general question. Our general question concerns with the sum of successive cubes $1^3+2^3+3^3+4^3+\dots+n^3$. We were led to it by the “particular instance” $n=4$. (Pólya 1945: 115)

How would the mathematician procede at this point? What would he do? Pólya's answer is that the mathematician would do what the naturalist would do—investigate other special cases! And realise that:

$$\begin{aligned} 1^3 &= 1^2 \\ 1^3 + 2^3 &= 9 = 3^2 \\ 1^3 + 2^3 + 3^3 &= 36 = 6^2 \\ 1^3 + 2^3 + 3^3 + 4^3 &= 100 = 10^2 \\ 1^3 + 2^3 + 3^3 + 4^3 + 5^3 &= 225 = 15^2 \\ &\dots \end{aligned}$$

¹ Pólya here refers to the question as to whether it is a coincidence or a general rule that the sum of the cubes of the first n natural numbers is equal to a square.

The mathematician might subsequently notice that the results on the right side of the equations, i.e. the squares, follow a regularity, a certain pattern too. Namely:

$$1^2=1^2$$

$$3^2=(1+2)^2$$

$$6^2=(1+2+3)^2$$

$$10^2=(1+2+3+4)^2$$

$$15^2=(1+2+3+4+5)^2$$

...

Interestingly enough, the sum of the cubes of the first n natural numbers is equal to the square of the sum of the first n natural numbers. Given that this regularity seems to be general too, the assertion finally obtains the form:

$$1^3+2^3+3^3+\dots+n^3=(1+2+3+\dots+n)^2$$

This initial procedure, as pointed out by Pólya, is based on observation and induction and as such corresponds to the procedures of investigation in the natural sciences where the naturalist “may also reexamine the facts whose observation has led him to his conjecture; he compares them carefully, he tries to disentangle some deeper regularity, some further analogy” (Pólya 1945: 116).

In mathematics as in the physical sciences we may use observation and induction to discover general laws. But there is a difference. In the physical sciences, there is no higher authority than observation and induction but in mathematics there is such an authority: rigorous proof. (Pólya 1945: 117)

The difference Pólya is referring to is however beyond the scope of this article. This idea of the mathematics-natural sciences analogy is meant to be confined to the epistemic descriptive context, while the disanalogy enters the picture in the context of justification, which I am not addressing in this paper.

To summarise, in this paper I have taken the underlying ontology to be a version of standard platonism. I choose not to refrain from endorsing the platonic perception as one of the possible epistemic paths and hence from endorsing a version of standard platonism since I have hopefully showed that platonic perception is not to be banned from the epistemology of mathematics domain given that we do have good reasons for endorsing it.

I have then argued that, in the domain of mathematical entities and within the descriptive epistemic context, there is however a plurality of platonic epistemic paths and that such paths in the mathematical domain are analogous to the epistemic paths in the natural sciences.

Last but not least, I analysed the mathematics vs. natural sciences analogy from the perspective of the underlying logic. I have claimed the importance of keeping in mind the distinction between the context for discovery and the one of justification. The former being correlated with

the so called *preformal* development of statements or theories in mathematics, while the latter being connected with the *formal* articulation of branches of mathematics. When concentrating on the mathematics-natural sciences analogy and the underlying logic, it is important to take into consideration that the analogy holds in the context of discovery, in which the *preformal* development plays the major role. Such a development is based, both in mathematics and in the natural sciences, mostly on induction. It is certainly true that in mathematics the basic logical apparatus is deduction and mathematical induction (and that differs from the logical apparatus used in the natural sciences). It is however crucial to take into account that the logical apparatus based on deduction enters the picture not before we take into account the context of justification. The context of justification, however, being outside the scope of this paper.

Hence—to conclude—if the claims about analogy hold ground and I hope that they do, they vindicate both a pluralist view on the epistemology of mathematics and a thorough analogy between the epistemic paths in mathematics and in the natural sciences (given the descriptive epistemic context). Given that the underlying ontology is taken to be (a version of standard) Platonism, the presented mathematics-science epistemic analogy will hopefully offer a new perspective in the platonistic epistemology debate.

References

- Benacerraf, P. 1965. "What Numbers Could Not Be." *Philosophical Review* 74 (1): 47–73.
- Benacerraf, P. 1973. "Mathematical Truth." *Journal of Philosophy* 70 (19): 661–679.
- Brown, J. R. 1991. *The Laboratory of the Mind. Thought Experiments in the Natural Sciences*. New York: Routledge.
- Brown, J. R. 1999. *Philosophy of Mathematics. An Introduction to the World of Proofs and Pictures*. New York: Routledge.
- Franklin, A. and Perović, S. 2016. "Experiment in Physics." *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/win2016/entries/physics-experiment/>>.
- Kitcher, P. 2011. "Epistemology without history is blind." *Erkenntnis* 75 (3): 505–524.
- Pölya, G. 1945. *How to Solve It. A New Aspect of Mathematical Method*. Princeton: Princeton University Press.
- Shapiro, S. 1997. *Philosophy of Mathematics. Structure and Ontology*. Oxford: Oxford University Press.

The Function and Limit of Galileo's Falling Bodies Thought Experiment: Absolute Weight, Specific Weight and the Medium's Resistance

RAWAD EL SKAF
IHPST, Université Paris 1 Panthéon-Sorbonne, CNRS, Paris, France

The ongoing epistemological debate on scientific thought experiments (TEs) revolves, in part, around the now famous Galileo's falling bodies TE and how it could justify its conclusions. In this paper, I argue that the TE's function is misrepresented in this a-historical debate. I retrace the history of this TE and show that it constituted the first step in two general "argumentative strategies", excogitated by Galileo to defend two different theories of free-fall, in 1590's and then in the 1638. I analyse both argumentative strategies and argue that their function was to eliminate potential causal factors: the TE serving to eliminate absolute weight as a causal factor, while the subsequent arguments served to explore the effect of specific weight, with conflicting conclusions in 1590 and 1638. I will argue thorough the paper that the TE is best grasped when we analyse Galileo's restriction, in the TE's scenario and conclusion, to bodies of the same material or specific weight. Finally, I will draw out two implications for the debate on TEs.

Keywords: Scientific thought experiments, Galileo's falling bodies, *De Motu* (1590) and *Discorsi* (1638), eliminating causal factors, absolute and specific weights, medium's resistance.

1. Introduction

Galileo's *Discorsi* (1638) falling bodies TE has become a key case study in the epistemological literature on TEs, especially since Brown (1986) famously claimed that it is canonical case of what he labelled "platonic TE": it is both destructive and constructive. It is destructive since it refutes an old theory (*i.e.* Aristotle's theory of free-fall), it is also construc-

tive since it establishes, in *a priori* fashion, a new law of nature (*i.e.* in void, *all* bodies free-fall at the same speed). Brown's analysis was met by Norton (1996) who denied this "platonian" power of the TE and argued that it is reducible to a deductive argument, a TE-argument, that is an argument with *irrelevant* and even *eliminable* particulars. Both Norton and Brown agree that the TE perfectly leads to its destructive conclusion in a deductive manner. In addition, if the TE leads to its constructive conclusion, Norton claims that the TE-argument could deductively lead to this conclusion as well. However, Norton argues that the TE-argument shows us that the TE only leads to its constructive conclusion if we add the following hidden assumption 8a: the speed of a falling body depends *only* on its weight (see 4). Which for Norton amounts to assuming vacuum, something Galileo could not do in the context of the TE, and thus this constructive conclusion is "at worst, a fallacious inference to a falsehood [when assumption 8a does not hold]; or, at best, valid only insofar as it is invoked *in special cases* in which assumption 8a holds, such as the fall of *very heavy, compact objects in very rare media*. This final step now looks more like a clumsy fudge or a stumble than a leap into the Platonic world of laws." (Norton 1996: 345, my emphasis).

Norton's concluding remark, apart from being in tension with his "elimination thesis"¹, since he seems to grant some important role for the particulars involved in this TE, elicit the need to analyse the function of the particulars involved in Galileo's TE: very heavy, compact spherical objects of the same material falling in a rare medium such as air. More generally, the literature on TEs suffers from a major omission in analysing this TE: it does not tackle Galileo's restriction, in the TE's scenario and conclusion, to bodies of the *same material*. This omission is not proper to the Norton/Brown debate but is found in most of the literature. Of course, we find here and there some mention. For instance, Gendler underlines in a footnote the "somewhat unfortunate practice of considering this thought experiment outside of both its historical and textual contexts" (Gendler 1998: 402, ft 8). She then briefly mentions this restriction, however without analysing it, since she believes that for the purpose of her discussion "this constraint is irrelevant" (403, ft 13). Even if she rightly concludes that the TE's function is refutational and claims that she doesn't "think that the thought experiment in question shows anything more than that natural speed is independent of weight" (419), this restriction should not be left unanalysed if we want to understand the function and limit of Galileo's TE and its role in both argumentative strategies.

¹ "Thought experiments are arguments which contain particulars irrelevant to the generality of the conclusion. Thus any conclusion reached by a good thought experiment will also be demonstrable by an argument which does not contain these particulars and therefore is not a thought experiment" (Norton 1991: 131).

This restriction has even puzzled many Galilean scholars. For instance, at the end of his historical analysis of Galileo's arguments and TEs ("experiences imaginaires") since the *De Motu* (1590), Alexandre Koyré states that "Galileo's mention of specific gravity—and this, in a reasoning in which it has nothing to do—is extremely curious. And even, historically, very important."² More recently, Palmieri (2005) and Van Dyck (2006) analysed the historical development of this TE and its restriction to bodies of the same material, which brought Palmieri to conclude that "[p]erhaps we need a new approach to the question of thought experiment, capable of integrating results from different disciplinary areas, such as, for instance, the history and philosophy of science and cognitive science. [...] The all too clean baby of today's debate on the most beautiful thought experiment in the history of science [Galileo's TE] should definitely be thrown out, and the bathwater carefully analyzed." (Palmieri 2005: 238). Regrettably, this was not taken into account by most philosophers working on TEs. For instance, we still find in the Stanford entry on TEs (2017, substantively updated in 2014) and in Brown's second version of his book (2010), that "Galileo showed that *all* bodies fall at the same speed with a brilliant thought experiment" (my emphasis).

I am in total agreement with Palmieri that history of science should play a central, at least a much greater role in the philosophical debate on TEs. Indeed, we have the general impression that the epistemological literature on scientific TEs is mainly built on a-historical analysis of case studies. This is especially lamentable for Galileo's falling bodies because the epistemological literature takes this TE as a canonical case study, while the a-historical analysis of this TE yields wide disagreements about its conclusion(s), leading to divergences pertaining to its epistemic function. Thus, leading the epistemological literature on TE astray and turning an important debate into a red herring: the Norton/Brown debate on TEs revolves, in part, around how Galileo's TE justifies its conclusions, by direct a priori access to laws of nature or by being a deductive argument. Nevertheless, the TE's function is misrepresented as *revealing* and *justifying* a law of nature (Brown since 1986 and even in a sense in Norton's 1996 reply).

The philosophical literature is thus in need of a more careful historical analysis of Galileo's TE and the following questions answered, before trying to analyse if and how the TE could justify its conclusion(s): What is the TE's function (or intended conclusion) for Galileo? What is its role in Galileo (1590 and 1638)'s argumentative strategies? What is the function of the particulars involved in its scenario? What are the idealisations involved? Are these idealisations justified? Since vacuum could not be explicitly assumed in the TE and thus its scenario takes

² "La mention par Galilée de la gravité spécifique — et ce, dans un raisonnement où elle n'a que faire — est extrêmement curieuse. Et même, historiquement, très importante." (my emphasis and translation, Koyré 1960: 203).

place in plenum, then how did Galileo take into account the multiple effects of the medium's resistance? In case we assume vacuum (for modern readers), what conclusion could the TE lead to? Is this conclusion justified? All these questions could be easily answered once we tackle the more general one: *why is the TE restricted to bodies of the same material?*

This paper aims at analysing the function and limit of Galileo's falling bodies TE, which will provide answers to these questions. First in (2), I show that the TE's function is only refutational; it aims at refuting Aristotle's theory of free-fall, one of its two principles to be precise, by showing that the falling body's *absolute weight* could *not cause* divergences in the speed of free-falling bodies. I thus retrace Galileo's TE to its first occurrence in the *De Motu* (1590) which explicitly indicates Galileo's intention of "seeking causes of effects". Second in (3), I analyse Galileo's both argumentative strategies that led him to two incompatible theories of free-fall. It will be shown that the TE's restriction to bodies of the same material is best understood when placing the TE in both 1590 and 1638 argumentative strategies. I will argue that both strategies aimed at exploring potential causal factors affecting divergences in speed of free-falling bodies: the TE aimed at eliminating *absolute weight* as a causal factor, which explains Galileo's restriction to bodies of the same material, while both 1590 and 1638 subsequent arguments aimed at exploring *specific weight* as a causal factor, *with conflicting conclusions*. Third in (4), I analyse one small effect of the medium's resistance that could not be taken into account in the TE, even by Galileo's choice of particulars; *i.e.* the medium's disproportionate effect on the free-falling body's surface to absolute weight ratio. This shows that the TE only works either if we can assume vacuum or by placing the TE in the whole argumentative strategy, where this small effect of the medium's resistance is subsequently explained (which Galileo does in 1638) and thus could be ignored in the TE. Finally in (5), I summarize, draw out two implications for the debate on TEs and restate answers to the above questions.

2. *Absolute weight in the De Motu (1590) and the Discorsi (1638): same TE, same conclusion*

Galileo³ first introduced his TE in the *De Motu*, an unpublished manuscript usually dated from the 1590's. The TE appears in a larger argumentative strategy intended to first refute Aristotle's theory of free-fall and then defend Galileo's own early theory.

Galileo starts by clarifying the concepts of "heaviness" and "lightness". He stresses that both should be understood by what we could

³ Prior to Galileo, we find a similar TE in the work of Jean Baptiste Benedetti (1553) who imagines a scenario involving the fall of two equal bricks, by themselves and then attached (cf. Koyré 1960: 203).

call “specific weight” (even if Galileo is comparing equal volumes, not unit volumes of bodies), without explicitly defining the concept in the *De Motu*. Indeed, Galileo tells us that “a thing should be called heavier than another, if when a piece of it is taken, equal to a piece of the other, it is found to be heavier than the piece of the other” (Galileo 1590: 1).

Then Galileo distinguishes different ways in which “greater or lesser swiftness of [natural] motion comes about” (Galileo 1590: 14). This is best understood when we divide, following Galileo, Aristotle’s theory of free-fall into two principles⁴: (i) natural speed is proportional to weight, (ii) natural speed is inversely proportional to the medium’s resistance or “density”:

[I]nequalities in the slowness and swiftness of motion occur in two ways: for either the same mobile is moved in different media [*i.e.* according to Aristotle’s principle (ii), the speed of a mobile is inversely proportional to the medium’s resistance]; or the medium is the same, but the mobiles are different [*i.e.* according to Aristotle’s principle (i), the speed of the mobile is proportional to its weight]. We will demonstrate shortly that in both cases of motion the slowness and swiftness depend on the same *cause*, namely, *the greater or lesser heaviness [i.e. specific weight] of the media and of the mobiles*; but first we will show that *the cause of such an effect* which has been conveyed by Aristotle is *insufficient*. (Galileo 1590: 14, my emphasis⁵)

Galileo’s aim is to be found in this passage: he is seeking the causes of inequalities of slowness and swiftness of motion. He first aims at showing that the causes conveyed by Aristotle are either false—principle (i)—or insufficient—principle (ii). Galileo then aims to propose an early theory of free-fall, according to which the speed of a free-falling body is *proportional to its specific weight* (to be precise, minus the specific weight of the medium, see 3.1). This is how he proceeds.

Galileo starts by arguing against principle (ii) which states “that the cause of the slowness of motion is the thickness of the medium, and that of the speed, its subtlety” (p.14). Galileo aims at showing that this cause is insufficient, and he demonstrates this by appealing to examples where bodies, such as an inflated bladder, fall slowly downwards in air, but fly very swiftly upward when let go from deep in water.

Then Galileo moves to principle (i), which is the purpose of the TE.

⁴ This division will be again introduced in the *Discorsi*. For instance when Simplicio explains Aristotle’s argument against motion in void, he claims that Aristotle: “first supposes bodies of different weights to move in the same medium; then supposes, one and the same body to move in different media. In the first case, he supposes bodies of different weight to move in one and the same medium with different speeds which stand to one another in the same ratio as the weights; so that, for example, a body which is ten times as heavy as another will move ten times as rapidly as the other. In the second case he assumes that the speeds of one and the same body moving in different media are in inverse ratio to the densities of these media; thus, for instance, if the density of water were ten times that of air, the speed in air would be ten times greater than in water.” Galileo 1638/1914: [105–106] of the National Edition.

⁵ All emphasis in the subsequent quotes from the *De Motu* and the *Discorsi* are mine.

This principle aims at describing the speed of mobiles falling in the same medium. For these mobiles, Galileo further distinguish between two cases:

[D]ifference between two mobiles can happen in two ways: for *either they are of the same species*, as, for example, both lead, or both iron; and they differ in size: *or they are of different species*, e.g. one iron, the other wood; they then differ from one another either in size and heaviness, or in heaviness and not in size, or in size and not in heaviness. (Galileo 1590: 15)

This distinction is crucial for what follows. Galileo will first limit his arguments against Aristotle to the first case; to bodies of the same species that differ only in size. For these bodies the difference in size is directly translated into a difference in absolute weight and most—not all (see 4)—of the effects of the medium's resistance are the same. While for bodies of different species things are more complicated; they could differ in the three different ways enumerated above (see Fig.1). After his TE, Galileo will analyse bodies of different specific weights and bodies falling in different media simultaneously (see 3.1).

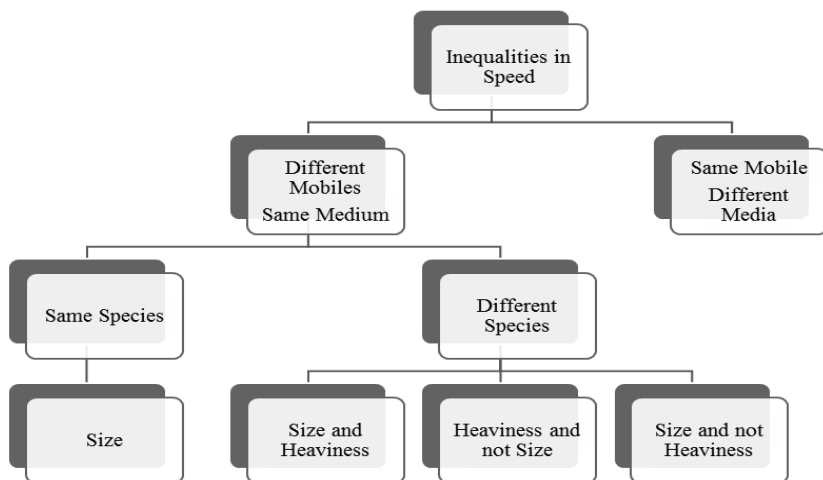


Fig.1:Galileo's *De Motu* analysis of the different ways inequalities in speed could come about

For bodies of the same species differing only in size (bottom left of Fig.1), Galileo starts by explicitly stating Aristotle's principle (i):

Concerning those mobiles that are of *the same species* Aristotle has said, that the larger is moved faster [...] Aristotle wants mobiles of the *same genus* to observe between themselves in the speed of motion the ratio of the sizes that these mobiles have: and he says that very openly [...], by affirming that a large piece of gold is carried more swiftly than a small one. (Galileo 1590: 15–16)

For bodies of the same species, principle (i) amounts to saying that the speed of a free-falling body is proportional to its volume or *absolute* weight. Galileo first dismisses this principle on empirical, or semi-empirical observations:

How ridiculous this opinion is, is clearer than daylight: for who will ever believe that if, for example, [...] from a high tower, two stones, one being double the size of the other, were thrown at the same moment, that, when the smaller was at mid-tower, the larger would already have reached the ground?" (Galileo 1590: 16)

Then Galileo moves away from empirical examples to several arguments, the last one being his famous TE. Galileo starts by explaining his preference to appealing to non-empirical arguments in "seeking the causes of effects":

[I]n order that we may always make more use of reasons than of examples (for we *are seeking the causes of effects, which are not reported by experience*), we will bring forth our way of thinking, whose confirmation will result in the downfall of Aristotle's opinion. (Galileo 1590: 16)

Galileo's entire thought process is somehow nested in two Archimedean analogies. The first concerns bodies floating on water, while the second concerns bodies heavier than water and thus sinking (see 3.1). Concerning the first Archimedean analogy, Galileo claims that the reason why bodies of the same species fall at the same speed is analogous to why a large beam and a small piece of wood float on water:

We say, then, that mobiles of the same species (let those things be said to be of the same species that are constituted of the same material, such as lead or wood, etc.), though they may differ in size, are however moved with the same swiftness, and a larger stone does not go down more swiftly than a smaller one. Those who are surprised by this conclusion will also be surprised that a very large beam can float on water, just as well as a small piece of wood: for the reasoning is the same. (Galileo 1590: 16–17)

Before introducing his TE, Galileo first proposes a three steps argument⁶ to explore this Archimedean analogy against Aristotle's principle (i):

In the first, Galileo invites us to think about the behaviour of a wooden beam and a stick of the same wood floating on the surface of the water. Galileo asks to imagine that the water's specific weight decreases to the point that it becomes lighter than the wood's. Then he asks, "who would ever say that the beam would go down first or more swiftly than the small piece of wood?" (Galileo 1590: 17).

In the second stage, Galileo reverses the strategy of the first. Instead of the medium's specific weight decreases, now the body's specific weight increases. He asks to imagine a volume of wax that is gradually filled with sand until the mixture's specific weight becomes bigger than the water's. Galileo then asks: "who would ever believe, if we took a particle of such wax, say one hundredth of it, either that it would not

⁶ Cf. Palmieri (2005: 226–227) for a similar analysis

go down or that it would go down a hundred times more slowly than the totality of the wax?" (Galileo 1590: 17)

These analogies show Galileo's emphasis on specific weight rather than absolute weight when analysing the speed of free-falling bodies. This is especially reflected in the third stage, where he explores the analogy between a balance and bodies floating on water:

And it will be possible to experience the same thing in the balance: for if very large, equal weights are placed on each side, and then to one of them something heavy, but only modestly so, is added, the heavier will then go down, but not any more swiftly than if the weights had been small. *And the same reasoning holds in water*: for the beam corresponds to one of the weights of the balance, while the other weight is represented by an amount of water as great in size as the size of the beam: if this amount of water weighs the same as the beam, then the beam will not go down; if the beam is made slightly heavier in such a way that it goes down, it will not go down more swiftly than a small piece of the same wood, which weighed the same as an [equally] small part of the water, and then was made slightly heavier. (Galileo 1590: 17)

This third step can be interpreted as an exploration of the analogy between the role of absolute weight on the balance⁷ and the role of specific weight in the floatability of bodies on water:

In the first case, the equilibrium of the balance is broken if one adds a weight on one arm of the balance in equilibrium. Whatever the material of these two weights or the added weight, what matters is the difference between the absolute weight that is already on the scale and the added weight. The mobile "falls", so to speak, on an arm of the balance when an extra weight is added. This speed of fall does not depend on the initial body's absolute weight, but on the difference between the absolute weight of the initial body and that of the added body.

In the second case when analysing bodies sinking in water, this difference must be understood in terms of specific weight. It is when one changes the body's specific weight that the equilibrium, which existed between the body and the water, is broken, and the floating body then sinks. It sinks with the same speed, whatever its volume or absolute weight, a beam or a stick of wood. The speed of fall does not depend on the initial body's absolute weight, but on the difference between the specific weight of the floating body and the specific weight of the added body. Galileo will indeed defend at the end of his argumentative strategy that the speed of a free-falling body is proportional to the specific weight difference between the mobile and the medium (see 3.1). But first, Galileo will argue against Aristotle's principle (i) with his TE.

⁷ Galileo will separate, in the *Discorsi*, from the idea that we could understand falling bodies by analogy to what happens in a balance, since bodies become weightless during their fall, a balance falling with a body cannot measure its weight. cf. Van Dyck 2006 for the analysis of the evolution of the role of the balance in Galileo's reasoning on free-fall.

2.1 *The TE in the De Motu*

Galileo introduces his TE as follows “[b]ut it is pleasing to confirm this by another argument”. Its scenario is almost the same as that of the *Discorsi*, five decades later. The difference is in Galileo’s justification of the following mediativity principle:

And first, let the following be presupposed: namely, if there are two mobiles, one of which is moved faster than the other, the combination of the two is moved more slowly than that part which was moved faster than the other, but more swiftly than the remaining part, which, alone, was carried more slowly than the other. (Galileo 1590: 17–18)

In the *Discorsi*, Galileo justifies this supposition with the following theoretical axiom “each falling body acquires a definite speed fixed by nature, a velocity which cannot be increased or diminished except by the use of force” (Galileo 1638/1914: [107]). In the *De Motu*, this supposition is justified by appealing to two examples taken from empirical observations: the first concerns two mobiles ascending in water⁸, while the second concerns two mobiles of different material falling in air:

[I]f [...] two mobiles go down, one of which is carried more slowly than the other, as, for example, if one is wood, the other a bladder, which go down in air, the wood more swiftly than the bladder, we presuppose this: if they are combined, the combination will go down more slowly than the wood alone, but more swiftly than the bladder alone. For it is manifest that the swiftness of the wood will be retarded by the slowness of the bladder, while the slowness of the bladder will be accelerated by the speed of the wood; and similarly a certain motion intermediate between the slowness of the bladder and the swiftness of the wood will result. (Galileo 1590: 18)

Note that this justification may seem to be out of place in the context of the TE, since Galileo considers, as in the version in the *Discorsi*, two bodies of the same material. Nevertheless, this mediativity supposition is not weakened: Galileo, through these examples, seems to give it an empirical, or semi-empirical justification resulting from our daily experience.

Galileo then aims, with the following TE’s scenario, at showing an inconsistency between this mediativity principle and Aristotle’s principle (i):

Let there be two mobiles *of the same species*, the larger a, and the smaller b; and, if it can be done, as our adversaries hold, let a be moved more swiftly than b. There are then two mobiles one of which is moved more swiftly than

⁸ “As, for example, if we understand two mobiles, such as a piece of wax and an inflated bladder, both of which are carried upward from deep water, but the wax more slowly than the bladder, we ask that it be conceded, that if they are combined, the combination will go up more slowly than the bladder alone, but more swiftly than the wax alone. Indeed this is very clear: for who doubts that the slowness of the wax will be diminished by the speed of the bladder, and, on the other hand, that the speed of the bladder will be retarded by the slowness of the wax, and that a certain motion intermediate between the slowness of the wax and the speed of the bladder will result?” (Galileo 1590: 18).

the other; hence, according to what has been presupposed, the combination of the two will be moved more slowly than a alone: but the combination of a and b is larger than a alone: hence, contrary to our adversaries' view, the larger mobile will be moved more slowly than the smaller; which would certainly be unsuitable. (Galileo 1590: 18)

That is in unfolding⁹ the TE's scenario according to these two principles, we arrive at an absurd result describing the composite body falling, at the same time, both faster and slower than the larger body a. Which brings Galileo to conclude:

Accordingly, let it be sufficiently confirmed that *there exists no cause*, per se, why mobiles of the *same species* should be moved with unequal speeds *but there certainly is one why they should be moved with equal speed*. But if there were some *accidental cause*, such as, for example, the shape of the mobile, it must not be classified amongst the causes per se. (Galileo 1590: 18)

This I submit is the function of the TE which is reflected in Galileo's own words in the *De Motu*: Galileo is isolating absolute weight in order to analyse it as a potential factor that could cause divergence in speed of free-falling bodies. He concludes, from his TE, that absolute weight could not be a causal factor and thus, contrary to Aristotle's principle (i), bodies of the same material do not fall proportionally to their absolute weight. Thus, bodies of the *same material* will fall with the same speed, if all "accidental" causes are accounted for. However, *the TE remains silent concerning the effects of other causal factors, in particular specific weight*.

I add that the TE's function will remain the same in the *Discorsi* (see 4), 5 decades later. The difference between these two occurrences of the same TE is to be found in the subsequent arguments that aimed at exploring *specific weight* as a potential causal factor, with *two conflicting conclusions*. To see that, let us compare how Galileo defended two incompatible theories of free-fall with two different argumentative strategies in 1590 and then in 1638.

3. *Specific weight in the De Motu (1590) and the Discorsi (1638): two argumentative strategies, two theories of free-fall*

In this section I aim at showing how Galileo defended two different theories of free-fall with different arguments that followed the same TE. I will show that this difference could be traced to Galileo's treatment of specific weight as a causal factor. I will first expose Galileo's *De Motu* second Archimedean analogy which led him to defend his early theory of free-fall: in void *all* bodies fall with a speed *proportional to*

⁹ Cf. El Skaf and Imbert (2013) for a defence of unfolding as a general task of science involved in several tools (computer simulations, real experiments and TEs). Cf. El Skaf Rawad (2016) ch.7 for an account of how TE reveal and resolve inconsistencies through a common structure which involves mentally unfolding TEs' scenarios.

their specific weights. Second, I will expose how in the *Discorsi*, Galileo eliminated specific weight as a causal factor to defend his final theory of free-fall: in void, *all* bodies of *any* material, fall at the same speed.

3.1 De Motu's Archimedean analogy: *specific weight is a causal factor*

The *De Motu* provides a very interesting manuscript to understand the evolution of Galileo's thought process, the limited function of his TE and his struggle with the causal role of specific weight, especially when compared with the *Discorsi's* argumentative strategy. In the *De Motu*, having eliminated absolute weight as a causal factor with the TE, by restricting its scenario to bodies of the same material, Galileo now wants to show that for bodies of different species, Aristotle's principle (i) is also false. First, Galileo—by building on the equality of speed for mobiles of the same species differing only in size—reduces his analysis of mobiles of *different species*, which differ in the three ways listed in Fig. 1, only to those differing “in heaviness and not in size”. He argues that “if the ratio of the motions of those mobiles that differ only in heaviness and not in size is given, the ratios of those that differ in any other way are also given” (Galileo 1590: 19). Then Galileo tackles both principles simultaneously:

And so, in order that we may find this ratio and, against Aristotle's way of thinking, show that in no way do mobiles observe the ratio of their heavinesses, *even if they are of different species*, [*i.e.* principle (i)] we will demonstrate things on which depends the answer not only to this investigation, but also to the investigation of the ratio of the motions of the same mobile in different media [*i.e.* principle (ii)]; and we will examine both questions simultaneously. (Galileo 1590: 19)

Galileo will examine both principle simultaneously with his second Archimedean analogy. Recall the first (see 2) Archimedean analogy concerned a large beam and a small piece of wood floating on water and Galileo used it to refute principle (i) for bodies of the same material. Galileo will build on the following second analogy to refute both principles and to defend his early theory of free-fall:

[A]ll these things will easily be drawn from the following demonstration. I say, then, that a solid magnitude heavier than water is carried downward with as much force as that by which a quantity of water, having a size equal to the size of the same magnitude, is lighter than this magnitude.” (Galileo 1590: 23)

Galileo provides several proofs of this latter claim (cf. Palmieri 2005: 229) and then concludes following this Archimedean analogy and contrary to principle (ii) that:

[T]he same mobile going down in different media, observes in the swiftness of its motions, *the ratio to one another of the excesses of its own heaviness over the heavinesses of the media*: thus if the heaviness of the mobile is 8, but the heaviness of a size of one medium, equal to that of the mobile, is 6,

then the swiftness of this body will be 2; if the heaviness of an amount of the other medium, equal to the size of the mobile, is 4, then the swiftness of the mobile, in this medium, will be 4. It is therefore evident that these swiftnesses will be to one another as 2 and 4; and not as the thicknesses or the heavinesses of the media, which is what Aristotle wanted, which are to one another as 6 and 4 (Galileo 1590: 24)

Galileo then applies the same reasoning to the fall of different bodies in the same medium, and concludes contrary to principle (i):

Similarly the answer to the other question is evident: namely, what ratio the speeds of mobiles equal in size, but unequal in heaviness, observe with one another in the same medium. For the speeds of such mobiles will be to one another as the *excesses by which the heavinesses of the mobiles exceed the heaviness of the medium*: thus, for example, if two mobiles are equal in size, but unequal in heaviness, the heaviness of one being 8, and of the other 6, but the heaviness of an amount of the medium, equal in size to the size of one of the two mobiles, is 4, then the swiftness of the former mobile will be 4, and that of the latter will be 2. Hence these speeds will observe the ratio of 4 to 2; and not that which is between the heavinesses, namely 8 to 6. (Galileo 1590: 24)

Put differently, Galileo's early theory describes the speed of free-falling bodies as an Archimedean ratio (a subtraction), not geometric (a division) as Aristotle's wanted: The speed is not W/R (Weight/Resistance), but proportional to $W_b - W_m$ (W_b and W_m being the specific weights of the body and the medium respectively)¹⁰. Thus, according to this early theory, in void where $W_m = 0$, mobiles fall proportionally to their specific weight:

Thus, in a void also *a mobile will be moved in the same way as in a plenum*. For in a plenum a mobile is moved swiftly according to the excess of its own heaviness over the heaviness of the medium through which it is moved; and thus *in a void it will be moved according to the excess of its heaviness over the heaviness of the void*: since this is null, the excess of the heaviness of the mobile over the heaviness of the void will be the total heaviness of this same mobile; *thus it will be moved swiftly according to its own total heaviness*. (Galileo 1590: 32)

That is in the *De Motu*, Galileo defends that specific weight is a causal factor affecting speed, *even in void*. In addition, this theory is consistent with the TE since specific weight is a mediative property. That is combining in the TE's scenario two bodies of different specific weights results in a body whose specific weight lies between the specific weights of the two constituent bodies. Which means that—according to the mediativity principle, but also according to the *De Motu's* theory—the combined body should fall at an intermediate speed. To see this, consider Gendler's reconstructed argument which aims at revealing a contradiction between 3 premises; *i.e.* (1) speed is mediative, (2) absolute weight is additive (3) natural speed is directly proportional to absolute weight. This shows a contradiction since a “mediative property cannot

¹⁰ Cf. Koyré (1960) and Van Dyck (2006) for a similar formula.

be directly proportional to one that is additive" (Gendler 1998, p.404). But we can rewrite the argument as follows without any contradiction: (1) speed is mediative, (2) specific weight is mediative (3) natural speed is directly proportional to specific weight. As we will see (3.2), in the *Discorsi* Galileo needed additional arguments to eliminate specific weight as a causal factor.

Finally, it should be noted that Galileo, immediately after defending the Archimedean ratios in plenum, notes that "a very great difficulty arises here: it will be found that these ratios are not observed by one who has made a test." However, without exploring this further, since he is convinced that "[i]t is necessary first to examine certain things which have not yet been inspected. For it is necessary, first, to see why natural motion is slower at the beginning." (Galileo 1590: 24)

3.2 *Discorsi's limiting case argument: specific weight is probably not a causal factor*

Five decades later, the same TE was reused by Galileo for the same purpose. However, the TE now constituted the first step of a more complex argumentative strategy which spans for 30 pages. Following the TE, Galileo will now propose two additional arguments, which will bring him to *eliminate specific weight as a causal factor* and to defend his final theory of free-fall: in void, *all* bodies fall at the same speed. This is how he argued. Following the TE, the second step consisted of a limiting case argument:

SALV. [...] in a medium of quicksilver, gold not merely sinks to the bottom more rapidly than lead but it is the only substance that will descend at all; all other metals and stones rise to the surface and float. On the other hand the variation of speed in air between balls of gold, lead, copper, porphyry, and other heavy materials is so slight that in a fall of 100 cubits a ball of gold would surely not outstrip one of copper by as much as four fingers. *Having observed this I came to the conclusion that in a medium totally devoid of resistance all bodies would fall with the same speed.* (Galileo 1638/1914: [116])

It is thus this "observation", not the TE, that brought Galileo to the conclusion that in void, *all* bodies would fall at the same speed. Having observed that the variation of speed, of bodies of different specific weights, becomes less and less important with the ratification of the medium, we could extrapolate what will happen at the limit in a medium totally devoid of resistance:

[...] if we find as a fact that the variation of speed among bodies of different specific gravities is less and less according as the medium becomes more and more yielding, and if finally in a medium of extreme tenuity, though not a perfect vacuum, we find that, in spite of great diversity of specific gravity [peso], the difference in speed is very small and almost inappreciable, then we are *justified in believing it highly probable* that in a vacuum all bodies would fall with the same speed. (Galileo 1638/1914: [117])

But, as it is made explicit by Galileo, we are justified in believing that in void *all* bodies fall at the same speed, only as “*highly probable*”¹¹. In fact, this is the case since this limiting case argument is also consistent with the *De Motu* early theory: at the limit, and therefore in a vacuum, the differences in speed between two falling bodies with different specific weights could only be *small*, not null. As shown in Palmieri’s diagrams (Fig.2), both Galilean theories predict divergences in the speed of falling bodies of different materials in plenum. In addition, both theories are consistent with this limiting case argument since they also predict that this divergence decreases with the ratification of the medium. However, in void, the two theories give two different predictions: in the diagram to the left, with the “restricted” *De Motu*’s theory, when the resistance of the medium becomes null, there will always be a small difference in the speed of falling bodies of different material. In void, the sphere made of gold falls faster than the one made of gold + silver, since specific weight is a causal factor according to the *De Motu*’s theory. While in the diagram on the right, with the “general” *Discorsi* theory, the speed of all falling bodies will be identical in void, since specific weight is no longer a causal factor.

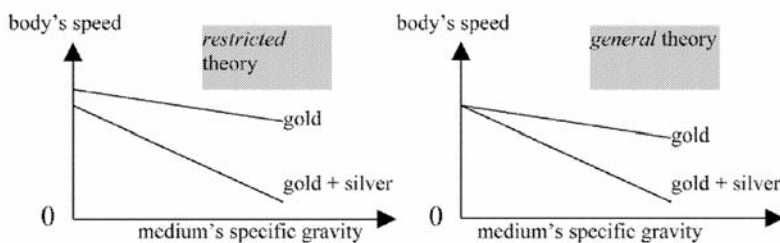


Fig.2: Inequalities in speed according to both theories
(Palmieri 2005: 234)

Galileo needed one additional step in his argumentative strategy in order to pass from “highly probable” to “confirming” his theory of free-fall. Which will provide him with a way to choose between his early and final conflicting theories, that is to make a theoretical choice. Galileo will provide an argument which aims at eliminating specific weight as a causal factor affecting the speed of free-falling bodies in void.

3.3 Discorsi’s *constant cause, constant effect argument or fall from small and high altitudes: specific weight is not a causal factor*

Galileo starts by setting up the stage for his analysis of different bodies falling from different altitudes. Salviati first raises and answers the following question:

¹¹ This is also explicit in Galileo’s *Postils to Rocco* (ca. 1634–1635) where he also uses the same TE and arguments as in the *Discorsi* (cf. Palmieri 2005: 232–233).

SALV. [...] Now, Simplicio, if we allow these two bodies [an inflated bladder and a mass of lead having the same size] to fall from a height of four or six cubits, by what distance do you imagine the lead will anticipate the bladder? You may be sure that the lead will not travel three times, or even twice, as swiftly as the bladder, although you would have made it move a thousand times as rapidly. (Galileo 1638/1914: [117])

To which Simplicio agrees, but adds that if they fall from a high altitude the difference will be bigger:

SIMP. It may be as you say during the first four or six cubits of the fall; but after the motion has continued a long while, I believe that the lead will have left the bladder behind not only six out of twelve parts of the distance but even eight or ten. (Galileo 1638/1914: [117])

Which will provide Salviati with the opportunity to analyse this divergence in speed of fall from different altitudes, all the while confirming that specific weight could not be a causal factor:

SALV. I quite agree with you and doubt not that, in very long distances, the lead might cover one hundred miles while the bladder was traversing one; but, my dear Simplicio, *this phenomenon* which you adduce against my proposition *is precisely the one which confirms it*. (Galileo 1638/1914: [117–118])

Galileo passes thus from “highly probable” following his limiting case argument, to “confirms” now. Here is how he argues with a *constant cause, constant effect argument*¹²:

SALV. [...] Let me once more explain that the variation of speed observed in bodies of different specific gravities *is not caused by the difference of specific gravity* but depends upon external circumstances and, in particular, upon the *resistance of the medium*, so that if this is removed all bodies would fall with the same velocity; and this result I deduce mainly from the fact which you have just admitted and which is very true, namely, that, in the case of bodies which differ widely in [specific] weight, their velocities differ more and more as the spaces traversed increase, *something which would not occur if the effect depended upon differences of specific gravity*. For since these specific gravities remain *constant*, the ratio between the distances traversed ought to remain *constant* whereas the fact is that this ratio keeps on increasing as the motion continues (Galileo 1638/1914: [118])

That is, when observing two bodies of different specific weights free-falling from a small and a high altitude, we realize that from a small altitude the difference in speed is so small that is barely observable, while from a high altitude the difference in their speed increases as the spaces traversed increase. Since from a *constant cause we should get a constant effect*, differences in specific weights, which remains constant, should cause the same variation of speed from small and high altitudes. Having observed that this variation of speed is not constant, but increases during the fall, we could conclude that differences in specific weights cannot cause this variation of speed, which should be caused by external factors; *i.e.* the resistance of the medium. Thus in void,

¹² Cf. Koyré (1960: 213) for a similar analysis

when the medium's resistance is removed, all bodies would fall with the same speed.

But Simplicio remains unpersuaded that this difference in speed should be caused by the medium's resistance, in such a way that if removed, all bodies would fall at the same speed. He will thus question the reason as to why the *same* medium produces *different* effects with the increase of the altitude of fall. Since the medium does not change as well, it should also produce a constant effect:

SIMP. Very well: but, *following your own line of argument*, if differences of weight in bodies of different specific gravities cannot produce a change in the ratio of their speeds, *on the ground that their specific gravities do not change*, how is it possible for the medium, *which also we suppose to remain constant*, to bring about any change in the ratio of these velocities? (Galileo 1638/1914: [118])

Which provides Galileo the opportunity to meet this "clever" objection by explaining how the effect of the medium's resistance increases with acceleration:

SALV. [...] There is [...] an increase in the resistance of the medium, not on account of any change in its essential properties, but on account of the change in rapidity with which it must yield and give way laterally to the passage of the falling body which is being constantly accelerated. (Galileo 1638/1914: [119])

That is, the medium's resistance is treated differently in the *De Motu* and the *Discorsi*: In the latter, the medium not only makes the body lighter as in the *De Motu*, it also has a frictional effect, which keeps on increasing until the falling body reaches its terminal velocity: "the speed [of the falling body] reaches such a point and the resistance of the medium becomes so great that, balancing each other, they prevent any further acceleration and reduce the motion of the body to one which is uniform and which will thereafter maintain a constant value." (Galileo 1638/1914: [118]).

The effect of specific weight is also treated differently in the *De Motu* and the *Discorsi*, in plenum and in void. In the former, the young Galileo defended that speed of fall is proportional to the specific weight difference between the mobile and the medium, which brought him to conclude that in void, where the medium's specific weight is null, speed is proportional to the mobile's specific weight. While in the *Discorsi*, the difference in speed that we observe in plenum for two free-falling bodies, of different specific weights, does not translate to a difference in speed in void, since:

[Specific] weight is the means employed by the falling body to open a path for itself and to push aside the parts of the medium, *something which does not happen in a vacuum* where, therefore, *no difference [in speed] is to be expected from a difference of specific gravity*. (Galileo 1638/1914: [118])

Put differently, while the *De Motu's* theory could be written as follows $V \sim W_b - W_m$ (see 3.1), the *Discorsi's* theory could be written as follows:

$V = V_0 [W_b - W_m]/W_b$ (W_b and W_m are as before the specific weights of the body and the medium, V_0 is the speed in void)¹³. Thus in void where $W_m = 0$, all bodies fall at the same speed V_0 .

Finally, Galileo will provide an empirical test: since measuring this variation in speed of two bodies falling from small heights was technically impossible at his time, and so “if there be a difference it will be inappreciable”, Galileo will propose to substitute these observations by observations on pendulums with equal bobs made from different material. In these experiments Galileo could “repeat many times the fall through a small height” in such a way that they become “not only observable, but easily observable” (Galileo 1638/1914: [128]). But this, as Palmieri notes, is a different story.

4. *Ignoring the medium's resistance without assuming vacuum*

We could now answer the question asked in the introduction, why is the TE restricted to bodies of the same material, as follows: Galileo, in restricting his TE's scenario to bodies of the same material, was able to isolate and eliminate absolute weight as a causal factor and to postpone his analysis of specific weight and the medium's resistance. That is, Galileo in his TE only addressed principle (i), without making any reference to the effects of the medium's resistance, which is described in Aristotle's principle (ii). In this section I aim to analyse if Galileo was justified in ignoring the medium's resistance, *without assuming vacuum*. In fact, this assumption which is usually legitimate if the context were different, could not be explicitly made by Galileo: the TE appears in a larger discussion concerning the existence of vacuum and the possibility of motion in void, which makes any explicit assumption of vacuum inadmissible in the TE.¹⁴

Atkinson and Peijnenburg (2004) set out to analyse the speed (acceleration and terminal velocity) of fall of bodies in different situations: bodies of different material falling in plenum and in void, from the same altitude and different altitudes, from small and high altitudes, etc. This analysis, even it is irrelevant to and consistent with Galileo's TE as analysed here (since I defend that Galileo's TE is only refutation-al¹⁵), remains interesting in its own right: it shows the complexity of

¹³ Cf. Koyré (1960) and Van Dyck (2006) for a similar formula.

¹⁴ Indeed, assuming vacuum at this point in the TE could invite the Aristotelian to reject the TE on the ground that void is impossible. cf. El Skaf (2016) ch.7 for an analysis of how TEs could fail and El Skaf (2017) for an analysis of the notion of possibility at play in TEs.

¹⁵ Which seems in line with Atkinson and Peijnenburg analysis as summarized in 2007: “As a destructive thought experiment, refuting the Aristotelian theory of falling bodies, we deem Galileo's thought experiment to be unparalleled, one of the best. But as a constructive thought experiment, claiming that all bodies fall at the same rate, it has a serious flaw. For it fails to make explicit a hidden assumption

taking into account all relevant causal factors—known (e.g. medium's resistance in plenum) or even unknown (e.g. inhomogeneous gravitational field of the earth, even in void) by Galileo—that could affect the speed of free-falling bodies.

Some interesting parts of this analysis, which are directly related to Norton's quote in the introduction concerning Galileo's "hidden" assumption 8a—*i.e.* speed of fall depends *only* on the body's weight –, are in fact explicitly addressed by Galileo in the *Discorsi*. This is first reflected in Galileo's choice of particulars involved in his TE's scenario—bodies made of the same material and have the same shape differing only in size—and second by Galileo's analysis of a small effect of the medium's resistance affecting even these particulars.

First, directly after the TE, Galileo underlines that:

Aristotle declares that bodies of different weights, in the same medium, travel (in so far as their motion depends upon gravity) with speeds which are proportional to their weights; this he illustrates *by use of bodies in which it is possible to perceive the pure and unadulterated effect of gravity [i.e. absolute weight]*, eliminating other considerations [...] which are greatly dependent upon *the medium which modifies the single effect of gravity alone* (Galileo 1638/1914: [109]) (Galileo 1638/1914: [109])

Put differently, Galileo is making Norton's hidden assumption 8a *but without assuming vacuum*. By his choice of particulars Galileo is considering a situation, like Aristotle did, in which absolute weight is the only causal factor and it is possible to perceive its pure and unadulterated effect. But Galileo is not making this assumption to argue from his TE that *all* bodies fall at the same speed, as in Norton's TE-argument (Norton 1996: 341–343)—since even assuming 8a, bodies could fall proportionally to their specific weights as we have seen in (3.1) –, but to show that absolute weight could not be a causal factor, contrary to Aristotle's principle (i) or any other theory linking differences in speed to differences in absolute weight.

Second and more subtly, Galileo knew that even for these bodies, *most*, not all of the effects of the medium's resistance could be taken into account in his TE. Most, not all since one small effect of the medium's resistance remains disproportional for larger and smaller bodies. Indeed, just before the above quote, Galileo makes reference to this small effect when he claims that “[y]ou find, on making the experiment, that the larger outstrips the smaller by two finger-breadths” and then dismisses this difference on the account that Simplicio would “not hide behind these two fingers the ninety-nine cubits of Aristotle”.

Finally and most importantly, at the end of his argumentative strategy Galileo comes back to this small effect and sets out “to explain how one and the same medium produces such different retardations in bodies *which are made of the same material and have the same shape, but dif-*

that is not always applicable, namely that the rate of fall of a body depends only on its weight, and on nothing else.” (207)

fer only in size" (Galileo 1638/1914: [132]). For Galileo this explanation "requires a discussion more clever than" the previous explanations of the different effects of the medium's resistance. Galileo's solution lies in:

[T]he roughness and porosity which are generally and almost necessarily found in the surfaces of solid bodies. [...] in the motion of falling bodies these rugosities strike the surrounding fluid and retard the speed; and this they do so much *the more in proportion as the surface is larger, which is the case of small bodies as compared with greater.* (Galileo 1638/1914: [132])

That is, the medium affects disproportionately even the two falling mobiles involved in the TE's scenario. The medium's resistance is more important the bigger the mobile's surface to absolute weight ratio is. The medium thus affects less the speed of fall of the larger mobile than that of the smaller one, since the former have a smaller surface to absolute weight ratio than the latter (for which Galileo provides a geometrical proof, Galileo 1638/1914: [133–134]). The larger mobile will be less retarded by the medium and thus will have a greater speed. Which means that, in the context of the TE where Galileo is in no position to assume vacuum, the larger mobile falls faster than the smaller one. If we take this effect into account in the TE's scenario, then it is hard to see how even the destructive conclusion could be obtained. Galileo thus needed to ignore this small effect of the medium's resistance in order to refute Aristotle's principle (i).

If we don't ignore this effect of the medium's resistance, then how should we analyse the TE. One option, which rather complicates things, is to analyse how the two bodies are tied together: are they merely united or smelted together. This point has already been raised by Gendler (1998) and Atkinson and Peijnenburg (2004). In analysing Norton's hidden assumption 8a and Gendler's reconstructed argument, Atkinson and Peijnenburg correctly explain that if we take into account this small effect of the medium's resistance, then things get complicated, since "two lead spheres of different weights (and therefore with different volumes), will have different terminal velocities. If they are tied together side-by-side, the terminal velocity of the united system will lie between the terminal velocities of the constituents [... and Galileo is justified in refuting Aristotle's principle (i)]. If, on the other hand, the spheres are melted and recast as one sphere of weight equal to the sum of the weights of the two original spheres, then the terminal velocity of the united system will be greater than those of either of the constituents. The reason [as we saw Galileo was aware] is that the retarding viscous force is a function of both the velocity and of the surface area of the falling body. The smelted sphere falls more quickly than the united spheres because the surface of the former is smaller than the combined surfaces of the latter." (p. 123) In this latter case, Galileo is no longer able to refute Aristotle's principle (i), since the smelted body falls faster than its constituents. That is, the mediativity principle no longer applies for smelted bodies.

I submit that there is no need to complicate the TE's analysis: Galileo is in a position to ignore this small effect of the medium's resistance since the TE appears in a larger argumentative strategy in which Galileo comes back to this effect and explains it.

5. Conclusion

Let us summarize and conclude. In this paper I aimed at clarifying the reasons behind Galileo's restriction, in the TE's scenario and conclusion, to bodies made of the same material. This restriction turned out to be of central importance to understanding the function and limit of the TE. I retraced the history of this TE to its first occurrence in the *De Motu* and showed that the TE is only refutational; it aimed at refuting Aristotle's principle (i) by showing that absolute weight could not be a causal factor.

I then exposed how Galileo, following the same TE, defended two incompatible theories of free fall and I argued that both theories could be traced to Galileo's analysis of specific weight: following a hasty Archimedean analogy, the young Galileo maintained specific weight as a causal factor and defended an early theory of free-fall according to which speed is proportional to specific weight, even in void. Five decades later and with two new arguments, Galileo eliminated specific weight as a causal factor and defended his final theory of free-fall according to which in void, *all* bodies fall at the same speed.

I finally showed that Galileo, in the TE, needed to ignore one small effect of the medium's resistance affecting even the kind of particulars involved in the TE's scenario: bodies made of the same material and having the same shape, differing only in size. This effect either complicates the TE if it is taken into account, or Galileo is justified in ignoring it only when we analyse the TE as part of a general argumentative strategy in which Galileo comes back to this small effect and explains it.

In conclusion, we could draw out from this historical analysis at least the following two implications for the debate on TEs.

First contra Norton's "irrelevant particulars" and elimination thesis (see ft 1), in appraising TEs it seems crucial to analyse the function(s) of some of the details involved in their scenarios, instead of trying to eliminate them. Indeed, some particulars involved in a TE's scenario have a crucial function. In Galileo's TE, they permit to isolate the effect of absolute weight *without assuming vacuum*. Of course some particulars are irrelevant, for instance the two falling bodies could be yellow or blue, weight 8 and 4 kg or 12 and 6 kg, however they should be made of the same material and have the same shape, if not absolute weight could no longer be isolated as a causal factor. In addition, following the TE and Norton's quote in the introduction, the equality in speed only applies to "special cases", that is to the kinds of particulars involved in the TE's scenario for which we could ignore the effect of the medium's resistance.

Which is not the case for *any* two free-falling bodies, *even* of the same material. For instance, for bodies having a different shape, such as a nugget of gold and a leaf of gold (example given by Galileo), the medium's resistance could no longer be ignored without assuming vacuum.

Second contra Brown, it is clear that Galileo's TE could not, and even did not reveal, and a fortiori justify a law of nature—*i.e.* in void, *all* bodies fall at the same speed—platonically or otherwise. For the simple reason that following the TE, and even if we assume vacuum, we still don't know if bodies fall proportionally to their specific weight (1590) or not (1638). The TE, restricted to bodies of the same material having the same shape, allows Galileo to isolate and eliminate absolute weight as a causal factor, but remains silent concerning other causal factors, in particular specific weight (see 3). At most, the TE could lead to a weaker reading of the law; *i.e.* *all* bodies *of the same material* fall at the same speed. However, this conclusion follows if the medium's resistance is the only remaining causal factor (see answer to the last question below) and if we are justified in idealizing *all* of its effects, in such a way that we could say, following the TE, that a nugget of gold falls at the same speed than a leaf of gold, which amounts to assuming vacuum.

Finally, answers to the questions formulated in the introduction are found explicitly throughout the paper, let me restate them briefly here for clarity: What is the TE's function (or intended conclusion) for Galileo? To show that absolute weight is not a causal factor, thus refuting Aristotle's principle (i). What is its role in Galileo (1590 and 1638)'s argumentative strategies? To eliminate absolute weight as a causal factor, thus paving the way to analyse specific weight and the medium's resistance. What is the function of the particulars involved in its scenario? To isolate absolute weight in order to eliminate it as a causal factor. What are the idealisations involved? Void is not explicitly assumed, but one small effect of the medium's resistance is ignored. Are these idealisations justified? Yes, if we analyse the TE in the general strategy where this small effect is subsequently explained. Since vacuum could not be explicitly assumed in the TE and thus its scenario takes place in plenum, then how did Galileo take into account the multiple effects of the medium's resistance? Most of them were taken into account by Galileo's choice of particulars, one small effect was ignored but then later explained. In case we assume vacuum (for modern readers), what constructive conclusion could the TE lead to? At most, the TE could lead to the following: *in void*, *all* bodies *of the same material* fall at the same speed. Is this conclusion justified? Yes, if there are no remaining causal factors. But as shown in Atkinson and Peijnenburg (2004: 124–125) and unbeknown to Galileo, different causal factors could affect speed, even in void. For instance, the earth inhomogeneous gravitational field affects disproportionately the acceleration of bodies of the same material when they are dropped from different heights.

References

- Atkinson, D., and Peijnenburg, J. 2004. "Galileo and prior philosophy." *Studies in History and Philosophy of Science* 35: 115–136.
- Atkinson, D., and Peijnenburg, J. 2007. "On poor and not so poor thought experiments. A reply to Daniel Cohnitz." *Journal for General Philosophy of Science* 38: 159–161.
- Brown, J. R. 1986. "Thought Experiments since the Scientific Revolution." *International Studies in the Philosophy of Science* 1: 1–15.
- Brown, J. R. 1991a [2010]. *Laboratory of the Mind: Thought Experiments in the Natural Sciences*. [2nd edition] London: Routledge, Second Edition.
- Brown, J. R. and Fehige, Y. 2017. "Thought Experiments." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/sum2017/entries/thought-experiment/>>.
- El Skaf, R. and Imbert, C. 2013. "Unfolding in the empirical sciences: experiments, thought experiments and computer simulations", *Synthese* 190: 3451–3474.
- El Skaf, R. 2016. *La structure des expériences de pensée scientifiques*. Doctoral dissertation. Université Paris 1 Panthéon-Sorbonne.
- El Skaf, R. 2017. "What notion of possibility should we use in assessing scientific thought experiments?" *Lato Sensu* 4 (1): 19–30.
- Galileo, G. 1590. *De Motu Antiquiora*. E-version http://echo.mpiwg-berlin.mpg.de/content/scientific_revolution/galileo/englishttranslation.
- Galileo, G. 1638/1914. *Dialogues Concerning Two New Sciences by Galileo Galilei*. Translated from the Italian and Latin into English by Henry Crew and Alfonso de Salvio. With an Introduction by Antonio Favaro. New York: Macmillan.
- Gendler, T. S., 1998. "Galileo and the Indispensability of Scientific Thought Experiment." *The British Journal for the Philosophy of Science* 49: 397–424.
- Koyré, A. 1960. "Le *De Motu Graviorum* de Galilée. De l'expérience imaginaire et de son abus." *Revue d'histoire des sciences et de leurs applications* 13 (3): 197–245.
- Norton, J. D. 1991. "Thought experiments in Einstein's Work", in T. Horowitz and G. Massey (eds.), *TEs in Science and Philosophy*, Lanham: Rowman & Littlefield, 129–148.
- Norton, J. D. 1996. "Are Thought Experiments Just What You Thought?" *Canadian Journal of Philosophy* 26: 333–366.
- Palmieri, P. 2005. "Spuntur lo scoglio più duro: did Galileo ever think the most beautiful thought experiment in the history of science?" *Studies in History and Philosophy of Science* 36: 305–322.
- Van Dyck, M. 2006. *An Archeology of Galileo's science of motion*. Doctoral dissertation. University of Gent.

Is the Antipathetic Fallacy Responsible for the Intuition that Consciousness is Distinct from the Physical?

FRANÇOIS KAMMERER*

*Ecole Normale Supérieure, PSL Research University,
Institut Jean Nicod (ENS/EHESS/CNRS), Paris, France*

Numerous philosophers have recently tried to defend physicalism regarding phenomenal consciousness against dualist intuitions, by explaining the existence of dualist intuitions within a purely physicalist framework. David Papineau, for example, suggested that certain peculiar features of some of our concepts of phenomenal experiences (the so-called “phenomenal concepts”) led us to commit what he called the “Antipathetic Fallacy”: they gave us the erroneous impression that phenomenal experiences must be distinct from purely physical states (the “intuition of distinctness”), even though they are not. Papineau’s hypothesis has been accepted, though under other names and in different forms, by many physicalist philosophers. Pär Sundström has tried to argue against Papineau’s account of the intuition of distinctness by showing that it was subjected to counterexamples. However, Papineau managed to show that Sundström’s counterexamples were not compelling, and that they could be answered within his framework. In this paper, I want to draw inspiration from Sundström, and to put forth some refined counterexamples to Papineau’s account, which cannot be answered in the same way as Sundström’s. My conclusion is that we cannot explain the intuition of distinctness as the result of a kind of “Antipathetic Fallacy”.

Keywords: Consciousness, dualism, physicalism, introspection, concepts, intuition.

Introduction

Many philosophers recognize that phenomenal consciousness seems to pose a metaphysical problem. On the one hand, we have various

* I would like to thank Samuel Webb and Joseph Levine for their comments and their help. I acknowledge the financial support of the two following grants: ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*

reasons to suppose that physicalism is true. Physicalism is the thesis that phenomenal states are fully identical with physical states (broadly construed, as to include physically realized functional states), such as brain states. On the other hand, the identity of phenomenal states and physical states appears very counter-intuitive. This is rendered manifest when we focus introspectively on one of our current experiences: how can *this* (say, this sensation of pain) be *the same thing* as some electrochemical activity that takes place in my brain?

Some dualist philosophers have argued that this intuition, once elaborated and transformed into arguments, simply shows that phenomenal states are *really* distinct from physical states. Others have tried to defend physicalism against this intuition, by giving an explanation of this intuition in a purely physicalist framework. For example, they have tried to show that this intuition is a by-product of certain (purely physical) features of some of our *concepts* of phenomenal states – concepts of phenomenal states which are notably applied through introspection, and which are called “phenomenal concepts”.

One particular line of thought has emerged as especially popular: some philosophers, such as David Papineau (Papineau 1993, 2002, 2007), have tried to explain this dualist intuition (which Papineau labelled the “intuition of distinctness”) as being the result of a peculiar feature of phenomenal concepts. These concepts, according to Papineau, display a “use/mention feature”: whenever a subject uses them, she tends to *activate* the very experience thought about *via* this concept, or at least a “faint copy” of this experience. For this reason, when we think about phenomenal experiences *qua* experiences, *i.e.* with phenomenal concepts, our thought has a distinctive feeling, which it has not when we think about the same states *qua* brain states, using purely physical concepts. We then succumb to a fallacy that Papineau calls the “Antipathetic Fallacy”, when we infer that this phenomenological difference between the two *thoughts* indicates that the *thing* thought about with a phenomenal concept *must be itself different* from the thing thought about with a purely physical concept.

Pär Sundström (Sundström 2008) addressed an objection to this physicalist account of the intuition of distinctness. He tried to show, on the basis of an imaginary counterexample, that this account cannot be correct, as it makes *false predictions*. It predicts that an intuition of distinctness should arise in a case in which it obviously doesn't. David Papineau answered this objection, by showing that it was possible to reinterpret Sundström's counterexample in order to make it harmless for his own account. In this paper, I want to draw inspiration from Sundström. I will formulate new counterexamples, inspired by Sundström's, which I think cannot be answered in the same way. I will then use these counterexamples to make a case for the idea that we cannot explain the intuition of distinctness as resulting from the “Antipathetic Fallacy” described by Papineau.

In a first section, I will explain how, in Papineau's theory, the antipathetic fallacy is supposed to account for the existence of the intuition of distinctness in a physicalist framework. In a second section, I will expose Sundström's criticism, as well as Papineau's answer to this criticism. In a third section, I will present two thought experiments, one of which is a clear counterexample to Papineau's account but cannot be answered in the same way as Sundström's objection. In a fourth section, I will present more thought experiments—with the aim of showing that the intuition of distinctness really has little to do with a hypothetical "use/mention" feature of phenomenal concepts. In a fifth section, I will consider one possible response to my argument, and I will try to counter it. The sixth section will be devoted to concluding remarks

1. *The antipathetic fallacy and the intuition of distinctness*

Phenomenal states are states such that *there is something it is like* to be in these states. A headache, a visual sensation of red, an olfactory sensation of honeysuckle, are typical examples of phenomenal states. These states are said to be endowed with *phenomenal properties*, which are properties in virtue of which these states are such that there is something it is like to be in them, and which are properties that determine *what it is like* to be in these states. *Being a visual sensation of red*, for example, that we can also label "phenomenal redness", is a typical example of a phenomenal property.

We have numerous reasons to think that these properties must be wholly identical with physical properties—broadly construed, as to include physically realized functional properties.¹ However, we are often deeply puzzled when we focus on our phenomenal states, and when we then try to think that they are fully identical with physical states, merely endowed with physical properties. How can *this* (thought by focusing, say, on a current visual sensation of red) be *the same thing* as a certain electrochemical activity in my visual cortex? Many of us, even convinced physicalists, admit that it seems to be a mystery. It has been said that in this kind of situation we face an *explanatory gap* (Levine 1983, 2001). David Papineau described this situation by saying that, in these cases, we encounter a strong *intuition of distinctness* (Papineau 2002): the intuition that our phenomenal states *are not* identical with physical states,² but truly are distinct from them. This explanatory gap

¹ These reasons have generally mostly to do with causal considerations (Levine 2001: Chapter 1; Papineau 2002, Chapter 1). I won't expound them here, as my goal is not to argue in favor of physicalism.

² I take these two descriptions of the issue to be roughly equivalent. This is confirmed by Levine's own words: "Whether we think of [the explanatory gap] as an explanatory gap or a distinctness gap, the problem is really the same" (Levine 2007: 148). Also see (Papineau 2011) for the idea that the explanatory gap has to be interpreted as constituted by the intuition of distinctness.

(or intuition of distinctness) fuels, in one way or in another, many anti-physicalists arguments regarding phenomenal consciousness (Chalmers 1996; Jackson 1982; Kripke 1980).

Some physicalists have suggested that it is possible to defend physicalism against this intuition, and against the arguments that it supports, by providing an *explanation* of this intuition within a physicalist framework. This explanation is supposed to rely on certain special features of some of the *concepts* we use to think about our phenomenal experiences. These concepts are called “phenomenal concepts”, and they are the concepts we notably (but not only) use when we focus introspectively on our phenomenal experiences. In this view, phenomenal states indeed *seem* distinct from physical states. However, this happens merely in virtue of some features of the way in which we *think* about phenomenal states—features which are themselves purely physical. And, from a metaphysical point of view, phenomenal states really are identical with physical states. This kind of defense of physicalism has been labelled the “Phenomenal Concept Strategy” (Stoljar 2005). Numerous versions of this Strategy have been developed in the recent years (Aydede and Güzeldere 2005; Balog 2012; Hill 1997; Levin 2007; Loar 1997; Papineau 2002; Sturgeon 1994; Tye 1999).

Many of the theories belonging to the Phenomenal Concept Strategy have a common way to tackle the intuition of distinctness. They interpret it as a result of what David Papineau called the “Antipathetic Fallacy” (Papineau 1993, 2002, 2011). According to Papineau indeed, the intuition of distinctness arises because of a special feature of phenomenal concepts. Phenomenal concepts present a “use/mention feature”: each occurrence of a given phenomenal concept involves the *instantiation* of the phenomenal property (identical to a physical property) this concept refers to, or at least of a property resembling it. This means that every time I think about a type of phenomenal experience using phenomenal concepts, I crucially activate a version of this experience, or at least what Papineau calls a “faint copy” of this experience. Phenomenal concepts, in this view, are peculiar because they *make use* of the property they *mention*.³

Why does this feature give rise to an intuition of distinctness? The explanation, according to Papineau, goes as follows. When we try to consider that a given phenomenal state (say, a visual experience of red) and a given physical state are *identical*, we make use of two different concepts. The first of them, being a phenomenal concept, brings the instantiation of phenomenal redness whenever we use it, while the other does not. Therefore, the phenomenal way of thinking about this property has itself a distinctive “feeling”: *it is like* something to think about phenomenal redness with a phenomenal concept. On the other hand, there is no distinctive feeling when I think about one of my brain

³ The details of this theory have changed over the years in Papineau’s work (Papineau 1993, 2002, 2007), but the general idea has remained the same.

states using physical concepts. So, according to Papineau, “there is an intuitive sense in which exercises of material concepts ‘leave out’ the experience at issue. They ‘leave out’ [...] the technicolour phenomenology, in the sense that they don’t activate or involve these experiences” (Papineau 2002: 170). And this is where we commit what Papineau calls the “Antipathetic Fallacy”: we can’t help thinking that the fact that our physical conception “leaves out” something when compared with our phenomenal conception shows that these two conceptions simply *are not* about the same thing. This is why it seems to us that phenomenal states and physical states are distinct; this is how we get the intuition of distinctness.

Although David Papineau has been an early and a forceful defender of this kind of explanation, he is not the only one who has proposed something in the vicinity. An explanation of this type can indeed be found in the work of numerous philosophers proponents of the Phenomenal Concept Strategy. Brian Loar, one of the other main defenders of this Strategy, writes:

A phenomenal concept exercised in the absence of the phenomenal quality it stands for often involves, not merely a recognitional disposition, but also an image. And so, as a psychological state in its own right, a phenomenal concept—given its intimate connection with imaging—bears a phenomenological affinity to a phenomenal state that neither state bears to the entertaining of a physical-theoretical concept. When we then bring phenomenal and physical-theoretical concepts together in our philosophical ruminations, those cognitive states are phenomenologically so different that the illusion may be created that their references must be different. (Loar 1997: 605)

Even if this feature is not the only feature that is supposed to account for the explanatory gap in Loar’s account, it still plays an important role. Besides, the Antipathetic Fallacy, though not by this name, also plays a role in Michael Tye’s and Katalin Balog’s theories of phenomenal concepts (Balog 2012: 30–31; Tye 1999: 712–713). For reasons of simplicity, I suggest to call the hypothesis according to which the Antipathetic Fallacy (or something roughly equivalent) explains the birth of the intuition of distinctness the “Antipathetic Fallacy Hypothesis”. I think that it is safe to say that this hypothesis constitutes one of the major lines of thought developed by proponents of the Phenomenal Concept Strategy in order to account for the explanatory gap in a physicalist framework.⁴

⁴ The other aspect that proponents of the Phenomenal Concept Strategy usually insist upon is the conceptual independence of phenomenal concepts and physical concepts, which cause an absence of conceptual derivation from physical truths to phenomenal truths. Papineau does not insist upon this trait in his theory, however, as he does not think that the explanatory gap is primarily a matter of lack of conceptual derivation (Papineau 2011). I tend to agree with him as well as with Joseph Levine (Levine 2001, 2007: 200) on this point, even though I won’t talk about it here.

2. Sundström's counterexample and Papineau's response

The Antipathetic Fallacy Hypothesis has been subjected to many criticisms. One of them, that I find quite compelling because it does not bear on many theoretical assumptions, relies on counterexamples. It has been developed by Pär Sundström (Sundström 2008).

The general idea of Sundström's criticism, as I understand it, can be exposed as follows. Let's accept that, as the Antipathetic Fallacy Hypothesis says, the use/mention feature of phenomenal concepts causes the intuition of distinctness.⁵ If this is the case, then we should expect that, whenever we consider an identity statement of a certain kind (which I will describe in detail), an intuition of distinctness arises.

The relevant identity statements are statements which relate two conceptions of the same phenomenal property (identical with a physical property, given that the hypothesis is physicalist), with only one conception being systematically accompanied by the instantiation of this very phenomenal property. Let's call statements of this kind "phenomenologically contrasted identity statements". So, if the Antipathetic Fallacy Hypothesis is true, whenever we consider phenomenologically contrasted identity statements, we should have an intuition of distinctness concerning the two things identified in the statement.

Sundström then shows that there are cases that can intuitively count as counterexamples to this prediction: cases in which we *do* consider phenomenologically contrasted identity statements and yet *do not* have an intuition of distinctness. Sundström puts forth two examples of this kind. The first one essentially relies on some particular details of Papineau's account of phenomenal concepts, and notably Papineau's hypothesis that there are "derived" phenomenal concepts.⁶ The second counterexample seems to me to be more compelling, as it does not bear on any specifics of the targeted theory, and therefore could apply to *any* theory that tries to explain the intuition of distinctness in a similar

⁵ Papineau seemed at first to imply that the Antipathetic Fallacy was the only cause of the intuition of distinctness, but he later explicitly stated that it was likely to be just *one* cause of this intuition amongst others (Papineau 2011: 17–19).

⁶ Roughly: Papineau says that there are, aside from "full-blown" phenomenal concepts, *derived phenomenal concepts* (Papineau 2007: 127–128). They are mental representations that are informationally deeply connected to "genuine" phenomenal concepts, so that they can refer to the same property, but whose instantiations do not necessitate the instantiation of the phenomenal property referred to (these concepts are required in order for me to be able to think thoughts such that: "I am not having an experience of this kind right now"). Sundström then builds a counterexample to Papineau's theory, crucially using these concepts. In a nutshell, his counterexample goes like this (Sundström 2008: 141): he notes that any identity statements relating a derived phenomenal concept and a genuine phenomenal concept will constitute what I called a phenomenologically contrasted identity statement, and then should cause an intuition of distinctness. However, according to him, this is obviously not the case.

manner as Papineau's—even one which is not committed to the existence of “derived” phenomenal concepts. For this reason, I will focus on this particular counterexample.⁷

The counterexample goes like this (Sundström 2008: 141–142). Consider an identity statement such as “My brother's most salient current experience = an experience of white”. Let's say that the second half of the identity statement is thought while focusing on my current experience of the whiteness of the background of my Word document. On the other hand, if we consider the first half of the identity statement, it seems that it can be thought without any instantiation of phenomenal whiteness. After all, I can think about my brother's most salient current experience without having in mind a particular experience—even without knowing what kind of experience it is. Therefore, this identity statement is a *phenomenologically contrasted identity statement*. If the Antipathetic Fallacy Hypothesis is true, an intuition of distinctness should arise. But, according to Sundström, it is obviously not the case. I have no trouble entertaining the hypothesis that my brother's most salient current experience is an experience of white. I am in no way puzzled by this statement—while I am puzzled when I think that an experience of white *is* a certain neural activation in my sensory cortex. So, this counterexample seems to show that the Antipathetic Fallacy Hypothesis makes false predictions, and should therefore be abandoned.

Papineau later responded to this counterexample (Papineau 2011: 16–17). Acknowledging that Sundström's point is “well-taken”, he seemed to agree with most of the premises of Sundström's objection. He notably seemed to accept that the Antipathetic Fallacy Hypothesis predicts that, when we face a phenomenologically contrasted identity statement, an intuition of distinctness should arise. He also recognized that we do not face such an intuition when we consider the identity statement: “My brother's most salient current experience = an experience of white”.

His defense strategy against Sundström's objection amounted to arguing that this identity statement is *not* a phenomenologically contrasted identity statement after all. Maybe, he says, we “tend surreptitiously to activate the experience” of white when we think the first half of the identity statement. Or maybe we *don't* activate the experience of white when we think the second half of the identity statement—for example, because we are making use of a derived phenomenal concept instead of a “genuine” phenomenal concept in order to think about the experience of white. We just have to stipulate that one of these two possibilities is the case in order for the Antipathetic Fallacy Hypothesis to be protected against Sundström's objection.

⁷ I also find it more interesting to focus on this counterexample as it is the only one (to the best of my knowledge) which has received an explicit response from Papineau.

Is this defense successful? I think that it can partially succeed, as one of the two possibilities described by Papineau could indeed be the case. It *could be* that, when I think about “My brother’s most salient current experience” (and then try to equate it to an experience of white), I “tend surreptitiously to activate” an experience of white. Nothing, in Sundström’s description of this situation, can guarantee that this is not the case. As for the other possibility, I don’t think (*pace* Papineau) that it constitutes a way out for the defender of the Antipathetic Fallacy Hypothesis. Indeed, Sundström explicitly supposed that, when I thought the second half of the identity statement, I focused on my *current* experience of the whiteness of the background of my Word document. It couldn’t be the case then that my thinking is not accompanied by an experience of white.

However, one possibility is enough to protect the Antipathetic Fallacy Hypothesis against Sundström’s counterexample. Therefore, I think that it allows Papineau to block Sundström’s objection.

My opinion is that Sundström’s point is mostly right, and that the Antipathetic Fallacy Hypothesis does not constitute a satisfying explanation of the intuition of distinctness. My goal is to draw on Sundström’s proposal and to propose some refined counterexamples, which do not allow for the same kind of defense move as the one suggested by Papineau. I will devote the rest of this paper to the description of these refined counterexamples.

3. *A refined counterexample to the Antipathetic Fallacy Hypothesis*

Suppose that I am sitting on a couch with my sister Elise, facing a large TV screen. Both of us have our eyes open, and we are visually paying attention to the screen. A computer feeds the screen with images—let’s say, for reasons of simplicity, that they are only images of colored geometrical shapes: a red triangle, a blue square, a green rectangle, etc. In my hand, I hold a remote control. Every time I press a button on the remote control, the image on the screen changes. The software that runs in the computer makes it so that the succession of images is “random”, in the sense that I have absolutely no way to predict what the next image will be.

I will now expose a few thought experiments which all use this device. Let’s start with a rather innocuous one. Suppose that I am trying to consider a kind of naïve version of representationalist physicalism concerning consciousness, which states that when I have a conscious experience of an object, this conscious experience is identical with the physical state of my brain when it detects this object. Say that I am trying to decide if such a position is plausible while I am facing the screen—which, at the time, displays an image of a blue square. So: while visually focusing on the square, I consider the identity “My experience of this

blue square = The state of my brain when it detects this blue square". Let's stipulate that I think the first half of the identity statement by introspectively focusing on my experience of the blue square, and that, when I think the second half, I think the "blue square" component on the basis of my visual perception of it, by focusing on it. That means that the identity statement I am considering is *not* a phenomenologically contrasted identity statement. Indeed, my thinking of *both* halves of the identity statement crucially relies on me instantiating the phenomenal property associated with an experience of a blue square. But I take it that, in this case, I will still have a clear intuition of distinctness: I will be puzzled, as in any other case, by the fact that my *experience* could be identical with *a certain state of my brain*. I think that it shows that phenomenologically contrasted identity statements are not *necessary* for intuitions of distinctness to arise⁸, which in turn shows that the Antipathetic Fallacy, if it can be *a* cause that gives rise to such intuitions, clearly cannot be their *only* cause. This point, as noted previously, has been clearly recognized by Papineau himself in recent papers.

Let's turn to a second thought experiment. Let's say that, still facing the screen, I close my eyes, and then press the button of the remote control. I now know that a new image is being displayed on the screen, and I know that my sister is visually experiencing it. However, I have no idea what the image is. Now let's say that, with my eyes still closed, I start thinking about "My sister Elise's most salient current experience". I assume that this experience is the experience that she is currently having as she watches the screen. However, given that I have my eyes closed, I have no idea what this experience is. This guarantees me that, when I think about Elise's most salient current experience with my eyes closed, I am *not* (surreptitiously or not) activating an experience of the same kind.⁹ Let's say that I now prepare myself to entertain the hypothesis that Elise's most salient current experience, about which I am thinking with my eyes closed, is type-identical with the experience that I would myself get if I opened my eyes. I then consider that Elise's most salient current experience is identical with... (and here I open my eyes) *an experience of a purple hexagon* (where this part is thought with a phenomenal concept, and while focusing on the very experience I got as soon as I opened my eyes).

In this situation, do I have an intuition of distinctness? I take it to be obvious that I don't. Of course Elise's experience *can be* an experience of a purple hexagon. Of course she can be in a state that *feels like*

⁸ Sundström also put forth a thought experiment aiming at showing that, though it relied on the details of Papineau's account regarding *derived phenomenal concepts* (Sundström 2008: 141).

⁹ Except if (1) I am imagining what this experience *could be* and, by chance, I just *got it right*—but this would happen only very rarely; or (2) We suppose that I am able to activate together and at the same time *hundreds* of different visual experiences (of a blue square, of a red diamond, of a yellow star, all in different sizes and hues, etc.). I take this to be completely implausible.

this. But here is the point: the identity statement I was considering was a phenomenologically contrasted identity statement, and it seems very difficult to deny it. Indeed, I forced myself to think about Elise's current experience *without any possibility of knowing what this experience was*, so that I was obviously *not* activating a copy of this experience (even "surreptitiously"). That guarantees that the first half of the identity statement was thought about *without* activating an experience of a purple hexagon. And then I forced myself to think about the experience of a purple hexagon by introspectively focusing on the very experience I got when I opened my eyes—which guarantees that, this time, I instantiated the corresponding phenomenal property when I thought about the experience.

So, in the situation described in the thought experiment, I consider a statement, which is quite certainly a phenomenologically contrasted identity statement, and I nonetheless get no intuition of distinctness. This case therefore constitutes a counterexample to the Antipathetic Fallacy Hypothesis, which draws inspiration from Sundström's case but cannot be answered in the same way.

4. *Pulling apart the intuition of distinctness and the phenomenologically contrasted identity statements*

I have shown previously, in my first thought experiment, that phenomenologically contrasted identity statements are not *necessary* for an intuition of distinctness to arise. This is something Papineau himself recognized. I then presented a thought experiment that gives a reason to think that they are not *sufficient* for an intuition of distinctness. I tend to think that this shows that we should pull apart the issue of intuitions of distinctness, and the issue of phenomenologically contrasted statements. I now want to quickly present a few more thought experiments that could bring our intuition in the same direction.

Suppose that I am still facing the same screen, with my sister Elise still by my side. I then close my eyes and press the button of the remote control. At this point I know that my sister is looking at an image, but I don't know which image it is. Let's say that, my eyes closed, I start thinking about "the current state of Elise's visual cortex". I then decide to consider the fact that it is identical with... (and then I open my eyes) the state of Elise's visual cortex when she looks at *this*—where *this* is, say, a red oval, about which I think *on the basis of my visual perception of it*. Do I then have an intuition of distinctness? I take it to be obvious that I don't: the identity I am considering seems perfectly reasonable. *Of course* the current state of her cortex can be identical with a certain state of her cortex! But the identity statement I was trying to consider at the time was nonetheless a phenomenologically contrasted identity statement. This *also* shows that the phenomenological contrast between two conceptions is not sufficient to create an intuition of distinctness concerning the two objects referred to by the conceptions.

Now, let's consider one more thought experiments in which an intuition of distinctness *does* arise while I consider a phenomenologically contrasted identity statement, but in such a way that it does not fit well with the Antipathetic Fallacy Hypothesis. Let's say that, while I am in the same kind of situation as described previously, I think, with my eyes closed, about Elise's most salient current experience. Again, I have no idea what it is at the time. I then consider that this experience is identical with... (and then I open my eyes) the state of Elise's visual cortex when she looks at *this*—where *this* is, say, a blue spiral, about which I think on the basis of my visual perception of it.

I take it that, in that case, an intuition of distinctness would arise: how could Elise's experience be identical with *a state of her cortex*? This identity statement would seem as strange as any other physico-phenomenal identity statement. I may *believe* it, but I will find it puzzling nonetheless. In that case, the identity statement I consider happens to be a phenomenologically contrasted identity statement. However, we can see here that the phenomenologically loaded conception, which is in the *second half* of the identity statement, is not at all the conception that seems to refer to an irreducibly phenomenal entity. In fact, that is exactly the contrary. The intuition of distinctness arises, but what strikes me as being irreducibly phenomenal is the thing thought about in a non-phenomenologically loaded way: eyes closed, and while *not knowing* what kind of experience is thought about.

Our previous examples had shown that phenomenologically contrasted identity statements were not necessary, nor sufficient, for intuitions of distinctness to arise. The further examples I just put forth are cases in which the two relevant factors (the intuitions of distinctness in the one hand, the phenomenologically contrasted identity statements on the other hand) can vary quite independently from each other. I hope they will incite the reader to completely pull apart these two things. An intuition of distinctness can arise, whether or not we are considering a phenomenologically contrasted identity statement, and I don't think that we have solid reasons to believe that one causes the other.

I don't intend to assert here that the specific way in which we grasp our phenomenal experiences through introspection is *not* crucial when it comes to explaining the arising of the intuition of distinctness. I actually think that our introspective grasp of consciousness has special features, which explain this intuition. But the Antipathetic Fallacy hypothesis, understood as one particular way to interpret in what way our introspective grasp of experiences contributes to the presence of this intuition, is mistaken. It is not true that we are reluctant to equate a phenomenal experience (thought about introspectively) with a purely physical state (thought about with purely physical concepts) because the first thought activates the concerned experience (or a copy of this experience), while the other doesn't.

5. *Objection: can we really separate two steps in our thinking of identity statements?*

The counterexamples I just presented may be subject to objections. I would like to consider one of them, and then try to answer it.

In order for my counterexamples to be immune to the kind of answer that Papineau gave to Sundström's objection, I had to describe them in such a way that it is guaranteed that the identity statements considered are indeed phenomenologically contrasted identity statements. This crucial task was fulfilled thanks to some specific features of the situation described in the counterexamples.

It is especially crucial to the counterexamples that the first half of the identity statements considered is always thought *with the eyes closed* (while ignoring the image displayed on the screen), and that the other half is thought *with the eyes open*, and on the basis of the visual perception I then get. But one could object the following: when we consider an identity statement, our thought cannot be *temporally divided* in such a clear and cut way. Thought is not like speech, in this respect. In fact, whenever we consider an identity statement, we have to think the two conceptions of the two things that are being identified *at the same time*, and thus *bring together* these two conceptions in our mind, so to speak. For this reason, when we look at the counterexamples I just described, there is no sense in saying that we only think the first half of the identity "the eyes closed", and without activating the relevant experience, because we also have to think it with the eyes open. Indeed, it is only when we have the eyes open that we can properly think the second half of the identity statement, and then "assemble" in our minds the first half and the second half of the identity statement. Therefore, when we think the relevant identity statement, we *must* think *even* the first half of the identity statement with our eyes open, visually attending at the screen, which means that we must think it while having the relevant experience. This means that the identity statements we are considering in the counterexamples are not phenomenologically contrasted identity statements after all. Therefore, they are not counterexamples to the Antipathetic Fallacy Hypothesis.

This objection has a certain appeal. However, I don't think that a defender of the Antipathetic Fallacy Hypothesis could make use of it in order to defend her theory against the counterexamples I put forth. Indeed, this objection crucially relies on the thesis that, when we think an identity statement, we have to activate *at the same time* the two conceptions (of the two things we try to identify) in order to "bring them together" in our mind. But this would destroy the very possibility of phenomenologically contrasted identity statements. For this would mean that, anytime I think a given identity statement, I have to entertain the *two conceptions* thought about at the same time, which means that, at least at that moment, both conceptions would be accompanied

by whatever phenomenology accompanies the other. That means that I could *never* think an identity statement in such a way that my thinking of one part of the statement would be accompanied by a given phenomenology, while my thinking of the other part would *not*. But, if there are no such things as phenomenologically contrasted identity statements, then the *explanans* posited by the Antipathetic Fallacy Hypothesis does not exist. Therefore, this hypothesis is false.

For this reason, even if this objection can seem appealing, I don't think a defender of the Antipathetic Fallacy Hypothesis can make use of it in order to repeal my counterexamples.

6. *Concluding remarks*

In this paper, I have devised thought experiments in order to show (conclusively, I hope) that phenomenologically contrasted identity statements are neither necessary, nor sufficient, for intuitions of distinctness to arise. The counterexamples I designed to show that they are not sufficient were inspired by Pär Sundström's counterexample to Papineau's theory. I tried to construct them in such a way that it was not possible to answer them in the same way Papineau answered Sundström's. I also tried to put forth a variety of cases, in which these two features of the situations (whether or not the identity statements considered are phenomenologically contrasted, and whether or not they create an intuition of distinctness) vary independently. My goal was to incite the reader to pull apart these two features of the situation: the phenomenological contrast that can exist between two halves of an identity statement, and the birth of an intuition of distinctness vis-à-vis the two objects identified in the statement.

In this paper, I argued against the Antipathetic Fallacy Hypothesis, which is a hypothesis that aims at defending physicalism against the dualist intuition (the "intuition of distinctness"), by giving an explanation of this intuition within a physicalist framework. However, I did not plan to argue against physicalism. I did not even plan to argue against the Phenomenal Concept Strategy, if we understand it as the general attempt to defend physicalism against dualist intuitions (such as the intuition of distinctness) by appealing to some physically explainable features of our way of thinking about conscious experience in order to explain the birth of this intuition. I therefore think that numerous versions of the Phenomenal Concept Strategy (Aydede and Güzeldere 2005; Hill 1997; Levin 2007; Sturgeon 1994) are left untouched by my argument. My own point of view is that physicalism is true, and that we can account for the intuition of distinctness within a purely physicalist framework, by showing how this intuition arises as a consequence of some of the features of our introspective grasp of consciousness. However, as I tried to show, the Antipathetic Fallacy Hypothesis does not give us a satisfying explanation of this intuition. My own view,

for which I did not argue here, is that the only satisfying physicalist theory of our introspective grasp of consciousness is an *illusionist* one, according to which we introspectively represent ourselves as conscious even though consciousness does not really exist (Frankish 2016). Illusionist views of consciousness escape the objection made here, as well as objections usually made against the Phenomenal Concept Strategy. Of course, they encounter problems of their own, whose solution may not be trivial (Kammerer 2018, 2016).

References

- Aydede, M. and Güzeldere, G. 2005. "Cognitive Architecture, Concepts, and Introspection: An Information-Theoretic Solution to the Problem of Phenomenal Consciousness." *Noûs* 39 (2): 197–255.
- Balog, K. 2012. "Acquaintance and the Mind-Body problem." In C. Hill & S. Gozzano (ed.). *New Perspectives on Type Identity: The Mental and the Physical*. Cambridge: Cambridge University Press.
- Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
- Frankish, K. 2016. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies* 23 (11–12): 11–39.
- Hill, C. 1997. "Imaginability, Conceivability, Possibility and the Mind-Body Problem." *Philosophical Studies* 87: 61–85.
- Jackson, F. 1982. "Epiphenomenal qualia." *Philosophical Quarterly* 32: 127–136.
- Kammerer, F. 2016. "The hardest aspect of the illusion problem—and how to solve it." *Journal of Consciousness Studies* 23 (11–12): 123–139.
- Kammerer, F. 2018. "Can you believe it? Illusionism and the illusion meta-problem." *Philosophical Psychology* 31 (1): 44–67.
- Kripke, S. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.
- Levine, J. 1983. "Materialism and qualia: the explanatory gap." *Pacific Philosophical Quarterly* 64: 354–61.
- Levine, J. 2001. *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press.
- Levin, J. 2007. "What is a Phenomenal Concept?" In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: Oxford University Press.
- Levine, J. 2007. "Phenomenal Concepts and the Materialist Constraint." In T. Alter and S. Walter (ed.). *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: Oxford University Press.
- Loar, B. 1997. "Phenomenal States (Revised Version)." In N. Block, O. Flanagan and G. Güzeldere (ed.). *The Nature of Consciousness*. Cambridge: MIT Press: 597–616.
- Papineau, D. 1993. "Physicalism, Consciousness, and the Antipathetic Fallacy." *Australasian Journal of Philosophy* 71: 169–183.
- Papineau, D. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.

- Papineau, D. 2007. "Phenomenal and Perceptual Concepts." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: Oxford University Press.
- Papineau, D. 2011. "What Exactly is the Explanatory Gap?" *Philosophia* 39 (1): 5–19.
- Stoljar, D. 2005. "Physicalism and phenomenal concepts." *Mind and Language* 20 (2): 296–302.
- Sturgeon, S. 1994. "The Epistemic View of Subjectivity." *The Journal of Philosophy* 91 (5): 221–235.
- Sundström, P. 2008. "Is the mystery an illusion? Papineau on the problem of consciousness." *Synthese* 163 (2): 133–143.
- Tye, M. 1999. "Phenomenal consciousness: the explanatory gap as a cognitive illusion." *Mind* 108 (432): 705–725.

On Understanding a Theory on Conscious Experiences

ERHAN DEMIRCIOĞLU
Koç University, Istanbul, Turkey

McGinn claims, among other things, that we cannot understand the theory that explains how echolocationary experiences arise from the bat's brain. One of McGinn's arguments for this claim appeals to the fact that we cannot know in principle what it is like to have echolocationary experiences. According to Kirk, McGinn's argument fails because it rests on an illegitimate assumption concerning what explanatory theories are supposed to accomplish. However, I will argue that Kirk's objection misfires because he misapprehends McGinn's argument. Further, I will articulate and briefly assess some ways in which McGinn's argument can be blocked.

Keywords: The mind-body problem, Concepts of consciousness, What it is like to be a bat, Colin McGinn, Robert Kirk.

McGinn (1989) claims, among other things, that we cannot understand the theory that explains how echolocationary (or batty) experiences arise from the bat's brain. One of the influential arguments McGinn develops for this claim appeals to the fact that we cannot know in principle what it is like to have batty experiences. According to Kirk (1991), McGinn's argument fails because it rests on an illegitimate assumption concerning what explanatory theories are supposed to accomplish. However, my main aim in this note is to show that Kirk's objection misfires because he misapprehends McGinn's argument. The objection that I will consider is somewhat old but I hope this does not by itself detract from its significance. The sort of misunderstanding of McGinn's argument that is encapsulated in Kirk's objection has not been sufficiently recognized in the literature, which might explain at least in part the general tendency many philosophers seem to have of rejecting McGinn's overall account out of hand. After answering Kirk's objection, I will articulate and briefly assess some ways in which McGinn's argument can be blocked.

1. It is widely assumed that we human beings cannot know in principle what it is like to have batty experiences. There is a clear sense in which the characteristic qualitative aspect, or the phenomenal character, of the experiences bats have when they navigate the environment by using their echolocationary techniques appears to be irredeemably beyond our cognitive grasp. Batty experiences will never be intelligible to us, it seems, in the way our experiences like smelling a skunk or tasting coffee are.¹

What follows from our lack of access to batty experiences? In particular, does our failure to access batty experiences provide any support for the following thesis?

(T) We cannot understand the theory that explains how batty experiences arise from the bat's brain.

McGinn argues that (T) is supported by the fact that we cannot access the phenomenal character of batty experiences:

Call this type of experience [batty experience] *B*, and call the explanatory property that links *B* to the bat's brain *P_r*. By grasping *P_r* it would be perfectly intelligible to us how the bat's brain generates *B*-experiences, we would have an explanatory theory of the causal nexus in question...But then it seems to follow that grasp of the theory that explains *B*-experiences would confer a grasp of the nature of those experiences: for how could we understand that theory without understanding the concept *B* that occurs in it? How could we grasp the nature of *B*-experiences without grasping the character of those experiences?...Our concepts of consciousness just are inherently constrained by our own form of consciousness, so that any theory the understanding of which requires us to transcend these constraints would ipso facto be inaccessible to us. (McGinn 1989: 355–6)²

McGinn's argument here is, roughly, this: (fully) understanding a theory that explains how *B*-experiences arise from the bat's brain requires us to grasp the concept *B* that (ineliminably) occurs in that theory,³ which in turn requires us to grasp the character of those experiences; however, we just cannot grasp the character of those experiences; therefore, we cannot understand the theory that explains how *B*-experiences arise from the bat's brain.

In a more explicit form, the argument runs as follows:

(1) Understanding a particular theory requires grasping all the concepts that occur in that theory.

¹ Nagel's seminal paper (1974) played the main role in forcefully bringing to the attention of philosophers the significance of our epistemic position with respect to batty experiences for the mind-body problem.

² All McGinn references are to this work.

³ McGinn does not explicitly state but presumably takes for granted that the qualifications in the parentheses ('fully' and 'ineliminably') are necessary for the argument to get off the ground. If understanding were taken as *partial* understanding, or if *B* were a concept that *eliminably* occurs in the theory, the first premise of McGinn's argument would be obviously false. Having noted that, I will suppress these qualifications in the remainder of the paper.

- (2) The concept *B* occurs in the theory that explains how *B*-experiences arise from the bat's brain.

From (1) and (2), it follows that

- (3) Understanding the theory that explains how *B*-experiences arise from the bat's brain requires grasping the concept *B*.

We also independently have the following:

- (4) We can grasp the concept *B* only if we can grasp the character of *B*-experiences.
(5) We can grasp the character of *B*-experiences only if we can have *B*-experiences.⁴
(6) We cannot have *B*-experiences.⁵

From (3)–(6), it follows that

- (T) We cannot understand the theory that explains how *B*-experiences arise from the bat's brain.

Let's call (1) *the grasping requirement* (GR); and, by taking special note of (1), let's call this argument *the argument from grasping*.⁶

How is the argument from grasping related to the argument McGinn develops for the sort of "mysterianism" he is best known for? According to McGinn's mysterianism, we cannot understand (or are "cognitively closed" with respect to) the theory that explains how *our* experiences arise from our brains. Clearly, a straightforward adjustment of the argument from grasping cannot support McGinn's mysterianism simply because we possess concepts of our experiences and thereby know what our experiences are like. McGinn's argument for his mysterianism, which I call *the closure argument by elimination*, runs roughly as follows: introspection and perception are the "two possible avenues open to us in our aspiration to identify *P* [the brain property that is responsible for our consciousness]" (397), and neither can help us identify *P*—therefore, we cannot identify *P*, in which case we cannot solve the mind-body problem in the case of humans. The argument from grasping is presented by McGinn as "a further point" (355) in (and hence it is a digression from) his main discussion of whether introspection can enable us to get to *P*; and as such it stands on its own and is independent of the closure argument by elimination. Given the implications of its conclusion, the closure argument by elimination

⁴ It is clear that the argument could have been stated by simply having 'We can grasp the concept *B* only if we can have *B*-experiences' as a premise instead of having both (4) and (5). However, I here stick with the way McGinn seems to prefer to state it.

⁵ Premises (5) and (6) are not explicitly stated in the passage quoted from McGinn, but the context surrounding the passage leaves no doubt that McGinn holds them.

⁶ I claim that the argument from grasping is one plausible and textually supported interpretation of McGinn's argument in the relevant passage, while I do not wish to claim that it is the correct one. Thanks to an anonymous reviewer for pressing on this issue.

is bound to be more controversial than the argument from grasping⁷; however, it seems to me that the latter also brings out some interesting issues and deserves a separate and focused investigation.

What, one might reasonably wonder, if the argument from grasping succeeds in establishing (T)? What is the significance of the purported truth that we cannot understand the theory that explains how *B*-experiences arise from the bat's brain? As McGinn sees it, nothing less than the possibility of *our* achieving "a general solution to the mind-body problem" (356) is at stake. If (T) is true, then, according to McGinn, "even if we could solve [the mind-body problem] for our own case, we could not solve it for bats or Martians" (356). Of course, such a result would be especially worrisome for a variety of reductionist views on consciousness: if *B*-experiences are nothing but some physical features of the bat's brain, as physicalism claims them to be, or if they are nothing but some causal-role properties of the bat's brain, as functionalism claims them to be, then what can possibly obstruct our path to solving the mind-body problem for bats? The question, I believe, is forceful: it certainly seems that the truth of (T) would be a mystery if some form of reductionism were true.

I believe McGinn's account has not received the due attention it deserves, and I am largely in agreement with Kriegel's following observation: "The literature on mysterianism has so far been somewhat dogmatically dismissive. Critical discussions of the merits and demerits of the view are few and far between. In particular, McGinn's argument is rarely if ever engaged" (2009: 455). The current paper may be read as a modest attempt to remedy this unfortunate situation by focusing on a particular but significant strand in McGinn's position. In what follows, I will argue against the objection Kirk develops against McGinn's argument. Once a misunderstanding like Kirk's is eliminated, the ground is cleared for drawing out the full implications of McGinn's argument. I will conclude by suggesting a trilemma, one that captures the options available for resisting the argument and thereby functions as an invitation for the potential dissidents to clarify their stand.

2. Kirk argues that McGinn's argument in the passage quoted above assumes that "a satisfactory theory of explaining the nature of subjective experience must be capable of actually conferring concepts of experience on those who start off without them" (Kirk 1991: 20).⁸ Let us call this assumption *the conferring requirement* (CR). So, we have the following:

(CR) A theory of conscious experience is explanatorily satisfactory for us (or for cognitive beings in general) only if it is capable of conferring a grasp of the concepts of conscious experience involved in that theory to (those of) us (or cognitive beings in general) who start off without (or who do not have an independent grasp of) them.

⁷ For a critical discussion of McGinn's mysterianism, see Sacks (1994).

⁸ All Kirk citations are to this work.

By Kirk's lights, CR "is illegitimate" (19) because it requires the explanatory theory to achieve something no theory can possibly achieve and can therefore be reasonably expected to achieve. Grasping concepts of experience requires "an actual sort of experience of the right sort," which is "something no theory could supply" (19). Holding that the explanatory theory should confer a grasp of such concepts as *B* is, Kirk argues, setting up "an insuperable hurdle" (21) for that theory; and, if we fail to get over it, then it is not our cognitive powers but that very hurdle itself that should deserve the blame.⁹

According to Kirk, if CR were legitimate, i.e. if it were required for a satisfactory theory of conscious experience that it confer concepts of experience on those who start off without them, then the proper conclusion to be drawn would be that no theory can meet that requirement and hence there cannot *be* such a satisfactory theory, one that is a possible object of our understanding. Therefore, Kirk finds it "puzzling" (19) that McGinn assumes the legitimacy of that requirement while holding that there *is* a satisfactory theory of conscious experience.

Let us assume, for the moment, that McGinn endorses CR. How, then, would the argument for (T) proceed? It is evident that in the passage above, McGinn's main intention is to develop an argument for (T), but it is not at all clear how the argument can possibly be intended to move from CR. CR states that the explanatory theory in question should confer a grasp of the concept *B*, and (T) states that we cannot understand that theory. However, if the explanatory theory *confers* a grasp of *B*, as CR says it must, then what reason can we possibly have to think that (T) is true? If anything, just the opposite appears to be correct: holding CR provides a good reason for thinking that (T) is *not* true. That the explanatory theory confers a grasp of *B* supports the conclusion that we *can* understand that theory.

The moral is that if McGinn were really to endorse CR, then what is, from his point of view, a formidable obstacle to our understanding the theory explaining batty experiences (namely, our apparent failure to grasp the concept *B* that occurs in that theory) would be overcome by what that theory confers. So, assuming that McGinn endorses CR

⁹ I would like to note in passing that Patricia Churchland attributes to McGinn some other requirement that is relevantly similar to CR. In a rather belligerent response to McGinn's (2014) review of her (2013), Churchland writes that McGinn's account suffers from "a whopping flaw" and that "no causal explanation for a phenomenon...should be *expected* to actually produce that phenomenon" (2014, emphasis original). On a natural interpretation, Churchland attributes to McGinn the requirement that a satisfactory explanation of a particular subjective experience must actually produce that experience in those who understand that explanation (*the producing requirement*, PR). PR follows from CR on the reasonable assumption that an explanation of a particular subjective experience can confer concepts of experience on those who start off without them only if that explanation can produce that subjective experience in them. I will not discuss Churchland's attack on McGinn's position separately but I am confident that what I have to say below about Kirk's critique applies *mutatis mutandis* to it.

faces the difficulty of giving a reasonable (and charitable) account of how McGinn's argument for (T) can possibly proceed.

The argument intended by McGinn for (T), I claim, is the argument from grasping, and as such, it has nothing much to do with (more specifically, neither assumes nor needs to assume) CR. So, regardless of whether Kirk shows the illegitimacy of CR, he fails to properly address McGinn's argument from grasping. Furthermore, if McGinn neither assumes nor needs to assume CR in his argument for (T), then there need not be anything "puzzling" in McGinn's commitment to the thesis that there *is* a satisfactory theory of conscious experience.¹⁰

3. Despite what I have argued for above, there is a particular statement in the passage above that might well give the impression that McGinn endorses CR. To quote again, McGinn writes: "it seems to follow that grasp of the theory that explains *B*-experiences would *confer* a grasp of the nature of those experiences." Kirk (19) places a special emphasis on this statement and takes it as evidence for the claim that McGinn endorses CR.

However, McGinn's statement at hand can be interpreted in a different way, and in a way that gives further support to interpreting McGinn's argument along the lines of the argument from grasping. In the relevant passage, McGinn can be plausibly taken as raising the following question: "Assuming that we cannot have *B*-experiences, how can we possibly understand the theory that involves the concept *B*?" Having *B*-experiences would enable us to form the concept *B* that occurs in that theory. Barring that, McGinn argues, the relevant explanatory theory can be understandable *by us* only if it confers the concept *B on us*. According to McGinn, we can understand a theory only if we can grasp the concepts it involves (GR); and, if we do not have an independent grasp of some of those concepts (i.e. if we do not have a grasp of those concepts prior to our exposure to the theory), then we can understand the theory only if it confers on us such a grasp. That theories must confer on us a grasp of some concepts that we do not have an independent grasp of if

¹⁰ It might be objected that McGinn is committed to the thesis that CR holds (if not for us) at least for *some* beings (e.g., those beings that are "cognitively open" to the relevant theory) and therefore that Kirk's objection to McGinn's account stands untouched by my point above. However, there is at least one good reason to think that McGinn is not committed to the idea that CR holds for some beings. It is important here to note the distinction McGinn makes between *absolute* and *relative* cognitive closure: "A problem is absolutely cognitively closed if no possible mind could resolve it; a problem is relatively closed if minds of some sorts can in principle solve it while minds of other sorts cannot" (360). McGinn also writes: "It certain seems to be at least an open question whether the problem is absolutely insoluble; I would not be surprised if it were" (361). Now, if McGinn grants the possibility of *absolute* cognitive closure, as he clearly does, then there might well be no minds that are cognitively open to the theory that explains *B*-experiences. However, given that McGinn holds that there is such a theory, then from his point of view, there being such a theory has nothing to do with its potential to endow *some* beings with a grasp of *B* because there might well be no such beings.

they are to be understandable by us follows from GR (more on this below). Therefore, GR is the claim that Kirk needs to attack if he wishes to undermine McGinn's argument for (T).¹¹ What McGinn has to say about conferring ultimately depends upon the GR he takes for granted.

Under this interpretation, the statement that Kirk focuses on does not commit McGinn to CR. CR is a requirement that theories that include some concepts that we don't have an independent grasp of must satisfy in order for them to be *explanatorily satisfactory*. According to Kirk, McGinn commits "the error of assuming that a *satisfactory* theory must actually endow us with concepts for characterizing experience" (22, emphasis mine). However, McGinn's statement at hand is about a requirement that those theories must satisfy in order for them to be *understandable by us* (*the understandability requirement*, UR). So, we have the following:

(UR) A given theory is understandable by us (or cognitive beings in general) only if it is capable of conferring a grasp of those concepts involved in that theory to (those of) us (or cognitive beings in general) who start off without (or who do not have an independent grasp of) them.

From UR, we can derive the following as one of its specific instances:

(URE) A theory of conscious experience is understandable by us (or cognitive beings in general) only if it is capable of conferring a grasp of those concepts of conscious experience involved in that theory to (those of) us (or cognitive beings in general) who start off without (or who do not have an independent grasp of) them.

For McGinn, if a theory that includes some concepts that we don't have an independent grasp of does not confer a grasp of those concepts on us, then that theory is not understandable by us. However, the fact that a theory is not understandable by us does not mean that it does not satisfactorily explain what it is intended to explain. So, a theory that does not confer on us a grasp of the concepts that we do not have any independent grasp of, according to McGinn, is *not* understandable by us but might still *be* explanatorily satisfactory.¹²

¹¹ Interestingly, Kirk explicitly accepts at one point GR. He writes: "Having concepts such as *B* is *necessary* in order to understand the theory that explains the character of *B*-experiences" (21). Given this, one might find it "puzzling" that Kirk accepts GR but still attacks what McGinn has to say about conferring. Of course, the puzzle dissolves once we realize that what McGinn has to say about conferring has nothing much to do with the CR Kirk attributes to him. This provides further support to my claim above that Kirk has misapprehended McGinn's statement on conferring.

¹² One might wonder what the argument from grasping would look like once McGinn's UR (or URE) is taken on board. The revision required is minimal: replace (5) above by (5') We can grasp the character of *B*-experiences only if either we can have *B*-experiences or the theory can confer on us a grasp of the character of *B*-experiences, and add as a new premise (7) The theory *cannot* confer on us a grasp of the character of *B*-experiences.

4. As stated above, Kirk's principal reason against CR is that there are no theories on conscious experiences that can possibly fulfill and can therefore be reasonably expected to fulfill it. Since McGinn does not embrace CR, I will not spend time on Kirk's argument against it. However, one might develop an analogous argument against URE and argue that because there are no theories on conscious experiences that can possibly fulfill and can therefore be reasonably expected to fulfill URE, it is *illegitimate*. If Kirk's reason against CR is forceful, and if it works equally effectively against URE, as one might reasonably argue, then pointing at the fact that Kirk misidentifies the requirement McGinn embraces does not fully circumvent the problem that Kirk thinks afflicts McGinn's account. If so, it might be maintained that the real issue with McGinn's position has not yet been brought into relief.

By Kirk's lights, CR is "illegitimate" because it requires an explanatory theory of conscious experience to achieve something no theory of conscious experience can possibly achieve and can therefore reasonably expected to achieve: given that *no* explanatory theory can confer on us concepts of conscious experience we do not have an independent grasp of, then CR sets an "insuperable hurdle" for such an explanatory theory to be satisfactory, which renders it illegitimate. The question I am concerned with now is how compelling an objection to URE, which is similar in spirit and form to Kirk's objection to CR, would be. The objection is this: given that *no* theory of conscious experience that involves some concepts of conscious experience that we do not have an independent grasp of can confer a grasp of those concepts on us, URE sets up an "insuperable hurdle" for such a theory to be understandable by us. Therefore, the objection goes, URE cannot be a requirement that we can reasonably expect theories of conscious experience should meet in order for them to be understandable by us.¹³

In response to this, I would like to concede two points. First, it is reasonable to derive the illegitimacy of a particular requirement from the fact that it cannot possibly be met by those on which it is placed as a requirement. This is, roughly, for the familiar reasons why many philosophers hold that *ought* entails *can*. Second, there are no theories on conscious experiences that can possibly satisfy URE. It is also worth noting that McGinn does not make any claim contradicting the former point and, more importantly, explicitly grants the latter point (p. 356) (just as would be expected given his intention to argue for (T): if there were theories that confer on us such a grasp, then that would be a reason for thinking that (T) is *not* true).

However, I would like to argue that it does not follow from these two concessions that URE is illegitimate. The crucial point here is that URE is derived from UR and UR is intended to be a *general* requirement for all theories but not a local requirement solely for theories on conscious

¹³ I would like to thank an anonymous reviewer for the comment that led to a clearer statement of this objection.

experiences. And, as such, the central rationale for UR derives from an eminently plausible general requirement about the conditions under which a theory is understandable by us, viz. GR: if we can understand a theory only if we can grasp all the concepts that occur in it (GR), then we can understand a theory that contains some concepts we do not have an independent grasp of only if it confers such a grasp on us (UR). Hence, pointing at the putative fact that there are no theories *on conscious experiences* that can possibly satisfy a particular instance of UR (viz. URE) cannot by itself show that UR as a general principle or URE as a specific instance of UR is illegitimate. The argument for the illegitimacy of URE needs to target the general principle from which it is derived (viz. UR) and needs to demonstrate that there are no theories on *any areas of inquiry* that can possibly satisfy it. And, the putative fact that UR cannot be satisfied by theories on *a particular area* cannot show that UR is illegitimate just as the fact that a particular person is incapable of fulfilling a particular moral requirement cannot show that the requirement is illegitimate. If Dexter has an irresistible urge to commit murder and is incapable of acting in accordance with a commandment like 'Thou shall not kill!', then the proper conclusion to be derived is not that the commandment is illegitimate. Similarly, if theories on conscious experiences cannot satisfy the relevant specific version of UR (viz. URE), then the proper conclusion to be derived is not that that specific version is illegitimate.

Further, there *are* certainly theories that satisfy UR. Take for instance Galileo's theory of motion, where, on a standard history of science, the concept *acceleration* is clearly introduced for the first time in physics by distinguishing it from *velocity*. Assuming that we do not have an independent grasp of the concept *acceleration*, we can understand Galileo's theory of motion only if it confers on us a grasp of that concept that occurs in it. And the theory actually confers on us such a grasp by defining it in terms of some concepts that we already have an independent grasp of such as *velocity* and *time*.¹⁴ Plainly, examples can easily be multiplied indefinitely.

The moral I draw is that UR is a general requirement that is supported by GR, which is itself another general requirement, and therefore, from the fact that there are no theories *on conscious experiences* that can satisfy one of its specific instances, it does not follow that that specific instance is illegitimate.

¹⁴ In this particular case, conferring takes the form of defining: the theory confers a grasp of a particular concept on the subject through defining it in terms of some of its other concepts. An interesting question is whether there are other forms the conferring relation in question might take, i.e., whether a theory can fulfill UR without defining the problematic concepts in terms of some other concepts. I myself cannot conceive any other way. That being said, however, I can fortunately sidestep the question for the purposes of this paper.

5. Let me highlight the central points that have emerged in our discussion. Firstly, the CR that Kirk thinks McGinn endorses is not an assumption of the argument intended by McGinn, namely the argument from grasping. The argument from grasping stands untouched even if Kirk is right that CR is illegitimate. Secondly, there is good reason to think that McGinn does not endorse CR. The truth of CR would count against the truth of (T). Thirdly, McGinn holds UR, i.e., that a theory that involves some concepts we do not have an independent grasp of is understandable by us only if understanding it confers on us a grasp of those concepts. The constraint placed here by McGinn on theories concerns their understandability by us and can be plausibly taken as following from GR; and as such, it does not entail CR. Fourthly, McGinn holds that there are not any theories of conscious experience that can satisfy UR.

The upshot is that McGinn does not endorse the CR Kirk attributes to him but endorses UR. Further, McGinn does not think the latter requirement is satisfied by theories on conscious experiences. All in all, Kirk's attack leaves McGinn's argument for (T) unscratched.

6. I would now like to close the paper with articulating and assessing the ways I find most plausible to block McGinn's argument from grasping. Despite its sketchiness, I believe that this will be of value for at least one good reason. The discussion below will further attest to the force of the argument from grasping and invite the potential dissidents to clarify their stand.

It is clear that the argument is valid. Further, I hold that premises (1) and (6) are virtually unassailable—any attempt to block the argument by denying one of these premises is *too* desperate:

- (1) Understanding a particular theory requires grasping all the concepts that occur in that theory.
- (2) We cannot have *B*-experiences.

This leaves us with premises (2), (4) and (5) as possible targets:

- (3) The concept *B* occurs in the theory that explains how *B*-experiences arise from the bat's brain.
- (4) We can grasp the concept *B* only if we can grasp the character of *B*-experiences.
- (5) We can grasp the character of *B*-experiences only if we can have *B*-experiences.

There are basically two different ways to attack these premises. First, one might deny (2). The most plausible way of attacking (2) is, it seems to me, to argue that it is based on the mistaken assumption that there *are* "concepts of consciousness" (McGinn) such as *B*. Concepts of consciousness, as McGinn conceives them, are formed through the introspective attention of the experiencing subject to the qualities of her experiences. One might reasonably endorse eliminativism about concepts of consciousness and argue that there are no concepts satisfying the

conditions McGinn articulates¹⁵ (perhaps because, one might say, the whole idea of concepts of consciousness stems either from a mistaken picture of introspection as “turning one’s gaze inward” or from a faulty view about the qualities of experiences).¹⁶

Another way to attack the argument from grasping is to deny what is entailed by (4) and (5), *viz.* that we can grasp the concept *B* only if we can have *B*-experiences. According to this objection, there *are* concepts of consciousness, as McGinn takes them to be, but grasping them does *not* require having the experiences that they refer to. The most viable version this objection might take, I believe, endorses reductionism about concepts of consciousness and maintain that those concepts can be analyzed in terms of some other concepts the grasping of which does not require one to have any specific experiences. Take, for instance, those varieties of functionalism according to which there are causally-based synonyms of concepts of consciousness. On such views, grasping a concept of consciousness is nothing more mysterious or demanding than grasping a concept describing a causal role. And, since grasping causal-role concepts does not require one to have any specific experiences, as one might reasonably argue, phenomenal concepts *qua* causal-role concepts do not also require one to have any specific experiences (including those experiences that they refer to). That is, if functionalism (or some form of reductionism about concepts of consciousness in general) is true, then either (4) or (5) is false.

So far as I can see, there are no other objections to the argument from grasping that are even remotely plausible. What I suggest, then, is a trilemma: *either* eliminativism *or* reductionism about concepts of consciousness, *or* (T). In slightly different words, if there are concepts of consciousness such as *B* that cannot be analyzed in terms of some other concepts that we possess, then (T) is inevitable—or so I have argued.¹⁷

¹⁵ An alternative way of denying premise (2) is to maintain that there are concepts of consciousness but they are not part of the relevant theory. As I see it, the idea here is not substantially different from eliminativism that I mention above. More specifically, this attack on (2) is eliminativism *so far as the theory that explains how B-experiences arise from the bat’s brain is concerned*. Therefore, at least for the purposes of this paper, the difference between the claim that there are phenomenal concepts but they are no part of the relevant theory and the claim that there are no concepts of consciousness *tout court* is not big enough to justify a separate treatment of the former.

¹⁶ Note what McGinn says about how concepts of consciousness are related to introspection: “Our acquaintance with consciousness could hardly be more direct; phenomenological description thus comes (relatively) easily. ‘Introspection’ is the name of the faculty through which we catch consciousness in all its vivid nakedness. By virtue of possessing this cognitive faculty we ascribe concepts of consciousness to ourselves; we thus have ‘immediate access’ to the properties of consciousness” (354). McGinn’s description of “introspectively ascribed concepts” (354) is anything except uncontroversial. See, for instance, Dennett (1988).

¹⁷ It is worth noting that the trilemma in question raises a serious challenge for a popular physicalist strategy to block various arguments for dualism, often called

References

- Churchland, P. 2013. *Touching a nerve*. New York: Norton.
- Churchland, P. 2014. "In response to: storm over the brain." *The New York Review of Books*.
- Demircioğlu, E. 2013. "Physicalism and phenomenal concepts." *Philosophical Studies* 165 (1): 257–277.
- Dennett, D. 1988. "Quining Qualia." In A. Marcel and E. Bisiach (eds.). *Consciousness in contemporary science*. Oxford: Oxford University Press.
- Kirk, R. 1991. Why shouldn't we be able to solve the mind-body problem? *Analysis* 51 (1): 17–23.
- Kriegel, U. 2009. "Mysterianism." In T. Bayne, A. Cleermans, and P. Wilken (eds.). *The Oxford companion to consciousness*. Oxford: Oxford University Press.
- Loar, B. 2004. "Phenomenal states." Originally published 1990/7. In P. Ludlow, Y. Nagasawa and D. Stoljar (eds.). *There is something about Mary*. Cambridge: MIT Press.
- McGinn, C. 1989. "Can we solve the mind-body problem?" *Mind* 98 (391): 349–366.
- McGinn, C. 2014. "Storm over the brain." *The New York Review of Books*.
- Nagel, T. 1974. "What is it like to be a bat?" *The Philosophical Review* 83 (4): 435–450.
- Papineau, D. 2004. *Thinking about consciousness*. Oxford: Clarendon Press.
- Sacks, M. 1994. "Cognitive closure and the limits of understanding." *Ratio* 7 (1): 26–42.
- Stoljar, D. 2005. "Physicalism and phenomenal concepts." *Mind and Language* 20: 469–494.

"the Phenomenal Concept Strategy" (PCS), a definitive thrust of which is a rejection of both conceptual eliminativism and conceptual reductionism (see for instance Loar (2004), Papineau (2004), Stoljar (2005), and Demircioğlu (2013)). The challenge for PCS is to show either that the trilemma I have just stated is a false one (i.e., it does not exhaust all the (plausible) options that one might have concerning the argument from grasping) or that, despite appearances, (T) can be adequately accommodated by a sort of physicalism, a doctrine that has enough bite to deserve that title. It appears that in either way, PCS faces a tall order.

‘Mais la fantaisie est-elle un privilège des seuls poètes?’ Schlick on a ‘Sinnkriterium’ for Thought Experiments

DANIEL DOHRN

Humboldt-Universität zu Berlin, Germany

Ever since the term ‘thought experiment’ was coined by Ørsted, philosophers have struggled with the question of how thought experiments manage to provide knowledge. Ernst Mach’s seminal contribution has eclipsed other approaches in the Austrian tradition. I discuss one of these neglected approaches. Faced with the challenge of how to reconcile his empiricist position with his use of thought experiments, Moritz Schlick proposed the following ‘Sinnkriterium’: a thought experiment is meaningful if it allows to answer a question under discussion by imagining the experiences that would confirm that the thought experimental scenario is actual. I trace this view throughout three exemplary thought experiments of Schlick’s.

Keywords: Thought experiment, Schlick, imagination, counterfactual, empiricism, verificationism.

Thought experiments are many and varied in science and philosophy. However, it is not so well understood how they contribute to our knowledge of the world. Unlike real experiments, they do not seem to interact with the world in a way that would allow us to gather new information. This concern is aggravated if one subscribes to a broadly empiricist view. According to such a view, any knowledge is eventually due to our more or less direct experiential contact with reality. Hence it is especially interesting to see how philosophers with a strong empiricist creed react to the practice of thought experimenting. In this article, I shall discuss one classical position within the movement of logical positivism which has hitherto been eclipsed by the contributions of contemporaries like Ernst Mach: Moritz Schlick’s proposal of how to make sense of thought experiments.

1. *What is a Thought Experiment?*

I shall start with a first take on thought experiments. The notion was coined in German by the physicist Hans-Christian Ørsted (1822), who used the word 'Gedankenexperiment' to refer to Kant's account of geometry in terms of the a priori use of imagination. Philosophers like Ernst Mach (1897) greatly expanded the scope of the term, including the scientific standard cases which also stand out in current debate, Galilei's falling bodies experiment, Newton's bucket experiment, and so on. Mach also gave a first empiricist account of how thought experiments can provide new knowledge. It makes tacit constraints imposed on imagination in the course of human evolution explicit. The role of retrieving and rearranging tacit knowledge plays a role in many recent accounts of thought experiments (e.g. Mišćević 1992).

Thought experiments do not only pervade science, they also abound in philosophy. I shall attempt at outlining some structural features of a typical philosophical thought experiment.

- 1) There is a question under discussion QUD.
- 2) An (as if) individual situation is described.
- 3) The situation is invented: we do not care whether it is actual.
- 4) Intuition: what would be the case in the situation?
- 5) The intuition is instrumental in answering the QUD.

Of course, this structure is only minimal. The aim is not to give necessary and sufficient conditions of thought experiments, but only to provide a first idea. I shall illustrate the structure by an example from the philosophical debate, so-called Gettier cases. I choose this example because it is one of the few successful thought experiments in philosophy, and it is used in many metaphilosophical debates (e.g. Williamson 2007).

1. QUD: Is knowledge justified true belief (JTB)?
- 2.–3. Invented scenario:

GC: At 8:28, Smith looks at a clock to see what time it is. The clock is broken; it stopped exactly twenty-four hours previously. Smith believes, on the basis of the clock's reading, that it is 8:28.(cf. Williamson 2009)

4. Intuition:
GC is possible.

If GC were actual, would Smith have JTB?—Yes!

If GC were actual, would Smith have knowledge?—No!

5. Hence knowledge is not (just) JTB.

This is only a schematic presentation of main structural features. If we look for a sound logical argument, the following formalization is plausible (Williamson 2007: 195):

- (i) Necessarily, for any subject S and proposition p, S knows p if and only if S has justified true belief in p.

- (ii) Possibly, some S is in GC.
- (iii) If some S were in GC, some S would have justified true belief in some p without knowing p.
- (iv) It is possible that some S has justified true belief in some p without knowing p.

Thus: not (i)

Having outlined a preliminary idea of thought experiments, I shall address what might be their most puzzling feature. In 'normal' experiments, we settle a question about independent reality by observation. Experiments are more sophisticated versions of observational practice. One arranges for standard observational conditions which are ideal for answering a question, and then one observes what happens under these conditions. For instance, in order to test whether there are Higgs bosons, physicists built a large hadron collider in which particles were accelerated until they had almost light speed. Under these conditions, Higgs bosons could be observed.

In thought experiments, one main ingredient of normal experiments is lacking. We do not observe an independent reality. One may try to frame thought experiments as observations of one's own reactions to certain considerations, but in my view this would be misleading. Thought experiments simply do not aim at observation in the way normal experiments do. The question becomes how merely imagined scenarios can be informative. Relatedly, when does it make sense to answer a question by using such a scenario?

Although there is a huge literature on thought experiments, these questions have not yet found a wholly satisfactory answer. Instead of trying to present one of my own, I shall consider an answer dating back to the first half of the 20th century. I find this answer interesting not only because it has been somewhat neglected in the literature but also because of the peculiar dialectical situation. While the point of thought experiments is difficult to appreciate in principle, the difficulty is much aggravated in a strongly empiricist framework. The answer I shall consider is bound to such a framework.

2. *Schlick's 'Sinnkriterium' for Thought Experiments*

Moritz Schlick is famous for being the founder of one of the most influential groups of philosophers to have flourished in the 20th century, the Vienna Circle. Schlick was also one of the main authors to set the circle's agenda. Among his tenets ranks the famous 'Sinnprinzip': the meaning of a statement consists in the conditions of its empirical verification. He also endorsed his own version of logical positivism, the view that any meaningful question has to be settled either by analysing one's use of language or by empirical means. Given these key convictions, it comes barely as a surprise that Schlick seems highly critical of thought experiments as far as they draw on merely invented scenarios:

...the philosopher is barely interested in merely fancied, invented objects; it is the real world that poses the big problems for him. ([1929] GAI/6: 162; all translations are mine)

The context of this passage from Schlick's 1929 article *Erkenntnistheorie und moderne Physik (Erkenntnistheorie and modern physics)* is the following: Schlick contends that the task of epistemology is complete if it can account for scientific knowledge of reality. There is no need for it to consider objects which are not actual. Now this seems precisely what thought experiments like Gettier's do: they make us consider invented situations which are not actual. Thus, Schlick seems to say that philosophers should not bother about thought experiments.

However, there are also remarks which point in the opposite direction, in particular the one originally in French from which I took the title of this article:

The representation of worlds departing from the real one requires a serious effort of imagination... But is fancy a privilege of poets? Don't we have a right to suppose it in philosophers? ([1935] GA I/6: 607)

Here Schlick seems to say that philosophers can be expected to use their imagination just as poets to represent worlds which are different from the real one. The context of this passage, the thought experiment to be considered in section 3.2. below, makes clear that Schlick does not oppose but endorses certain efforts of philosophers to come up with fictive scenarios which diverge from reality. The question becomes how to reconcile these two remarks.

There is one obvious way of reconciling the two quotes. The reconciliatory proposal is that philosophers may consider non-actual scenarios as long as considering them contributes to answering questions about reality, the ones philosophers are interested in. However, this requirement leads to new concerns: given Schlick's empiricist creed, we access reality by experience. We make observations and theorize about them. This is how we answer questions about reality. How could fictive scenarios contribute to such an access?

In order to solve this problem, I shall take inspiration from a key verificationist tenet of Schlick. Thought experiments should contribute to answering questions about reality. Schlick imposes a verificationist constraint on meaningful questions:

A question is in principle answerable (I should like to say: it is a „good question“) if we can imagine the experiences which we would have to have in order to give the answer ([1932] GA I/6: 404)

A question has to be answerable, perhaps not here and now, but in principle. Otherwise it would miss its point as a question. A sufficient (and presumably necessary) condition for a question to be answerable is that we can anticipate the experiences which would allow to answer it in imagination.

In assigning the role of anticipating experience to imagination, Schlick seems to subscribe to a simulationist view of imagination.

Here is a classical statement of this view:

Imaginative projection involves the capacity to have, and in good measure to control the having of, states that are not perceptions or beliefs or decisions or experiences of movements of one's body but which are in various ways like those states—like them in ways that enable the states possessed through imagination to mimic and, relative to certain purposes to substitute for perceptions, beliefs, decisions, and experiences of movements. (Currie and Ravenscroft 2002: 11)

In this quote, Currie and Ravenscroft present the imagination as a capacity of recreating or simulating mental states, among them perceptual states as we enjoy them in experience. Schlick assigns a key role to such a recreative imagination. The ability of using imagination is a prerequisite of our ability to ask and answer questions and thus of any intellectual activity. In order to grasp a question, we must be able to use imagination in anticipating the possible experiences which would serve to answer the question.

The notion of imagination needs to be clarified, though. Many philosophers bind imagination to a capacity to conjure up qualitative states like visual imagery as contrasted to states with purely non-qualitative content like propositions, concepts, and the like (cf. Kind 2001). Schlick seems to agree. This leads him to a qualification of his condition for meaningful questions:

I do not think, for instance, that we can be charged with talking nonsense if we speak of a universe of ten dimensions, or of beings possessing sense organs and having perceptions entirely different from ours; and yet it does not seem right to say that we are able to imagine such beings and such perceptions, or a ten-dimensional world. But we *must* be able to say under what *observable* circumstances we should assert the existence of the beings or sense-organs just referred to. ([1936] GA I/6: 730)

In this quote, Schlick seems to acknowledge that we can ask meaningful questions about a universe of ten dimensions or beings with completely different sensory experiences than ours. Such topics go far beyond our reality. Our reality might eventually turn out to be one with ten dimensions or beings with completely different sensory experiences, but we cannot simply presuppose that it will. Judging from Schlick's criterion for good questions, we should be able to imagine experiences which would make us answer the question whether such scenarios are real in the affirmative. However, we cannot imagine ten dimensions or what it would be like to have completely different sense perceptions, says Schlick. The reason, I surmise, is that imagination is bound to qualitative states we are in a position to recreate. We cannot qualitatively represent ten dimensions or sense experiences which are completely different from ours.

One may wonder why Schlick emphasizes these limits of imagination. One answer is that he considers the claim that we would have to imagine a universe with ten dimensions or beings with different sense

organs in order to imaginatively anticipate the experiences we would have to make to posit the existence of such items. This claim seems doubtful. After all, we might posit these objects by broadly abductive reasoning, i.e. as theoretical entities which explain experiences with quite a different qualitative content. Anything else would amount to an implausibly radical empiricism, according to which theoretical concepts somehow have to be built from qualitative experiences. Although there are some indications that Schlick at a certain point in his career flirted with such a radically empiricist view of conceptual content (Oberdan 1996: Sec. 2), I shall not further discuss this aspect of Schlick's work. Suffice it to say that, judging from the above quote, Schlick does not hold that we would have to imagine objects like a ten-dimensional universe or alien perceptions directly.

In order to make room for meaningful questions about a universe with ten dimensions etc., we do not have to imagine a universe with ten dimensions or what creatures with alien sense organs experience directly but only which experiences of ours would lead to positing such items by way of theorizing on our experiences, for instance by inference to the best explanation. For instance, the best explanation of our actual observations of the physical realm may be to posit ten dimensions, even if we can only perceive three of them. And the best way of accounting for the function of dog whistles may be to claim that dogs can hear sounds in the ultrasonic range, even if we cannot imagine hearing such sounds. Taking into account such indirect ways of imagining the pertinent experiences, we can uphold a definition of meaningful questions in terms of imagination.

In how far do these findings on meaningful questions bear on discerning useful thought experiments? I venture a constructive proposal: a thought experiment is useful precisely if it contributes to answering a meaningful question about reality. For a question to be meaningful, we have to be able to imagine the experiences which would allow us to answer it. I suggest that there is an analogous condition for thought experiments:

Sinnkriterium: a thought experiment is meaningful only if we can imagine the experiences which would confirm to us that the experimental scenario is real.

This is a constructive proposal. In order for us to use a thought experiment, it must contribute to answering a meaningful question about reality. A question is meaningful precisely if we can imagine the experiences that would lead to answering it. It does not follow that we have to consider the experiences which would confirm that the thought experimental scenario is real. Alternatively, one may think of a more indirect relationship between experience and the scenario, in particular that experience only confirms the scenario to be *possible* in some sense. This alternative seems even more plausible, or at least better

in tune with our current ways of thinking about thought experiments. However, Schlick's practice of thought experimenting supports that he indeed had something like my constructive proposal in mind. One may feel uncertain about my broad use of 'meaningful'. Shouldn't the notion be restricted to linguistic meaning? I suggest that we think of thought experimental descriptions which have a linguistic meaning and thus can be assessed as to whether they are meaningful.

I shall now present three thought experiments discussed by Schlick in which I see the *sinnkriterium* at work. In order to perform a good thought experiment, we have to imagine the experiences which would confirm that the experimental scenario is real. Moreover, this imagination should bear on answering the question about reality which the thought experimenter set out to answer.

3. *Three Exemplary Applications*

3.1. *Poincaré's Thought Experiment*

My first example is a thought experiment which Schlick adapts from the physicist Henri Poincaré. The question to be answered is whether physical magnitudes are absolute or only relative to other magnitudes, in particular whether there is absolute space as Newton had it. To answer this question, Poincaré invites us to imagine all spatial structures to suddenly grow by the same proportion:

Imagine that all bodies in the world over night grow to huge size, their dimension is enlarged by the factor 100... I am a Goliath of 180m and use a 15m fountain pen to draw letters on the paper which are several meters high, and in an analogous fashion all other magnitudes in the universe have changed, such that the new world, though enlarged, geometrically resembles the old one. ([1917] GA I/2: 198–199)

Schlick devotes a lengthy discussion to the precise general formulation of this thought experiment. As it stands, it does not answer Poincaré's question. It leaves open key issues like whether the masses of objects also change. Schlick eventually settles for a formulation in terms of a suitable mapping of spacetime points which is in tune with his philosophy of physics. These subtleties do not matter for my general topic. I shall consider the case as originally described under the simplifying assumption of a static universe which only undergoes the sudden transformation described by Poincaré. Moreover, all our measuring devices are assumed to be geometrical ones.

According to the *sinnkriterium*, in order to deal with this thought experiment, we have to imagine the experience which would confirm the scenario to be real, Schlick notes:

What would I feel like after such a change? I wouldn't notice the change. Since all objects have participated in the enlargement, all objects and instruments, we would lack any means to figure out the imagined change. ([1917] GA I/2: 199)

There can be no experience which would confirm that we are in Poincaré's scenario, says Schlick, adding:

This whole change exists only for those who mistakenly argue as if space were absolute... Hence the enlarged universe is not only indistinguishable from the original one, it simply is the same universe, there is no sense in talking of a difference as the absolute size of a body is nothing 'real'. ([1917] GA I/2: 199)

Schlick's remarks show that we need to be careful in applying the *sinnkriterium*. Originally, one would have thought that a meaningful thought experiment requires to imagine the experiences which confirm that the scenario is real. In contrast, Poincaré's thought experiment seems to work precisely by our inability to imagine such experiences. How could that be?

Here is my proposal how to construe the dialectics of Poincaré's thought experiment. Poincaré describes a scenario which his opponent who believes in absolute space is committed to accept as a meaningful thought experiment: all spatial magnitudes might change proportionally although we would be unable to detect that they do. It is not Poincaré but his opponent who must be prepared to perform this thought experiment. Applying the *sinnkriterium*, we notice that the scenario is not meaningful. For we cannot imagine the experiences which would confirm that the scenario is real. This shows that the idea of absolute space leads to absurd consequences: one must accept a scenario as meaningful which does not make sense. Poincaré's opponent would have to answer which experiences would confirm that the scenario is real, but there is no positive answer to this question.

I note that there is also a weaker interpretation of the dialectics. Sometimes it seems as if Schlick had only in mind that unobservable change is nothing that makes sense to a *physicist*, who underlies stronger obligations of supporting her claims by observation. I grant that Schlick at this point may not yet invoke a general *sinnkriterium*. But if we consider how to embed his take on Poincaré's experiment into his overall philosophical position, the more ambitious interpretation seems plausible.

3.2. *Private Psychological States*

My second example can be dubbed the thought experiment of private psychological states. Schlick envisions a scenario in which mental states are completely isolated from physical ones. The initial question which motivates this thought experiment is the following: can statements on mental states be reduced to statements about physical states? The question is motivated. Many pundits nowadays say things like 'the brain feels, thinks...'. If we reason this claim through, we may end with the claim that my state of, say, being in pain, is identical or reduces to a certain physical state of my brain. A related claim is that a sentence like 'some being is in pain' is a sentence about a certain iden-

tifiable physical state. Against this claim, Schlick develops his thought experiment. He imagines a scenario in which there are beings with qualitative mental states. They experience what it is like to be in such a state. But there is no correlation between these qualitative states and anything physical. To Schlick, true sentences about the qualitative states of these beings are not sentences about anything physical. This shows that statements about mental states do not in principle reduce to statements about physical facts.

My interest is not so much the success or failure of this thought experiment, but how the experiment is presented by Schlick. In Schlick's description of the experiment, we can see the *sinnkriterium* at work. Schlick says about his scenario:

We should perhaps talk of two realms, one physical, public, common, the other private, psychological, consisting entirely of monologues... the two worlds would be parallel *but they remain connected*. ([1935] GA I/6: 607, m.e.)

The final sentence is puzzling. Why do the two worlds, the physical and the psychological one, have to be connected? In what way are they connected? And how does this connection square with Schlick's stipulation that there is no correlation between private mental and physical states?

In my view, the answer again lies in the *sinnkriterium*. In order to perform the experiment, we have to ask which experiences would confirm that there are the two separate realms, the physical and the psychological one. Experience would have to confirm to *one and the same* cognizer both that there is the physical realm and the isolated psychological realm. This can only be done if the cognizer herself has access both to the physical and the psychological realm. The cognizer must on the one hand have suitable experiences of the physical world. These experiences must be systematically correlated with physical facts. On the other hand, she must also have a range of experiences which so completely lack any systematic correlation with physical states that she cannot even communicate them to others.

To illustrate the point, assume that there is a community of researchers who lack eyesight. However, only one of them, Mary, additionally has intense psychedelic colour experiences which occur randomly. Since the others do not have colour experiences and there are no physical correlates to Mary's experiences, there is no way she could communicate to the others what she is experiencing. At best the others could notice (if they take her avowals seriously) that she has a random pattern of experiences which they have no further access to.

One may question the experiment by doubting that one can have a private language to be used for monologues about private experiences. More importantly, one may ask what modal status we have to assign to the scenario for it to say something about psychological sentences in our language. Perhaps all true psychological statements in our language are perfectly correlated with physical statements. As a consequence,

one may wonder why the theoretical alternative of psychological statements which are not correlated in this way with physical statements matters if our main interest is to settle questions about reality. But it is not my purpose here to assess the success or failure of the thought experiment. My interest is only to show how Schlick's metatheory of thought experiments drives his interpretation of his own experiment. The combination of access to the physical and the private realm in one and the same cognizer is a direct result of the *sinnkriterium*.

3.3. *Immortality*

My third example is interesting especially due to its aftermath in the history of 20th century philosophy. It illustrates Schlick's empiricist take on the age-old question of human immortality. To appreciate the standing of the question: when we nowadays read Descartes's *Meditations*, we are primarily interested in his *Cogito ergo sum*, his radical doubt, his mind-body dualism. But Descartes himself titled the *Meditations* 'Meditations on the first philosophy, in which the existence of God and the immortality of the soul are demonstrated' (Descartes 1641). The question of immortality seems to have been a primary preoccupation of Descartes. Moreover, he approached the question by purely intellectual inquiry, detached from all sense experience.

As an empiricist, Schlick must take a completely different approach. He addresses the question by asking what experience might contribute to answering it. To Schlick, the main contribution would concern empirical evidence that one has survived one's own bodily death. Here is what this evidence might be like:

In fact I can easily imagine, e.g., witnessing the funeral of my own body and continuing to exist without a body, for nothing is easier than to describe a world which differs from our ordinary world only in the complete absence of all data which I would call parts of my own body. ([1936] GA I/6: 731–733)

Again Schlick develops a thought experimental scenario, one's survival of one's own death, by asking which experiences might confirm that this scenario is real. He envisions the experience of attending his own funeral without having any experiences confirming that he has a body. It is an interesting question how Schlick could imagine watching his funeral without thereby having data about his having an embodied perceptual system standing in physical contact with his physical surroundings. It also sounds strange to talk of parts of the body as data. A realist may insist that parts of the body are not simply bits of information, they are material beings out there. But again, my purpose is not to discuss the minutes of Schlick's experiment. Instead I note that, to Schlick, the scenario of surviving one's own death makes sense only if we can imagine the experiences which would confirm that the scenario is real.

Schlick's empirical twist of the issue of immortality has sparked criticism by Bernard Williams. Williams says:

Schlick famously claimed that survival after death must be a contingent matter, because he could imagine watching his own funeral. In order to make good this claim, Schlick would have had to give a coherent account of how, as participant at his own funeral, he could be himself, Schlick; all the problems of continuity, personal identity, and so forth are called up. (Williams 1973: 40)¹

Williams notes that Schlick took survival of one's own death to be a contingent matter, as contrasted, for instance, to Descartes, who considered it a consequence of the necessary metaphysical structure of the world, not to be ascertained by observation but by purely intellectual inquiry.² I do not think that Schlick aimed at establishing the contingency of survival after death, though. Rather he wanted to bring out the only way to make sense of the very question of immortality: by telling how we could find out whether we are immortal by empirical means.

Williams's second criticism is more interesting: Schlick might have devised a scenario of a subject watching Schlick's funeral, but what could confirm to the subject that it is the same person, Schlick, who is being buried and watching his funeral? I think that this criticism, whatever its ultimate plausibility, is telling. It shows the limitations imposed on thought experiments by Schlick's *sinnkriterium*. Either the question whether it is he himself, Schlick, who observes his funeral, is a meaningful question. Then the question must be principally answerable by experience, for instance by the experienced continuity of psychological states. Or the question does not make sense because it cannot be answered empirically. In this case, Williams criticism might still apply. It might just show that Schlick's thought experiment did not live up to his own standards. He would not have devised conditions that could empirically confirm his surviving his own bodily death. And this failure might indicate that the whole question of immortality becomes meaningless by Schlick's standards.

4. *In Conclusion*

I have illustrated Schlick's *sinnkriterium* for thought experiment by several applications. But what are we to make of this account? Is it just a curious footnote in the history of logical positivism? I shall close with two remarks.

On the one hand, Schlick's criterium surely imposes strong limitations on thought experiments. Nowadays these precise limitations do not seem overwhelmingly plausible. For instance, we presumably can conceive a universe which exists although there is no way of ascertaining its existence by observation, for instance our universe as

¹ Thanks to Bernhard Thöle for bringing this passage from Williams to my attention.

² More precisely, Descartes thought that the soul is necessarily immortal as far as its continuous existence only depends on God's continuous support and not on anything physical.

it would have been if there had been no intelligent life. This thought experiment tells us something about our reality, as witnessed by the discussion of the so-called anthropic principle among scientists. The argument from the anthropic principle roughly goes as follows: there is nothing remarkable about the universe giving rise to conscious life. If it were different, we would simply be unable to observe it. The opposite impression is just due to selection bias. The thought experiment makes sense. But if we apply Schlick's *sinnkriterium*, we would have to ask which observations could confirm the existence of a universe devoid of intelligent life, and, of course, there can be no such observations.

Moreover, there are doubts that Schlick's *sinnkriterium* elucidates our standard way of tackling thought experiments. For instance, surely we can answer which experiences would confirm us that a Gettier case is real. As for GC, we have to observe a person looking at a clock, observation must confirm us that the clock stopped precisely 24h earlier, and so on. But this answer is trite. It does not really capture anything that matters in dealing with this thought experiment. In particular, Schlick offers nothing like the sophisticated apparatus of alethic modalities which structures the current debate of thought experiments. This leaves us somewhat clueless about how merely imagining experience which might be arbitrarily unlikely to become real could tell us anything about our reality. The problem already surfaced at the end of section (3.2.). Schlick does not offer an answer to the crucial question how mere imagination can generate *new* information as Mach did in his proposal that tacit empirical constraints are written into our minds by evolution.

On the other hand, in recent times, empiricist tendencies in modal epistemology are on the rise (e.g. Bueno and Shalkowski 2015, Martínez 2015, Fischer and Leon 2017). Philosophers have become suspicious of lofty possibility claims which are not somehow grounded by empirical science. The same goes for thought experiments. Notwithstanding huge differences in detail, empirically-minded philosophers may be inspired by Schlick's thoroughly empiricist attitude, which also becomes manifest in his account of thought experiments.

References

- Bueno, O. and Shalkowski, S. A. 2015. "Modalism and Theoretical Virtues: Toward an Epistemology of Modality." *Philosophical Studies* 172 (3): 671–89.
- Currie G. and Ravenscroft, I. 2002. *Recreative Minds*. Oxford: Oxford University Press.
- Descartes, R. 1641. *Meditationes de prima philosophia*. Paris: Michel Soly.
- Fischer, B. and Leon, F. (ed.). 2017. *Modal Epistemology After Rationalism*. Cham: Springer.
- Kind, A. 2001. "Putting the Image Back in Imagination." *Philosophy and Phenomenological Research* 62: 85–110.

- Mach, E. 1897. "Über Gedankenexperimente." *Zeitschrift für den Physikalischen und Chemischen Unterricht* 1: 1–5.
- Martínez, M. 2015. "Modalizing Mechanisms." *The Journal of Philosophy* 112: 658–670.
- Miščević, N. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6: 215–226.
- Oberdan, T. 1996. "Postscript to Protocols: Reflections on Empiricism." In A. Anderson and R. Giere (eds.), *Origins of Logical Empiricism*. Minnesota Studies in the Philosophy of Science: Vol. XVI. Minneapolis: University of Minnesota Press: 269–291.
- Ørsted, H.C. 1822. "Oersted über das Studium der allgemeinen Naturlehre." *Journal für Chemie und Physik* 36: 482.
- Schlick, M. [1917]. *Raum und Zeit in der gegenwärtigen Physik*. GA I/2: 159–287 [Moritz-Schlick-Gesamtausgabe I/2. Wien-New York: Springer].
- ____ [1929] *Erkenntnistheorie und moderne Physik*. GA I/6: 161–172.
- ____ [1932] *A New Philosophy of Experience*. GA I/6: 397–414.
- ____ [1935] *De la relation entre les notions psychologiques et les notions physiques*. GA I/6: 583–609.
- ____ [1936] *Meaning and Verification*. GA I/6: 709–749.
- Williamson, T. 2009. "Replies to Ichikawa, Martin and Weinberg." *Philosophical Studies* 145: 465–476.

Thought Experiments in the Theory of Law: The Imaginary Scenarios in Hart's The Concept of Law

MIOMIR MATULOVIĆ
University of Rijeka, Rijeka, Croatia

H. L. A. Hart's The Concept of Law is an important and influential work in the modern philosophy and theory of law. In it, Hart introduced and discussed three imaginary scenarios: the absolute monarchy under the Rex dynasty; the pre-legal society governed by primary rules of obligation; and the worlds in which rules would be different from those in our actual world. Although Hart did not use the expression "thought experiments" in his work, some of his interpreters refer to the imaginary scenarios as thought experiments. However, interpreters do not go into the question of whether the imaginary scenarios in Hart's work do indeed satisfy a general characterization of thought experiments. In this article, the author first summarizes the three imaginary scenarios in Hart's work and points to the context within which we encounter each of them. Then, he makes use of a general characterization of thought experiments in the contemporary philosophical literature and briefly examines the way and the extent to which the imaginary scenarios in Hart's work can satisfy its requirements.

Keywords: Thought experiments, philosophy of law, H. L. A. Hart, imaginary scenarios.

1. Introduction

Are there any thought experiments in law? If we look for an answer to this question in the recently published prestigious *The Routledge Companion to Thought Experiments* (hereinafter *RCTE*), we will not find one. Its Part II contains discussions on thought experiments in particular disciplines, such as political philosophy, economics, theology, ethics, physics, biology and mathematics, but not in law.¹ Neither

¹ The editors of *RCTE* advocate a further expansion of the discussion to thought experiments in other disciplines, including law (see Stuart, Fehige and Brown (2018: 3, 5)).

does the ambitious work on the history of philosophy and theory of law, their orientations and main topics, *A Treatise of Legal Philosophy and General Jurisprudence* (hereinafter *TLPJ*) in 12 volumes, contain a discussion on thought experiments in law. It uses the expression “thought experiment” twice: first, to refer to Otfried Höffe’s state of nature thought experiment (Hofmann 2016: 335) and, second, to refer to F. K. von Savigny’s idea of the process of statutory interpretation (Chiassoni 2016b: 584). However, when discussing the contribution to the legal theory of the two great contemporary philosophers, John Rawls and Jürgen Habermas, *TLPJ* fails to inform us that the original position of the first and the idealized speech situation of the second are also thought experiments (Riley 2009). Considering that these two publications, each highly respectable in its field of research, are silent or scarcely informative about thought experiments in law, any further information about the topic is welcome.

I will be focusing here on H. L. A. Hart’s *The Concept of Law*, (hereinafter *CL*), (1961 [1994]). It is an important and influential work in the modern philosophy and theory of law.² In *CL*, Hart introduced and discussed, among other things, several imaginary scenarios: the absolute monarchy under the Rex dynasty (52–66); the pre-legal society governed by primary rules of obligation (91–99); and the worlds in which rules would be different from those in our actual world (193–200). These imaginary scenarios are designed by Hart to dismiss the theories of law of some other authors, such as John Austin (1832 [1954]),³ and to present some of the main ideas of his own theory. First, in his discussion of the imaginary absolute monarchy under the Rex dynasty, Hart demonstrates the inadequacy of the central notions of Austin’s theory of law, those of sovereignty and general habit of obedience, and the indispensability of the idea of rules for the understanding of law. Then, in his discussion of the imaginary pre-legal society governed by primary rules of obligation and the imaginary worlds in which rules would be different from those in our actual world, Hart presents two other main ideas of his theory of law, in addition to the idea of rules, that is, the idea of law as the union of primary rules of obligations and secondary rules of power, and the idea of minimum content of natural law. Thus, all the three main ideas of Hart’s theory of law turn around imaginary scenarios.

Although Hart did not use the expression “thought experiments” in *CL*, some of Hart’s interpreters refer to his imaginary scenarios as thought experiments and/or suggest that the method of thought experimentation is coequal to linguistic methods in his work. For ex-

² On Hart’s life and work, see the following books and collections of essays in English: d’Almeida, Edwards and Dolcetti (2013); Postema (2011); Simpson (2011); Kramer, Grant, Colburn and Hatzistavrou (2008); MacCormick (2008); Lacey (2004); Coleman (2001); Bayles (1992); Leith and Ingram (1988); Gavison (1987); Moles (1987); Hacker and Raz (1977). For other works on Hart’s *CL*, see references.

³ He should not be confused with Hart’s philosopher colleague J. L. Austin.

ample, Nicos Stavropoulos argues that Hart in *CL*, in addition to an examination of the semantics of imperatives,⁴

... further employs less obviously linguistic methods, namely, the pursuit of philosophical argument as to the true content of the key concepts, by means of the familiar techniques of drawing distinctions and defending them through thought experiments. (2001: 67)

Another author, Pierluigi Chiassoni, claims that “Hart regards the method of philosophical imagination as a major tool in the game of descriptive metaphysics”, the method that “[i]n Hart’s understanding ... requires the working out of thought experiments meant to explain how our actual conceptual and institutional structures are, and why, by comparing them with alternative imaginary situations” (2011: 65; 2013: 456).⁵

Unlike Stavropoulos, who does not give any concrete example of thought experiments in Hart’s *CL*, Chiassoni lists three thought experiments in Hart’s work: the simple model of law as coercive orders; the idealized picture of a primitive, pre-legal, society ruled only by a set of unconnected primary rules of obligation; and the theory of the minimum content of natural law. (They are the same as the imaginary scenarios that I have mentioned above under slightly different appellations.) However, Chiassoni does not discuss the thought experiments in detail. Still, other authors (Priel 2013: 544; Giudice 2015: 59; von Daniels 2016: 109–112; and Houlgate 2017: 51) refer to some of the imaginary scenarios in Hart’s *CL* as thought experiments.

These authors usefully draw attention to the imaginary scenarios in Hart’s *CL* as thought experiments and the thought experimentation as an integral part of the methodology that underlies his work.⁶ However, they do not go into the question of whether Hart’s imaginary scenarios in *CL* do indeed satisfy a general characterization of thought experiments. Until we have an answer to this question in the first place, we cannot say whether their interpretation of Hart’s work is on the right track.

⁴ According to Stavropoulos (2001: 67), Hart’s examination of semantics of imperatives consists of an analysis of “... the meaning of expressions such as ‘to order’ ... and ‘to give an order’ ... ‘to address’, as applied to commands ... and to laws ... ‘obedience’ ... ‘being obliged’, ‘having an obligation’, and ‘duty’ ... and ‘valid’. It also results in the truth conditions of propositions such as ‘a legal system exists’ ... ‘it is the law that X’ and ‘in England they recognize as law X’ ... ‘rule X is valid’ ... or those expressing the existence of obligations ... and a number of other expressions and propositions” (page references to Hart’s *CL* are omitted).

⁵ Chiassoni (2016: 64) also uses expressions “mental experiments” and “experiments in ‘philosophical imagination’”.

⁶ In her biography of Hart, Nicola Lacey tells us that Hart was famous for inventing games of wit or ingenuity that he and his wife Jenifer used to play with their friends Isaiah and Aline Berlin: “The game consisted in a thought experiment in which the Harts’ and the Berlins’ guests wake up to find themselves with the other family, and involved wry comparisons of the comportment of their respective guests” (Lacey 2004: 340). However, Lacey says nothing about Hart’s attitude towards thought experimenting in his professional work as opposed to his leisure-time activity.

In the section 2, I will first summarize the three imaginary scenarios in Hart's *CL* and point to the context of his work within which we encounter each of them. Then, in the section 3, I will make use of a general characterization of thought experiments in the contemporary philosophical literature, including the aforementioned *RCTE*, and briefly examine the way and the extent to which the imaginary scenarios in Hart's work can satisfy its requirements.

Before I do this, let me note that in other works of his, Hart introduced and discussed several other imaginary scenarios: the criminal law without excusing conditions (1958 [2008]: 47–8); the counterfactual causation (1959 [1985]) and the pure theory of imperatives (1970 [1983]: 312–13). Hart has also discussed some imaginary scenarios and thought experiments invented by other authors: F. W. Maitland's state of Nusquamia (1954 [1983]: 37–39); L. L. Fuller's Rex the lawmaker (1965 [1983]: 347–53); Rawls's original position (1973 [1983]); Robert Nozick's emergence of minimal state (1976 [1983]: 150–51) and Ronald Dworkin's Hercules the judge (1977 [1983]: 139, 154; 1961 [1994]: 264). Hart termed "Gedankenexperiment" ("thought experiment"), (1958 [2008]: 47) an imaginary situation in which criminal law is operating without excusing conditions, this being the only occasion in his works, as far as I know, that he used the expression. Of course, from the fact that Hart terms the criminal law without excusing conditions as a "thought experiment", it still does not follow that it really is a thought experiment. If we want to argue that the latter is a thought experiment, we will require a characterization of a genuine thought experiment. We will also require such a characterization, if we want to argue that other imaginary scenarios in Hart's works are thought experiments, even though Hart does not call them thought experiments. I have already said I will examine here only the imaginary scenarios in Hart's *CL* as possible candidates for thought experiments. Last but not least, it has to be noted that Hart is credited with having revived the contemporary debate on the doctrine of double effect (1967 [2008]) from which some famous thought experiments emerged, such as the terror bomber and strategic bomber thought experiment and the tram/trolley thought experiment (Di Nucci 2004).

2. *Three imaginary scenarios*

In this section, I summarize the imaginary scenarios in Hart's *CL* that this article deals with and point to the context of his work within which we encounter each of them. They are:

- the absolute monarchy under the Rex dynasty;
- the pre-legal society governed by primary rules of obligation; and
- the worlds in which rules would be different from those in our actual world.

The absolute monarchy under the Rex dynasty

In Chapters II–IV of *CL*, Hart states and criticizes Austin’s theory of law.⁷ On Austin’s theory, Hart writes in the first of these two chapters:

... there must, wherever there is a legal system, be some persons or body of persons issuing general orders backed by threats which are generally obeyed, and it must be generally believed that these threats are likely to be implemented in the event of disobedience. This person or body must be internally supreme and externally independent. If, following Austin, we call such a supreme and independent person or body of persons the sovereign, the laws of any country will be the general orders backed by threats which are issued either by the sovereign or subordinates in obedience to the sovereign. (25)

In Chapter III, Hart criticizes Austin’s notion of general orders backed by the threat of sanction by arguing that it is inadequate to account for the content, the mode of origin and the range of application of many laws that modern legal systems contain. He then goes on in Chapter IV to criticize Austin’s theory on the ground that its notions of sovereignty and general habit of obedience are inadequate to account for the continuity and the persistence of law, as well as legal limitations on legislative authority.

In this context, Hart introduces and discusses the imaginary absolute monarchy under the Rex dynasty. He asks the reader to imagine a society in which Rex is an absolute monarch or sovereign (52). Rex is habitually obeyed by the bulk of his subjects, but he habitually obeys no one else. He exercises power over his subjects by issuing general orders backed by the threat of sanction, requiring them to do various things and to abstain from doing certain other things. Although some incidents of disobedience took place during the early years of the reign of Rex, the problems have resolved themselves and the subjects settled into a habit of obeying his orders. Hart asks the reader to suppose further, namely, that after a long successful reign, Rex dies leaving a son, Rex II, who immediately starts to issue orders (53). The questions of continuity and persistence of law arise: “Would the orders of Rex II be already law?” and “Would the orders of the dead Rex still be law?” (62). Answering these questions in the manner Austin’s theory requires, leads to an absurd conclusion. First, Rex II has not reigned long enough for the subjects to have had time to develop a habit of obeying his orders. More importantly, the mere fact that the subjects habitually obeyed the orders of his father does not confer on Rex II any right to succeed him and issue orders in his place (59–60). Therefore,

⁷ Hart claims (1961 [1994]: 18) that he is considering and criticizing a modified version of Austin’s theory in its strongest form, and not the theory as Austin formulated it in *The Province of Jurisprudence Determined* (1832 [1954]). In a lengthy note (1961 [1994]: 282–283), Hart enumerates the additions, modifications and qualifications he made to Austin’s theory.

Rex II would not be a sovereign and his orders would not already be law. On the other hand, being dead, Rex is no longer habitually obeyed. Therefore, neither would he be a sovereign nor would his orders still be law. The absurd consequence is that, on Austin's theory, the imaginary absolute monarchy over which Rex has reigned would end up without a sovereign and without law, at least until the subjects settle into a habit of obeying Rex II and his orders.

However, even in an absolute monarchy, Hart claims, there must be some accepted fundamental rules specifying a class or line of persons whose word is to constitute a standard of behaviour for the society, i.e. who have the right to legislate. Such a rule, though it must exist now, may in a sense be timeless in its reference: it may not only look forward and refer to the legislative operation of a future legislator but it may also look back and refer to the operations of a past one. (62–3)

The same is true of the imaginary absolute monarchy under the Rex dynasty:

Each of a line of legislators, Rex I, II, and III, may be qualified under the same general rule that confers the right to legislate on the eldest living descendant in the direct line. When the individual ruler dies his legislative work lives on; for it rests upon the foundation of a general rule which successive generations of the society continue to respect regarding each legislator whenever he lived. In the simple case Rex I, II, and III, are each entitled, under the same general rule, to introduce standards of behaviour by legislation. (63)

The answer to the questions of continuity and persistence of law, "Would the orders of Rex II be already law?" and "Would the orders of the dead Rex still be law?", leads now to no absurd consequence. The absurdity is removed by assuming that there is a general rule recognizing the enactments of each legislator in the direct lineal succession, Rex I and Rex II, as law.⁸

In the rest of Chapter IV, Hart questions the necessity of a sovereign with legally illimitable power for the existence of law as well as the very possibility in modern legal systems of a sovereign in Austin's sense. As a substitute for Austin's notions of sovereignty, general orders backed by the threat of sanction and general habit of obedience, Hart introduces the idea of rules, without which "we cannot hope to elucidate even the most elementary forms of law" (80).

The pre-legal society governed by primary rules of obligation

After demonstrating the inadequacy of Austin's theory for the understanding of law, Hart announces in Chapter V of *CL* "a fresh start" (79, 80). The starting point of such a fresh start is the distinction between two types of rules:

⁸ On his discussion of the continuity and the persistence of law in the context of extended imaginary scenario where Rex dynasty has been overthrown in revolution and Brutus becomes new sovereign, see Hart (1965 [1983]: 362–63).

Under rules of the one type, which may well be considered the basic or primary type, human beings are required to do or abstain from certain actions, whether they wish to or not. Rules of the other type are in a sense parasitic upon or secondary to the first; for they provide that human beings may by doing or saying certain things introduce new rules of the primary type, extinguish or modify old ones, or in various ways determine their incidence or control their operations. Rules of the first type impose duties; rules of the second type confer powers, public or private. Rules of the first type concern actions involving physical movement or changes; rules of the second type provide for operations which lead not merely to physical movement or change, but to the creation or variation of duties or obligations. (81)

Hart claims that

in the combination of these two types of rule there lies what Austin wrongly claimed to have found in the notion of coercive orders, namely, ‘the key to the science of jurisprudence.’ (81)

In the rest of Chapter V, Hart examines the two types of rules. Firstly, he characterizes the primary rules of obligations in terms of seriousness of social pressure for compliance, importance for the preservation of social life, and conflict with self-interests of those controlled by them. In addition, Hart elaborates the distinction between the internal and external point of view (introduced earlier in *CL* (see 56–7). He then goes on to examine the secondary rules of power.

In this context, Hart introduces and discusses the imaginary society without sovereign and his subordinates (91). He is asking the reader to imagine a primitive society governed by a set of primary rules of obligation that forbid the free use of violence, theft, and deception, prescribe performance of various services and contributions to the common life, etc. (See more on the content of these rules below.) Some members of society reject these rules or conform to them only out of fear of sanction. They take the rules from the external point of view. However, the vast majority of society’s members accept the rules and obey them. They “live by the rules seen from the internal point of view” (92). Unless such a society is small, made up of a close-knit population sharing common sentiment and belief, and placed in a stable environment, Hart contends, its formal structure “must prove defective and will require supplementation in different ways” (92). One defect would be the uncertainty of the rules. If doubts were arisen as to what the rules are or as to the precise scope of a given rule, there would be no procedure for settling this doubt, either by reference to an authoritative text or to a society’s institution whose declarations on this point are authoritative. Another defect would be the static character of the rules. There would be no means of deliberately introducing new rules or adapting or eliminating old ones in the light of changing circumstances in a society. In an extreme case, which “never perhaps fully realized in any actual community”, the obligations, specified in the rules, “in particular cases could not be varied or modified by the deliberate choice of any individual” (93). The last defect would be the inefficiency of the rules. If

disputes were arisen as to whether a rule has or has not been violated, there would be no society's institution specially empowered to ascertain finally and authoritatively the fact of violation (93–4).

The remedy for each of these defects of primary rules of obligation, Hart claims, would consist in the introduction of secondary rules of power. First, the remedy for the uncertainty of primary rules would be the introduction of the rule of recognition. It “specify some feature or features possession of which by a suggested rule is taken as a conclusive affirmative indication that it is a rule of the group to be supported by the social pressure it exerts” (94). For example, in the imaginary absolute monarchy under the Rex dynasty, the rule of recognition would be that whatever Rex I enacts is law (96). Second, the remedy for the static character of primary rules would consist in the introduction of the rules of change. They empower “an individual or body of persons to introduce new primary rules for the conduct of the life of the group, or of some class within it, and to eliminate old rules” (95). Again, in the case of the imaginary absolute monarchy under the Rex dynasty, the rule of change would be that the eldest living male descendant, in the direct line, of Rex I, has the right to legislate. A further remedy for the static character of primary rules would be the introduction of rules that confer on individuals the power to vary their initial positions under the primary rules. These rules are akin to the rules of change involved in the notion of legislation (96). Third, the remedy for the ineffectiveness of the primary rules would consist in the introduction of the rules of adjudication. They empower “individuals to make authoritative determinations of the question whether, on a particular occasion, a primary rule has been broken” (97).

Hart marks the introduction of all these secondary rules of power into society as “the step from the pre-legal into the legal world,”⁹ and equates its importance for a society with the invention of the wheel (42).

In Chapter VI of *CL*, Hart examines the rule of recognition in more detail and still further elaborates the distinction between the internal and external point of view. Of particular interest here is his remark about a legal system in which only officials take the rules from the internal point of view:

The society in which this was so might be deplorably sheeplike; the sheep might end in the slaughter-house. But there is little reason for thinking that it could not exist or for denying it the title of a legal system. (117)

The point is, as Hart makes clear later in *CL*, that the step from the pre-legal society to one with law would bring with it not only gains (cer-

⁹ There would be borderline cases where some, but not all, secondary rules of power are introduced. Namely, Hart claims that the introduction of each of the secondary rule of power “might, in itself, be considered as a step from the pre-legal into the legal world”, since each one “brings with it many elements that permeate law”, while “certainly all three [secondary rules of power] together are enough to convert the regime of primary rules into what is indisputably a legal system” (94).

tainty, dynamism and efficiency of rules), but also the cost of risk that “the centrally organized power may well be used for the oppression of numbers with whose support it can dispense, in a way that the simpler regime of primary rules could not” (202). Does not the same point hold true for all inventions? Think of the above-mentioned analogy with the invention of the wheel.

*The worlds in which rules would be different
from those in our actual world*

After considering law as the union of primary and secondary rules, Hart redirects his attention in Chapters VIII and IX of *CL* to the consideration of the relation between law and morality.¹⁰

In the first of these two chapters, he discusses the relevance of the idea of justice to law, the main features (importance, immunity from deliberate change, voluntary character of offences, and the form of pressure) that distinguish moral rules from legal rules and other types of social standards, and the role that moral ideals and principles play in a society and life of individuals.

Hart then goes on in Chapter IX to consider the contention that there is no necessary connection between law and morality. This contention is most closely associated with the tradition of legal positivism. Hart distinguishes between two forms in which the contention has been rejected.

One of these is expressed most clearly in the classical theories of Natural Law: that there are certain principles of human conduct, awaiting discovery by human reason, with which man-made law must conform if it is to be valid. The other takes a different, less rationalist view of morality, and offers a different account of the ways in which legal validity is connected with moral value. (186)

Hart devotes most of the chapter to an examination of the traditional natural law theory. Although he rejects its teleological view of nature, he accepts its claim about human survival as a goal of law and morality:

We are committed to it as something presupposed by the terms of the discussion; for our concern is with social arrangements for continued existence, not with those of a suicide club. (192)

Proceeding on this assumption, Hart begins to specify what he calls the “*minimum content of natural law*” (193).¹¹

In this context, Hart introduces and discusses several imaginary worlds in which rules would be different from those in our actual world.¹² First, he asks the reader to imagine a world in which human

¹⁰ Although the “union of primary and secondary rules is at the centre of a legal system”, Hart claims, “it is not the whole” (99).

¹¹ Hart stresses (1961 [1994]: 303) that his idea of minimum content of natural law is based on Thomas Hobbes’s and David Hume’s accounts of laws of nature.

¹² See also Hart’s slightly earlier work (1958 [1983]: 79–81).

beings were to become invulnerable to attack by each other, were armored perhaps like “animals whose physical structure (including exoskeletons or a carapace) renders them virtually immune from attack by other members of their species” or were incapacitated like “animals who have no organs enabling them to attack” (194). In such a world there would be little point, Hart claims, “for the most characteristic provision of law and morals: Thou shalt not kill” (195). Second, the reader is asked to imagine a world of human beings “immensely stronger than others or better able to dispense with rest, either because they are far above the present average, or because most were far below it” (195). In such a world of “giants among pygmies”, there would be no special system of organized sanctions, but only a system “in which the weak submitted to the strong on the best terms they could make and lived under their ‘protection’” (198). Third, Hart asks the reader to imagine still another world with devils “dominated by a wish to exterminate each other” and an opposite world with angels “never tempted to harm others” (196). In the former world, rules requiring forbearances would be impossible, while in the latter one they would be unnecessary. Fourth, the reader is asked to imagine a world in which the “human organism ... have been constructed like plants, capable of extracting food from air, or what it needs ... have grown without cultivation in limitless abundance” (196). In such a world, there would be no point in having rules that protect property. The last imaginary world is a pre-legal world that we encountered above. In this world human beings were approximately equal in physical strength and vulnerability, and live under “a system of mutual forbearances”. Because of obvious advantages of submission to such a system, “the number and strength of those who would co-operate voluntarily” in its maintenance would “normally be greater than any likely combination of malefactors” (197–98; see also 218–19). In such a world there would be no need for a special system of organized sanctions.

Considerations of these imaginary worlds enable Hart to make five “very obvious generalizations” or “truisms” about the human nature and the character of physical world in which they live: human vulnerability; approximate equality; limited altruism; limited resources and limited understanding and strength of will. Given these five truisms, Hart claims, certain rules necessary for human survival can be determined. They include rules that “restrict the use of violence in killing or inflicting bodily harm” (194), require “mutual forbearance and compromise” (195) and respect for property (196), enable “men to transfer, exchange, or sell their products” and secure the recognition of promises as a source of obligation” (197), and create a special organization for the detection and punishment “of those who would ... try to obtain the advantages of the system without submitting to its obligations” (198). Hart calls these rules the minimum content of natural law.¹³

¹³ These rules, Hart writes, “are so fundamental that if a legal system did not have them there would be no point in having any other rules at all” (1958 [1983]: 80).

3. *Reconstructing legal theoretical thought experiments*

After having summarized, the three imaginary scenarios in Hart's *CL* and pointed to the context within which we encounter each of them in his work, I turn now to the main question of my article: Do the three imaginary scenarios in Hart's *CL* constitute thought experiments? In the discussion of this question, I rely on a general characterization of thought experiments in philosophy according to which they are:

- imaginary;
- counterfactual scenarios;
- designed for special cognitive purposes.¹⁴

Consider how Hart's scenarios satisfy the requirements of the above characterization.

First, Hart's scenario of the absolute monarchy under the Rex dynasty describes the imaginary monarchy. Hart claims that it is "probably far too simple ever to have existed anywhere" (53) and interchangeably calls it "imagined community" (52), "imaginary monarchy" (54), "imaginary simple world" (67), "imagined society" (68), and "imaginary kingdom" (96). As we saw in the previous section, the scenario reveals absurdities inherent in Austin's notions of sovereignty and general habit of obedience, and suggests that the reader would consider the idea of rules as the way out of the absurdities.

Second, there is a certain ambiguity in Hart's scenario of the pre-legal society governed by primary rules of obligation. Hart writes that

there are many studies of primitive communities which not only claim that this possibility is realized but depict in detail the life of a society where the only means of social control is that general attitude of the group towards its own standard modes of behaviour in terms of which we have characterized rules of obligation. (91)¹⁵

However, the scenario does not describe primitive communities without law which ever did or do now exist, but the imaginary pre-legal regime of primary rules of obligation. In "Postscript" to *CL*, Hart calls it "imagined simple regime consisting only of primary rules of obligation" (249) and "imagined pre-legal regime of custom-type primary rules of obligation" (251). As we saw in the previous section, the scenario's essential elements are defects of such a regime, even those that "never perhaps fully realized in any actual community" (93). These defects are uncertainty, unchangeability and inefficiency, and the remedies for them are secondary rules of recognition, change and adjudication. Furthermore, the introduction of all these secondary rules of power together makes the step from the pre-legal into the legal world. The scenario suggests

¹⁴ For more on this general characterization of thought experiments, see Gendler (2004: 1155), Roux (2011: 19–27), and Goffi and Roux (2018: 440–41). See also Mišćević's discussion on thought experiments in political philosophy (2018; 2017) and Brun's discussion on thought experiments in ethics (2018).

¹⁵ Hart cites several such works in an accompanying note (1961 [1994]: 291).

that the reader would consider the idea of law as the union of primary rules of obligations and secondary rules of power.

Several authors have noticed a certain similarity between Hart's scenario and John Locke's account of the state of nature (Sartorius (1966 [1971]: 140); Bobbio (1968 [1988]: 70); Hacker (1977: 11); Fitzpatrick (1992: 193); Postema (2011: 306); Simpson (2011: 174–77); Chiassoni (2013: 456)). Namely, in the *Second Treatise of Government*. Locke claims that the state of nature is defective, inconvenient to use his euphemism, because it lacks “an *establish'd*, settled, known *Law*”, “*a known and indifferent Judge*” and “*Power to back and support the Sentence when right, and to give it due Execution*” (1689 [1988]: 351). For these defects of the state of nature, Locke writes, “*Civil Government is the proper Remedy*” (276). However, Hart does not mention Locke nor take any notice of the state of nature in *CL*.

Third, Hart's scenarios of imaginary worlds in which rules would be different from those in our actual world are glaring examples of philosophical fantasy (195).¹⁶ As we saw in the previous section, these scenarios refer to specific features of animals, fantastic beings and natural conditions, such as invulnerability, inequality, unlimited altruism, unlimited selfishness, unlimited resources, unlimited understanding and strength of will, which make them different from actual human beings and the world in which they live. The scenarios suggest first that the reader would consider the most characteristic rules of law and morality to be different, if human beings and their natural conditions had any of the specific features above. Furthermore, they suggest that in considering this, she would also consider these rules as rooted in the physical world and our human nature. Finally, the scenarios suggest that the reader would consider the ongoing survival of a human society as contingent upon the most characteristic minimum content of natural law.

The germ of Hart's scenarios of imaginary words in which rules would be different from those in our actual world can be found in Plato's story of Gyges' ring. Namely, in *The Republic*, Plato has Glaucon tell story about the ring which makes its wearer invisible to others human beings. One of the lessons to draw from this story is that in a world in which one were invisible there would be little point for the most characteristic rules of law and morality.¹⁷ Hart in *CL* mentions Plato twice (162, 186), but does not make use of his story.

4. Conclusion

Taking all the above points together, I conclude that Hart's imaginary scenarios in *CL* fulfill the requirements of the general characterization of thought experiments that we can find in the contemporary philo-

¹⁶ The discussion of these worlds, Hart writes, “involves the use of a philosophical fantasy” (1958 [1983]: 79).

¹⁷ On Plato's story of Gyges' ring as a thought experiment, see Becker (2018) and Mišević (2012).

sophical literature. Thus, it is revealed that Hart's interpreters are right, namely those who draw attention to the imaginary scenarios in Hart's *CL* as thought experiments and the thought experimentation as an integral part of the methodology that underlies his work. Hart's work should really be considered as a great example of the thought experimentation in the contemporary theory of law. However, the question remains as to how much Hart's thought experiments fulfill the basic desiderata for good or successful thought experiments. I have to leave this question to be considered on another occasion. Its discussion would also require the consideration of various objections to Hart's ideas of rules, union of primary and secondary rules and minimum content of natural law that are contained in the almost immeasurable literature on Hart's *CL* published in the last fifty and more years after its first edition.

As always, the advice of Nenad Mišćević has proved to be more than useful, while on this occasion I am especially grateful to him for his immense patience shown while waiting for this article to be finished.

References

- Austin, J. 1832 [1954]. *The Province of Jurisprudence Determined and the Uses of the Study of Jurisprudence*. London: Weidenfeld and Nicolson.
- Bayles, M. D. 1992. *Hart's Legal Philosophy*. Dordrecht: Kluwer Academic Publisher.
- Becker, A. 2018. "Thought Experiments in Plato." In Stuart, Fehige and Brown 2018: 44–56.
- Bobbio, N. 1968 [1988]. "Ancora sulle norme primarie e norme secondary." *Rivista di Filosofia* 59: 35–53. Translated in, and cited from, Norberto Bobbio, *Eseji iz teorije prava*, edited by N. Visković. Split, Logos, 1988: 63–74.
- Brun, G. 2018. "Thought Experiments in Ethics." In Stuart, Fehige and Brown 2018: 195–210.
- Chiassoni, P. 2011. "The Simple and Sweet Virtues of Analysis. A Plea for Hart's Metaphilosophy of Law." *Problema. Anuario de Filosofía y Teoría del Derecho* 5: 53–80.
- Chiassoni, P. 2013. "The Model of Ordinary Analysis." In L. D. d'Almeida, J. Edwards and A. Dolcetti (eds.). *Reading HLA Hart's The Concept of Law*. Portland: Hart Publishing: 444–82.
- Chiassoni, P. 2016a. "Supporting The Force of Law: A Few Complementary Arguments Against Essentialist Jurisprudence." In Ch. Bezemek and N. Ladavac (eds.). *The Force of Law Reaffirmed. Frederick Schauer Meets the Critics*. New York: Springer: 61–71.
- Chiassoni, P. 2016b. "The Heritage of the 19th Century: The Age of Interpretive Cognitivism." In E. Pattaro and C. Rovarsi (eds.). *A Treatise of Legal Philosophy and General Jurisprudence, Volume 12: Legal Philosophy in the Twentieth Century: The Civil Law World, Tome 2: Main Orientations and Topics*. New York: Springer: 565–600.
- Coleman, J. (ed.). 2001. *Hart's Postscript*. Oxford: Oxford University Press.

- d'Almeida, L. D., Edwards, J. and Dolcetti, A. (eds.). 2013. *Reading HLA Hart's The Concept of Law*. Portland: Hart Publishing.
- Di Nucci, E. 2004. *Ethics Without Intention*. London: Bloomsbury.
- Fitzpatrick, P. 1992. *The Mythology of Modern Law*. London—New York: Routledge.
- Gavison, R. (ed.). 1987. *Issues in Contemporary Legal Philosophy. Essays for H. L. A. Hart*. New York: Oxford University Press.
- Gendler, T. S. 2004. "Thought Experiments Rethought—and Reperceived." *Philosophy of Science* 71: 1152–1163.
- Giudice, M. 2015. *Understanding the Nature of Law. A Case for Constructive Conceptual Explanation*. Cheltenham: Edward Elgar Publishing, Inc.
- Goffi, J.-Y. and Roux, S. 2018. "A Dialectical Account of Thought Experiments." In Stuart, Fehige and Brown 2018: 439–53.
- Hacker, P. M. S. 1977. "Hart's Philosophy of Law." In P. M. S. Hacker and J. Raz (eds.). *Law, Morality and Society. Essays in Honour of H. L. A. Hart*. Oxford: Oxford University Press: 1–25.
- Hart, H. L. A. 1954 [1983]. "Definition and Theory in Jurisprudence." *Law Quarterly Review* 70: 37–60. Reprinted in, and cited from, Hart 1983: Essay 1.
- Hart, H. L. A. 1958 [1983]. "Positivism and the Separation of Law and Morals." *Harvard Law Review* 71: 593–629. Reprinted in, and cited from, Hart 1983: Essay 2.
- Hart, H. L. A. 1958 [2008]. "Legal Responsibility and Excuses." In S. Hook (ed.). *Determinism and Freedom. In the Age of Modern Science*. New York: Collier Books: 95–116. Reprinted in, and cited from, Hart 2008: Chapter II.
- Hart, H. L. A., Honoré, T. 1959 [1985]. *Causation in the Law*. Second Edition. Oxford: Clarendon Press.
- Hart, H. L. A. 1961 [1994]. *The Concept of Law*. Second Edition with a Postscript edited by P. A. Bulloch and J. Raz. Oxford: Oxford University Press.
- Hart, H. L. A. 1965 [1983]. "Lon L. Fuller: *The Morality of Law*." *Harvard Law Review* 78: 1281–1296. Reprinted in, and cited from, Hart 1983: Essay 16.
- Hart, H. L. A. 1967 [2008]. "Intention and Punishment." *Oxford Review* 4: 5–22. Reprinted in, and cited from, Hart 2008: Chapter V.
- Hart, H. L. A. 1970 [1983]. "Kelsen's Doctrine of the Unity of Law." In H. E. Kiefer and M. K. Munitz (eds.). *Ethics and Social Justice*. Albany: State University of New York Press: 171–99. Reprinted in, and cited from, Hart 1983: Essay 15.
- Hart, H. L. A. 1973 [1983]. "Rawls on Liberty and Its Priority." *The University of Chicago Law Review* 40: 534–55. Reprinted in, and cited from, Hart 1983: Essay 10.
- Hart, H. L. A. 1976 [1983]. "1776–1976: Law in the Perspective of Philosophy." *New York University Law Review* 51: 538–551. Reprinted in, and cited from, Hart 1983: Essay 5.
- Hart, H. L. A. 1977 [1983]. "American Jurisprudence through English Eyes: The Nightmare and the Noble Dream." *Georgia Law Review* 11: 969–989. Reprinted in, and cited from, Hart 1983: Essay 4.

- Hart, H. L. A. 1983. *Essays in Jurisprudence and Philosophy*. Oxford: Clarendon Press, Essay 1.
- Hart, H. L. A. 2008. *Punishment and Responsibility. Essays in the Philosophy of Law*, Second Edition with an Introduction by J. Gardner. Oxford: Oxford University Press.
- Hofmann, H. 2016. "The Development of German-Language Legal Philosophy and Legal Theory in the Second Half of the 20th Century." In E. Pattaro and C. Roversi (eds.). *A Treatise of Legal Philosophy and General Jurisprudence*, Volume 12: *Legal Philosophy in the Twentieth Century: The Civil Law World*, Tome 2: *Main Orientations and Topics*. New York: Springer: Chapter 10.
- Houlgate, L. D. 2017. *Philosophy, Law and the Family. A New Introduction to the Philosophy of Law*. New York: Springer.
- Kramer, M., Grant, C., Colburn, B. and Hatzistavrou, A. (eds.). 2008. *The Legacy of H. L. A. Hart. Legal, Political and Moral Philosophy*. Oxford: Oxford University Press.
- Lacey, N. 2004. *A Life of H. L. A. Hart. The Nightmare and the Noble Dream*. Oxford: Oxford University Press.
- Leith, Ph. and Ingram, P. (eds.). 1988. *The Jurisprudence of Orthodoxy. Queen's University Essays on H. L. A. Hart*. London: Routledge.
- Locke, J. 1689 [1988]. *Two Treatises of Government*. Edited with an introduction and notes by P. Laslett. Cambridge: Cambridge University Press.
- MacCormick, N. 2008. *H. L. A. Hart*. Second edition. Palo Alto: Stanford University Press.
- Miščević, N. 2012. "Plato's Republic as a Political Thought Experiment." *Croatian Journal of Philosophy* 12: 153–165.
- Miščević, N. 2017. "Accounting for Thought Experiments—25 Years Later." In B. Borstner and S. Gartner (eds.). *Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević*. Newcastle upon Tyne: Cambridge Scholars Publishing: 11–29.
- Miščević, N. 2018. "Thought Experiments in Political Philosophy." In Stuart, Fehige and Brown 2018: 153–70.
- Moles, R. N. 1987. *Definition and Rule in Legal Theory: A Reassessment of H. L. A. Hart and the Positivist Tradition*. Oxford: Basil Blackwell.
- Plato 1937. *The Republic, Two Volumes*. Translated by P. Shorey. Cambridge: Harvard University Press.
- Postema, G. J. 2011. *A Treatise of Legal Philosophy and General Jurisprudence*, Volume 11, *Legal Philosophy in the Twentieth Century. The Common Law World*. Dordrecht: Springer: Chapter 7.
- Priel, D. 2013. "Law, Social Phenomenon of." In B. Kaldis (ed.). *Encyclopedia of Philosophy and the Social Sciences*. Los Angeles: Sage Publications: 543–46.
- Riley, P. 2009. *A Treatise of Legal Philosophy and General Jurisprudence*, Volume 10: *The Philosophers' Philosophy of Law from the Seventeenth Century to Our Days*. Dordrecht: Springer: Chapter 17.
- Roux, S. 2011. "Introduction." In K. Ierodiakonou and S. Roux (eds.). *Thought Experiments in Methodological and Historical Contexts*. Leiden: Brill: 1–33.

- Sartorius, R. 1966 [1971]. "Hart's Concept of Law." *Archiv für Rechts-und Sozialphilosophie* 52: 161–94. Reprinted in, and cited from, R. S. Summers (ed.). *More Essays in Legal Philosophy*. Oxford: Basil Blackwell: 131–61.
- Simpson, A. W. B. 2011. *Reflections on The Concept of Law*. Oxford: Oxford University Press.
- Stavropoulos, N. 2001. "Hart's Semantics." In Coleman 2001: 59–98.
- Stuart, M. T., Fehige, Y., and Brown, J. R. (eds.). 2018. *The Routledge Companion to Thought Experiments*. New York: Routledge.
- Stuart, M. T., Fehige, Y., and Brown, J. R. 2018. "Thought Experiments: State of the Art." In Stuart, Fehige and Brown 2018: 1–28.
- Szabó Gendler, T. 2000. *Thought Experiment. On the Powers and Limits of Imaginary Cases*. New York and London: Garland Publishing, Inc.
- von Daniels, D. 2016. *The Concept of Law from a Transnational Perspective*. London and New York: Routledge.

Intuiting Intuition: The Seeming Account of Moral Intuition

HOSSEIN DABBAGH*

*Doha Institute for Graduate Studies, Doha, Qatar
Institute for Cognitive Science Studies, Tehran, Iran*

In this paper, I introduce and elucidate what seems to me the best understanding of moral intuition with reference to the intellectual seeming account. First, I will explain Bengson's (and Bealer's) quasi-perceptualist account of philosophical intuition in terms of intellectual seeming. I then shift from philosophical intuition to moral intuition and will delineate Audi's doxastic account of moral intuition to argue that the intellectual seeming account of intuition is superior to the doxastic account of intuition. Next, I argue that we can apply our understanding of the intellectual seeming account of philosophical intuition to the moral intuition. To the extent that we can argue for the intellectual seeming account of philosophical intuition, we can have the intellectual seeming account of moral intuition.

Keywords: Philosophical intuition, moral intuition, intellectual seeming, Bealer, Bengson.

1. *Introduction*

Epistemological moral intuitionism is ordinarily thought of as an account of non-inferentially justified *moral* intuitions. In this paper, I deal with intuitionists' mental ontology. I defend the quasi-perceptualist account of *philosophical* intuition, which understands intuitions as *intellectual seemings*. According to this account, to have an intuition that *p* is to have the intellectual seeming that *p*. I will say more about intellectual seemings and certain shared phenomenological features between intuitions and perceptual experiences. In order to do so, I appeal to John Bengson's view about intuition. Following Bengson (2010),

* I would like to thank Philip Stratton-Lake, Brad Hooker, Sophie-Grace Chappell, David Oderberg, and the editor and anonymous reviewers, for their comprehensive critical comments.

I explain intellectual seemings in terms of “presentation” and “translucency”. Although Bengson echoes almost all that George Bealer (1998) believes, Bengson labels his account of intuition “Quasi-Perceptualism”. Bengson puts more weight on the shared phenomenological features between intuition and perceptual experience than Bealer did and Bengson claims that intuition is fundamentally just *like* perceptual experience but is still not *sensory* experience.

In the next section, I rely on Bengson’s view to outline a quasi-perceptualist account of philosophical intuition to explain intuition in terms of intellectual seemings. However, he recently argues in his paper, “The Intellectual Given” (2015), that his perceptualist account differs from a seeming account of intuitions, e.g. Bealer’s seeming account. For example, Bengson says that seemings are not non-voluntary, compare to presentations discussed in core quasi-perceptualist thesis. Or while a seeming is explicit, i.e. its content is available when the content seems true, presentations in core quasi-perceptualist thesis can be inexplicit. In this paper, however, I assume that what Bengson considers as core quasi-perceptualist thesis can be applicable to intellectual seemings. For the sake of argument, the distinction between core quasi-perceptualist thesis and seeming view is not at stake. I believe Bealer’s seeming account and Bengson’s quasi-perceptualist thesis can give us important features to explain how moral intuitions work in terms of seeming.

After having understood what philosophical intuition is in terms of seeming, I then shift from philosophical intuition to moral intuition. I will say about Audi’s doxastic account of moral intuition and alternatively explain moral intuitions in terms of quasi-perceptualist account which understands moral intuitions as intellectual seemings. In the meantime, I argue that the intellectual seeming account of intuition is superior to the doxastic account of intuition.

2. *Quasi-Perceptualist Account of Philosophical Intuition*¹

There are some similarities between intuition and perceptual experiences. By perceptual experiences, I assume the standard representational theory of perception. According to this view, to have a perception of an object *O* as having a property *F* is to be in a perceptual mental state with a phenomenal character which represents *O* as having the property *F*, i.e. it has representational content that *O* is *F*.

Perceptual experiences should be distinguished from *inference*. In making inferences, we often actively practice a number of steps of explicit reasoning, whereas in perceptual experiences something simply comes to us passively. Yet, we can use some inferences to explain why we have a particular perceptual experience. Thus, perceptual experiences, in this sense, give us a sense of directness, “givenness” and viv-

¹ In writing this section, I was influenced by the works of Dancy (2014) and Bengson’s doctoral thesis (2010), “The Intellectual Given”.

idness. Perceptual experiences are examples of non-doxastic states, so essentially can involve grounding non-inferential justification for our beliefs (see Chappell 2008).

Bengson is impressed by certain phenomenological features shared between intuition and perceptual experiences (see Bengson 2015). Of course, there are several differences between perceptual experience and intuition. For instance, intuition lacks the rich sensory phenomenology which most perceptual experiences have (see Williamson 2007: 217 and Sosa 2007: 48). Also, perceptual experiences are workable only in particular cases, while intuition deals both with the particular and the general cases (see Hintikka 1999: 137 ff.).

However, there are some abstract similarities between them that might be helpful in giving an account of the nature of intuition. For example, both perceptual experiences and *some* intuitions are direct, contentful and non-factive states. Suppose I have a sensory experience that there is a pen on the table in front of me. So, I am in a state with the direct content that there is a pen on the table. But, the experience might be non-veridical, i.e. not coincide with reality. Even more so, some of our intuitions, especially in moral cases, often fail to be correct. This must be true, since they so often contradict one another. And when one person's moral intuitions contradict another person's, at least one of these people must have incorrect moral intuitions.

What more can be said about abstract similarities between intuitions and perceptual experiences? If intuitions and perceptual experiences are, in a certain way, similar, what sort of mental state is intuition?

We can make a distinction among different contentful states in terms of representationality and presentationality. Some states such as beliefs, perceptual experiences and intuitions are representational in the sense that they represent the world in a certain way as if their content were true. For example, the belief that "Everest is the highest mountain in the world" *represents* the world in a certain way that its content is true. Or one's moral belief that p, e.g. "surrogate motherhood is wrong", *represents* the world as being such that p is true, i.e. it is not permissible to obtain or to be a surrogate mother. However, some states such as hopes, desires and wishes do not represent the world in a certain way as if their content were true, although they are contentful, since they do not aim to describe the world. For example, my hope that "World War III does not happen" does not represent the world in a way that its content describes the world. Spelling it out in terms of "direction of fit", we can say that beliefs aim to fit the world, but desires, hopes, intentions, and so on aim for the world to fit them (see Searle 1979).

There are also some contentful representational states that are also presentational, in the sense that not only do they represent the world in a certain way, but also they *present* the world in a certain way.² For

² We can also think of mere presentational states when we are in pain. Presentational states such as pain come to us non-voluntarily and without our conscious intention.

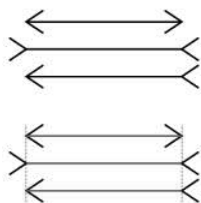
example, when I have a perceptual experience that there is a book in front of me, the world is represented to me in a certain way that it is true that there is a book in front of me. Furthermore, while I have this perceptual experience, it is presented to me (non-inferentially) that there is a book in front of me. In fact, I have the (non-inferred) impression or feeling that there is a book in front of me. Of course, we can have this (non-inferred) sense that there is a book in front of me even if it turns out that this is not so. For example, Jim Pryor writes about the presentationality of perceptual experience as

the peculiar “phenomenal force” or way our experiences have of presenting propositions to us. Our experience represents propositions in such a way that it “feels as if” we could tell that those propositions are true—and that we’re perceiving them to be true—just by virtue of having them so represented (Pryor 2000: 547).

William Tolhurst (1998: 298–299) also has the same idea in his mind when he writes about seeming states as “felt givenness”:

The real difference between seemings and other states that can incline one to believe their content is that seemings have the feel of truth, the feel of a state whose content reveals how things really are. Their felt givenness typically leads one to experience believing that things are as they seem as an objectively fitting or proper response to the seeming.

Consider now the famous picture of Mueller Lyer (below). Although the unequalness of the two lines is non-voluntarily *presented* to us, we still believe that they are equal as they *represent* to us in another way. In such cases where it is as if something has come to us, we are actually in a state that is presentational. This entails that unlike representational states, presentational states do not simply represent the world as being a certain way. Yet they present the world as being that way as if things are so.



Presentational states have at least three characteristics: they are gradable, non-voluntary and compelling (see Dancy 2014). They are gradable in the sense that their quality may vary from one situation to another situation, depending upon the way in which they are presenting. They are non-voluntary in the sense that, unlike decisions (which are active), presentational states are passive and happen to us (see Wittgenstein 1976: 632). They are compelling in the sense that it is hard to resist assenting to their contents when they are presented.

Having understood what the difference between presentationality and representationality is, we are able to agree that the presentation-

ality of perceptual experience is not a very challenging idea. We should accept that perceptual experiences are presentational states.

However, what about intuitions? Are they presentational states? There are some reasons, I believe, to think that intuition is a presentational state—a state that presents its content as being so. For instance, suppose that we have an intuition that it is not possible that both *p* and not-*p*. When we have this intuition, it is not simply to say that we are in a state that *represents* the world as if the principle of non-contradiction is true. We can have just the *sense* or *impression* that this principle is so. And just like perceptual error that an object *x* can present itself as a *y*, in case of intuitional error *p* can present itself as not-*p* or vice versa. For example, we might have an intuition that *p* because it seems to us that *p*. But after further reflection or getting confirmation from a third party, we find that we were wrong.

Thus, although intuition and perceptual experience are different and have different properties, they have some similarities and can be the same kind of state in terms of *presentations*. Following Dancy (2014) and Bengson (2010), I call this

*The First Quasi-Perceptualist Thesis: (i) Intuitions are **akin** to perceptual experiences in being **presentational**.*

Formulating intuitions in terms of presentationality has different virtues. First of all, this thesis simply makes a distinction between intuition and some other mental states such as guesses, hunches, hypotheses, conjectures or beliefs that are merely *representational*. Just as perceptual experiences are typically non-voluntary, intuition is a non-voluntary state and can oppose what we believe or are inclined to believe. Hence, insofar as intuitions are akin to perceptions in being presentations, they are *belief-independent*.

Moreover, by appealing to the first quasi-perceptualist thesis, we reveal another difference. We can make a distinction between intuition and *dispositions* or *inclinations* such as attractions and temptations. Intuitions are presentations, but inclinations are not. As happens in the case of wishful thinking, it is possible to have a feeling of being inclined to believe that *p*; however, *p* is not presented to one as true. Thus, the first quasi-perceptualist thesis identifies a difference between intuitions and other phenomena in terms of non-presentationality and presentationality.

As a second virtue, the first quasi-perceptualist thesis is able to provide us an account for *psychological* roles of intuition. In fact, the thesis *explains* how intuitions help us to come to believe something or form our beliefs. For example, in Jackson's thought experiment, we may form our *belief* that Mary does learn by having the *intuition* that she does learn. In this sense, intuition has the *explanatory* power with respect to beliefs, i.e. intuitions explain beliefs. For in different situations we can say "I believe that *p* on the basis of the intuition that *p*". Why do we believe that, for example, Mary does learn? Simply because

it *strikes* us that Mary does know or we have the intuition that Mary does learn. Therefore, the thesis may explain why we have the corresponding intuitive belief. That intuitions are presentations helps us to explain why intuitions are explanatory of belief.

Perceptual experiences also have another characteristic shared with intuition, namely, *translucency*. Bengson explains the idea of translucency in this way:

Let us call a presentational state σ of x *translucent* iff, in having σ , it is presented to x that p is so, and there is no content q (where $q \neq p$) such that it seems to x that p is presented as being so by q 's being presented as being so (2010: 38).

Yet, what does it mean when we say a mental state is translucent? According to Bengson, calling intuitions translucent is a way of saying that intuitions are direct (or non-inferred). However, there is a distinction in philosophy of perception between “translucent” and “transparent”. The distinction picks out as translucent a class of experiences that are not completely direct or non-inferred. Contrast this with transparent experiences, where this is not so. For example, when I look at a tree or when I introspect my visual experience, my experience is transparent to me (see Smith 2008). What Bengson must mean by translucent experiences is transparent, direct (or non-inferred) experiences.³ Let me explain.

Suppose one is sitting in front of a table and there is a pen on the table. One's vision of the pen directly presents the fact that there is a pen on the table. Contrast this with the situation that one suddenly notices that the pen's ink is empty by seeing that the pen does not work. It may be presented to such a person that the ink is empty even though she lacks perceptual experience of the ink (suppose the ink tank is covered up). That the pen is not working serves as her “perceptual guide” (see Bengson 2010). Most likely, in such a case, one infers that the ink is empty from the fact that the pen is not working. One thinks that the best explanation of the pen's not working is that the ink has run out, and so one makes the inference about the ink. This entails that the presentation of the pen as being out of ink is not direct (translucent).⁴

We can think of this distinction between direct or translucent presentation and indirect presentation, in the *intellectual* cases, as well. Consider the intuition that “identity is transitive”. This intuition is *translucent* in the sense that it is *presented* to one as being the case that identity is transitive. It does not present to one as being so by something else (other propositions) being presented so. It seems that one can “just see”, *directly*, that it is so. However, there are intuitions which do not have this *directness* or are not *translucent*, especially cases in which one may be presented with multiple contents, some of which hold in virtue

³ For the sake of argument, this distinction is not at stake here. I treat “translucent” as if it means “transparent” in this paper and use them interchangeably.

⁴ For another example, see Dretske (1969: 153 ff.).

of the others. In effect, translucency has two components: *presentation* and *directness*, which bring the *epistemic* status of being *un-inferred*.⁵ Note: that some presentation is translucent and thus un-inferred does not imply that its content *cannot* be inferred as well.

In the light of the discussion of translucency, we can now add another constituent to the quasi-perceptualist thesis. I call this

The Second Quasi-Perceptualist Thesis: (ii) Intuitions are intellectual translucent states.

But how is this “intellectual” state generated? Why do not we postulate intuitions as sensory or perceptual states? One might even object that intellectual states are *completely non-sensory* because they do not involve sense data. If this is the case, it seems that all we have said so far about the certain shared phenomenological features between intellectual seemings and perceptual experiences is redundant.

The answer is that, although intuitions are similar to perception in terms of translucency and presentation, intuitions *cannot* be just sensory perceptual states. We can also think of two negative and positive readings of an intellectual state: a negative reading of an intellectual state equates “intellectual” with *completely non-sensory*. However, a positive conception reads an intellectual state as a state that involves the *deployment* or *exercise* of *concepts*. The quasi-perceptualist does not need to choose between these two readings. Therefore, the quasi-perceptualist thesis is “neutral” on this issue (see Bengson 2010).⁶

Hence, combining the two constituents of the quasi-perceptualist thesis, i.e. (i) and (ii), yields the core idea of quasi-perceptualism about intuition. This can be formulated as

The Quasi-Perceptualist Theory of Intuition: Intuitions are translucent intellectual presentations.

Although quasi-perceptualism explains intuition with terminology different from that used by Bealer (2000), I think they are both saying the same thing. In other words, Bengson tries to elaborate what Bealer means when he uses intellectual seemings. And by seemings, in Bengson’s terminology, Bealer means something direct or translucent and presentational.

We should bear in mind that nothing we have said implies that intuitions must be *unreflective* or *gut feelings*.⁷ Rather, a translucent intellectual presentation with certain content may occur in the case of *substantial reflection*. But through this reflection, intuitions do not

⁵ I elsewhere argued for the epistemology of moral intuitionism on the basis of “non-inferred epistemological intuitionism”. See Dabbagh (2017).

⁶ I do not deny that there is a tradition of philosophers such as Kant, Sellars, and McDowell etc. who think that perceptual states involve the deployment of concepts. For example, when I see a tree in front of me I have deployed the concept of a tree. My claim here is compatible with what they said.

⁷ See Prinz (2006) for an alternative view.

need to make a *transition* from one proposition to the second one, because they are translucent.⁸

So far, I have given an explanation—and to some extent justification—of what a philosophical intuition is. It is now time to examine whether quasi-perceptualist account of philosophical intuition in terms of seeming is applicable to moral intuition. I argue we can have a seeming account of moral intuition as well.

3. *Shifting from Philosophical Intuition to Moral Intuition: The Seeming Account of Moral Intuition*

Having discussed what *philosophical* intuition is, we now direct our focus to what *moral* intuition is. Jennifer Nado (2012) distinguishes between *epistemological* intuition and *moral* intuition and argues that the mental states falling under the category of intuition are quite heterogeneous. In almost the same manner, I assume here that it is plausible to think of two separate types of intuition with different content as “philosophical intuition” and “moral intuition”. However, I do not believe that philosophical intuition and moral intuition are not two different types of mental state. For having different content does not make something a different mental state. The nature of moral intuitions and philosophical intuitions and how they work to justify our beliefs are the same. Thus, the characteristics that we attribute to philosophical intuitions can also be attributed to moral intuitions. Yet, we can make a distinction between philosophical and moral intuition, in terms of their different content, and this distinction between philosophical and moral intuition helps us to focus solely on moral intuition.

The term “ethical intuition” or “moral intuition” has often raised difficulties in the history of moral philosophy. Some moral philosophers think that the term “moral intuition” refers to a moral judgement shared by philosophers and scholars. Some of these philosophers think that moral intuition is just immediate or non-inferential moral judgements. Some others think of moral intuition as a pre-theoretical judgement. Another understanding refers to philosophers who think about moral intuitions as apparent and self-evident truths.⁹ For example, notably, Robert Audi describes moral intuition as a doxastic state about a self-evident proposition.¹⁰

Below, I will delineate Audi’s doxastic account of moral intuition to argue that the seeming account of intuition is better than the doxastic account of intuition. I will partly argue against Audi’s account of moral

⁸ This translucency is like non-inferentiality in the case of propositional belief.

⁹ I have used Lillehammer’s (2011) various “conceptions of ethical intuition” here.

¹⁰ This does not entail that, Audi believes, we cannot have intuitions about non-self-evident propositions. We can have intuitions that are not intuitions of self-evident propositions. See Audi (1996: 109–110).

intuition that the seeming account of moral intuition can do better a job than his doxastic account. I then discuss an alternative.

3.1. *Audi on the Nature of Moral Intuition*

According to Audi, a moral intuition should have at least four conditions (listed below). Although Audi talks about four conditions, it seems that the “pre-theoretical” condition entails the “directness” condition. For, in Audi’s view, if an intuition is not held or believed on the basis of a premise or theoretical hypothesis, it must be non-inferential.

(1) First, a moral intuition must be non-inferential (*directness requirement*). This means that “the intuited proposition in question is not—at the time it is intuitively held—believed on the basis of a premise” (2004: 33). (2) Second, moral intuitions must be firm cognitions (*firmness requirement*). This means that “intuitions are typically *beliefs*, including cases of knowing”; however, “[a] mere inclination to believe is not an intuition” (2004: 34). A moral intuition must have some degree of epistemic weight, i.e. conviction. (3) Third, a moral intuition must be shaped by an adequate understanding of its propositional object (*comprehension requirement*). An adequate understanding for a belief “tends both to produce cognitive firmness and to enhance evidential value” (2004: 34–35). (4) Fourth, moral intuitions must be pre-theoretical (*pre-theoretical requirement*). Moral intuitions are not like theoretical hypotheses, nor do they depend being inferred from theories. So, “... an intuition *as such*... is held neither on the basis of a premise nor as a theoretical hypothesis” (2004: 35).¹¹

Nevertheless, moral intuitions are defeasible. They can be defeated by some theoretical results that are incompatible with the moral intuition (see Audi 1996: 110). By accepting the defeasibility of moral intuitions, Audi tries to distinguish between *reliable* and *unreliable* moral intuitions through entering the notion of *justification*. According to him, reliable moral intuitions are those that “we can rationally hope will remain credible as we continue to reflect on them” (1996: 121). Of course, Audi does not suggest that what makes certain moral intuitions reliable is that we rationally hope they will remain credible as we reflect on them. What he must have meant is that we rationally hope that the moral intuitions that are reliable will remain credible as we reflect on them. For a moral intuition to remain credible as we reflect on it is for it to be “stable under reflection”. Reliability and stability under reflection are different things. What makes a moral intuition reliable, in Audi’s view, is that it normally or nearly always leads to the *truth*. In fact, some moral intuitions are reliable, as having initial credibility and as themselves being *prima facie* justified. Audi says that insofar as moral intuitions

¹¹ Audi elsewhere states that his focus on intuitions is on empirical quasi-perceptual intuitions: “My concern will be only empirical intuitions and mainly quasi-perceptual intuitive moral judgments” (2007: 201). But how are moral intuitions empirical ones? I am not sure what Audi means by this, especially when he thinks intuitions are identified with *a priori* ones!

are like certain perceptual beliefs (e.g. in being non-inferential, “natural,” and pre-theoretical)—and perhaps more important—insofar as they are based on an understanding of their propositional objects, there is reason to consider them *prima facie* justified (Audi 1996: 116).

So understood, in Audi’s view, moral intuition *simpliciter* can be reliable, has initial credibility, and can be considered as *prima facie* justified, but on one condition: moral intuitions must be formed in light of an adequate understanding. If they are not based on sufficient reflection, we lack reason to consider them *prima facie* justified.

Moreover, pre-theoreticality of moral intuitions does not imply that the propositional content of a moral intuition is not capable of proof and inferential justification. It is not true that a non-inferential cognition cannot be believed as a theoretical hypothesis (see Audi 2004: 35–36; 1996: 112 and 1998: 23).

In Audi’s view, to give a plausible account of moral intuition, we need the idea of “reflection”. In order to do that, he distinguishes between a *conclusion of inference* and *conclusion of reflection*. A conclusion of inference is “premised on propositions one has noted as evidence” (1998: 19). Simply put, a conclusion from one or more evidential premises is a conclusion of inference. In contrast, a conclusion of reflection “emerges from thinking ...but not from one or more evidential premises” (1998: 19). To give a better idea of what the conclusion of reflection is, Audi compares it to looking at a painting or seeing a facial expression. When a conclusion of reflection emerges, one can obtain a view of the whole and characterise it (see Audi 2004: 45–47). Moral intuitions, Audi holds, should be known as conclusions of reflection. The conception of moral intuition, then, is that moral intuition is a non-inferential cognitive capacity, not a non-reflective one (see Audi 1996: 112 and 1998: 20). However, as Audi rightly observes, this does not imply that “every intuitive moral judgment need be a conclusion of reflection” (2007: 204).

It is clear from Audi’s definition that moral intuitions have an epistemological feature as well as a normative one. A moral intuition is something that is totally dependent on the level of understanding of each person and is not necessarily obvious to all. Rather, it may be rejected or become clearer in the course of theorising. Also, as Audi puts it, moral intuitions must be understood here in a cognitive sense (see Audi 2004: 32). Moral intuitions have an epistemic role in our judgments and they have effects on our beliefs, i.e. they lead us to know some moral principles and believe in them.

Audi sees a sort of connection between moral intuition and self-evident propositions. He believes that moral intuitions are typically our beliefs about some self-evident moral principles, and there are some moral self-evident principles that we have moral intuitions (beliefs) about (see Audi, 1996; 1998; and 2004). For example, in Audi’s view, the moral intuition that “promise-keeping is permissible” is typically our belief about the self-evident principle that “promise-keeping is *pro*

tanto right”. Furthermore, we have a moral intuition about the self-evident proposition “promise-keeping is *pro tanto* right”, which is intuitively true. Of course, this does not entail that all moral intuitions are self-evident propositions.

Although adopting the doxastic account of intuition has different advantages, I believe, the seeming account is superior. The seeming account of moral intuition can help us to distinguish intuition from certain similar mental states, such as guesses, gut reactions, hunches and common-sense beliefs. The reason that I advocate the seeming account is that it looks more fundamental than the doxastic account. We can explain why we believe various things by saying that they *seem* true to us. In other words, even in cases where we believe something, we actually believe it because it *seems* true to us. Although seeming *p* true to me is a decent reason for my believing *p*, believing *p* is not an enough reason for me to believe *p*.

3.2. *Moral Intuitions as Seeming States: Can Bengson’s Account be Applied to the Moral Domain?*

Moral intuitionists like Michael Huemer understand intuition in terms of *seeming* states or as an “initial intellectual appearance” (2005: 101–105). Moral intuition, on the basis of this understanding, is an initial intellectual appearance with moral content.¹² In what follows, I will focus on the psychology of moral intuition generally and try to answer the question of “what are moral intuitions like” specifically.

Three main questions about moral intuitions can be distinguished, Sidgwick believes. One is a question about existence (psychology). The second is a question about validity (epistemology). The third is a question about origin (see Sidgwick 1967: Book 3, Ch. 1, at 211). The question about existence is a psychological question asking whether it is possible for people to ever have a moral intuition. The question about validity is an epistemological question seeking truth in such moral intuitions. The question about origin, finally, is a question of what the nature of moral intuition is and how moral intuition is developed.

Although an answer to the question about existence can be affected by what we think a moral intuition is, Sidgwick rightly thought that the question about existence and the question about moral intuition’s nature should be kept separate. By listing some states of mind that can be confused with intuitions, Sidgwick directs our attention to the question of what exactly intuition, and specifically moral intuition, is. He starts off by asking what the difference between moral intuition and blind impulses or vague preferences is. And he finally ends up talking about moral intuition as “judgment or apparent perception that an act is in itself right or good” (1967: Book 3, Ch. 1, §4). However, he cannot endorse that this is the definition of moral intuition. For there are some

¹² See also McMahan (2000: 93–4). For a discussion of two views on moral intuitions, see Bedke (2008).

examples that even Sidgwick knows of as fundamental intuitions, but they do not have such content! Consider his intuition that it cannot be right for person A to treat others in a certain way and not right for person B to treat others in that same way unless there is some relevant difference between A and B or their situations, beyond the bare fact that A is A and B is B. Another is his intuition that from the point of view of the universe no one person's good can matter any more or less than any other person's good, apart from their effects on others. Neither of these intuitions holds that an act is in itself right or good.¹³

But, if it is true that moral intuitions are not judgement or apparent perception that an act is in itself right or good, what phenomenon, event or state is moral intuition? As I discussed before, recent work in philosophical intuition by Bengson (and Bealer) understands philosophical intuitions as intellectual seemings. Intellectual seemings are similar to perceptual experiences, though important differences should be taken into account. For instance, perceptual experiences are conscious, contentful, non-factive and presentational. And since they are presentational, they differ from belief or judgement. In being presentational states, they are baseless, gradable, fundamentally non-voluntary and compelling. Bealer, whose works in this area are seminal, discusses philosophical intuitions only in four domains: the conceptual, the logical, the mathematical and the modal. Perhaps we do not need to determine whether Bealer's account of intuition is suitable for the four domains of philosophical intuitions he is interested in. Our question is whether his discussions are appropriate for intuitions in ethics. Bealer does not apply his account explicitly to ethics. Nevertheless, Bealer's four domains can be used as an argument for the view that moral intuitions can be treated as evidence, as I will explain below.

Bengson's work, which is a development of Bealer's ideas, also does not clearly discuss moral intuition's mental ontology. Although both Bealer and Bengson do not clearly apply their theory to ethics, I think Bealer's or Bengson's account of intuition can make space for the ontology of moral intuitions. In the next paragraphs, I run through Bealer's and Bengson's accounts of the ontology of philosophical intuitions to provide an account of moral intuition.

Let us start with the psychological question about moral intuition. Borrowing Bengson's account of philosophical intuition, we can ask whether we have conscious, contentful, non-factive and presentational states with moral content. Do we have such mental states in ethics? The answer to these questions, I think, is simple. It is clear for us that at some point we have conscious, contentful, non-factive and presentational states with moral content (e.g. promise breaking is wrong). These are states with moral content that fit the general account of the presentational state.

¹³ Sidgwick believes that at least some intuitions can occur, in principle, without being true. He, for example, admits "the possibility that any such "intuition" may turn out to have an element of error" (1967, Book 3, Ch. 1, §4).

Now if moral intuitions so defined exist, what is the nature of moral intuition? According to Bengson's view, intuitions are intellectual seemings, or to put it more accurately, intuitions are translucent presentational states. But are moral intuitions intellectual seemings?

There is a decisive reason for believing in seemings with moral content, I believe.¹⁴ We should admit that at least some moral intuitions are intellectual seemings. Consider, for example, the following propositions: "it seems that killing innocent people for no reason is absolutely wrong"; "it is wrong to torture someone for one's own amusement"; "that an act would hurt an innocent person must count morally against it"; "that an act would reduce the pain an innocent being is suffering counts morally in favour of it". Insofar as one adequately understands the conceptual constituents, one can be struck by the seeming rightness of these propositions. Consideration of these propositions produces intellectual seemings with moral content. In effect, what makes an intuition a moral one is an intellectual seeming with moral content.

Yet, one might object that what get produced are beliefs not intellectual seemings. The reply should be that these propositions match the contents of the intellectual seemings. For example, suppose we have a proposition (P) and suppose further that the proposition seems to us to be true. Then the proposition matches the content of our intellectual seeming.

Moral intuitions are similar to philosophical intuitions in that they are seeming states but with different contents. And in so being, one might believe that moral intuitions can present a consideration as evidence (which provides reason). And to present a consideration as evidence (which provides reason) is to present it as favouring a response of a certain sort.¹⁵ On this account, some moral intuitions or intellectual seemings present propositions as true (facts) and generate evidences for this or that sort of response. The seemingness of moral intuition, which is associated with some phenomenological features such as a feeling or appropriateness, can provide evidence for us.

How intuitions, generally, can be treated as evidence? Bealer (1992) famously writes about the three Cs to answer the question of whether intuition can be a source of evidence. Here are the three criteria: Consistency, Corroboration and Confirmation. The consistency test explores whether one intuition is consistent with other intuitions. The corroboration test asks whether one person's intuition is corroborated by others' intuitions. And the confirmation test seeks to establish whether those intuitions are confirmed by observation or experience.

But if the seemingness of moral intuition can provide evidence, a plausible objection could be raised. The objection is that there is no difference between moral intuitions and emotions in presenting evidence

¹⁴ Huemer (2007: 30–35), for instance, believes that any epistemological theory which denies the justificatory power of seemings with moral content is self-defeating.

¹⁵ See Dancy (2014).

(which provides reason), since at least certain emotions do this. So, if emotions seem to present the person who has them with evidence (which provides reason), are at least some moral emotions in fact moral intuitions? The answer is that moral intuitions are not emotions at all. This is because intuitions are purely intellectual seemings and hence truth-apt, yet emotions are not. Although further investigation is needed to have an account of emotion, my conjecture would be that some emotions can be like seeming states, and in being so they are similar to moral intuitions; and among those, the moral ones are similar to moral intuitions.

Understanding moral intuition with reference to seeming states, I think, can easily bring us at least some degree of justification. For example, if you have seemings about *p* and there is no defeater against *p*, you are to some degree justified in believing *p* based on that seeming.¹⁶ And if there is an explanation of how moral intuitions can serve as evidence in philosophy, this at least can give us a *prima facie* justification for using them.

4. Conclusion

I have investigated about the nature of moral intuition. I started with explaining the quasi-perceptualist account of philosophical intuition in terms of seeming. I focused on Bengson's (and Bealer's) intellectual seeming account and elaborated what the intellectual seeming is by appealing to Bengson's theory of quasi-perceptualism. Following Bengson, I considered presentational states as "immediate apprehensions" and allowed the notion of translucence to serve as an explanation of the notion of "directness". Consequently, I now have a conception of philosophical intuition as a kind of direct, immediate apprehension *akin* to perceptual experience, though it includes intellectual concepts. I also showed that intellectual seemings are translucent intellectual presentations. I then argued for reading moral intuitions in terms of Bengson's (and Bealer's) account of intellectual seemings. I showed that the quasi-perceptualist account of philosophical intuition which understands intuitions as seemings can be applicable to moral intuition. Therefore, we now can have a conception of moral intuition in terms of intellectual seemings similar to perceptual experiences.

References

- Audi, R. 1996. "Intuitionism, Pluralism, and the Foundation of Ethics." In W. Sinnott-Armstrong and M. Timmons (ed.). *Moral Knowledge?: New Readings in Moral Epistemology*. Oxford: Oxford University Press: 101–36.

¹⁶ Michael Huemer calls this "the principle of phenomenal conservatism". See Huemer (2007).

- _____. 1998. "Moderate Intuitionism and the Epistemology of Moral Judgment." *Ethical Theory and Moral Practice* 1: 15–44.
- _____. 2004. *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton: Princeton University Press.
- _____. 2007. "Intuition, Reflection and Justification." In M. Timmons, J. Greco, A. Mele (ed.). *Rationality and the Good*. Oxford: Oxford University Press: 201–221.
- Bealer, G. 1998. "Intuition and the autonomy of philosophy." In M. DePaul and W. Ramsey (ed.). *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry*. New York: Rowman and Littlefield Publishers, Inc.
- _____. 1992. "The incoherence of empiricism." *Aristotelian Society Supplementary Volume* 66 (1): 99–138.
- _____. 2000. "A theory of the a priori." *Pacific Philosophical Quarterly Journal* 81: 1–30.
- Bedke, M. 2008. "Ethical Intuitions: What They Are, What They Are Not, and How They Justify." *American Philosophical Quarterly* 45 (3): 253–270.
- Bengson, J. 2010. *The Intellectual Given*. Dissertation. University of Texas at Austin.
- _____. 2015. "The Intellectual Given". *Mind* 124 (495): 707–760.
- Chappell, T. 2008. "Moral perception." *Philosophy* 83 (4): 421–437.
- Dabbagh, H. 2017. "Sinnott–Armstrong Meets Modest Epistemological Intuitionism." *Philosophical Forum* 48 (2): 175–199.
- Dancy, J. 2014. "Intuition and Emotion." *Ethics* 124 (4): 787–812.
- Dretske, F. 1969. *Seeing and knowing*. Chicago: University of Chicago Press.
- Hintikka, J. 1999. "The emperor's new intuitions." *Journal of Philosophy* 96: 127–147.
- Huemer, M. 2005. *Ethical intuitionism*. New York: Palgrave MacMillan.
- _____. 2007. "Compassionate phenomenal conservatism". *Philosophy and Phenomenological Research* 74: 30–55.
- Lillehammer, H. 2011. "The Epistemology of Ethical Intuitions." *Philosophy* 86: 181–184.
- McMahan, J. 2000. "Moral Intuition." In H. LaFollette (ed.). *The Blackwell Guide to Ethical Theory*. Oxford: Blackwell.
- Nado, J. 2012. "Why Intuition?" *Philosophy and Phenomenological Research* 86 (1): 15–41.
- Prinz, J. 2006. "The Emotional Basis of Moral Judgment." *Philosophical Explorations* 9 (1): 29–43.
- Pryor, J. 2000. "The Skeptic and the Dogmatist." *Noûs* 34: 517–549.
- Searle, J. 1979. "A Taxonomy of Illocutionary Acts." In J. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press.
- Sidgwick, H. 1967 [1874]. *The Methods of Ethics*. 7th edition. London: Macmillan.
- Smith, A. D. 2008. "Translucent Experiences." *Philosophical Studies* 140 (2): 197–212.
- Sosa, E. 2007. "Experimental philosophy and philosophical intuition." *Philosophical Studies* 132: 99–107.

- Tolhurst, W. 1998. "Seemings." *American Philosophical Quarterly* 35 (3): 293–302.
- Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.
- Wittgenstein, L. 1976. "Cause and Effect; Intuitive Awareness." Translated by P. Winch. *Philosophia* 6 (3–4): 409–425.

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, Mišćević (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 *Nicomachean 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, Weatherston 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 epistemically 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 ‘ ϕ -ing is sometimes/often/always wrong’ fairly easy and natural to make, but what, if anything, warrants them? What are the epistemically 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 epistemically 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 suggest, 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 epistemically 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 intuition, 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 principle²⁰ 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 Mišević (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 Pizaro 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 under-described 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 station²⁷ (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 Mišćević (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. Sinnott-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. Sinnott-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 Mišćević*. 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.
- Mišćević, N. 2004. "The explainability of intuitions." *Dialectica* 58 (1): 43–70.
- Mišćević, N. 2007. "Modelling intuitions and thought-experiments." *Croatian Journal of Philosophy* 7: 181–214
- Mišćević, N. 2013a. "In search of the reason and the right: Rousseau's social contract as a thought experiment." *Acta Analytica* 28 (4): 509–526.
- Mišćević, 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.
- Mišćević, N. 2013c. "The ontology of secondary and tertiary qualities." *Balkan Journal of Philosophy* 5 (1): 45–58.
- Mišćević, 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. Sinnott-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.
- Weatherston, B. 2014. "Centrality and Marginalization." *Philosophical Studies* 171 (3): 517–533.

Simulation and Thought Experiments. The Example of Contractualism

NENAD MIŠČEVIĆ

University of Maribor, Maribor, Slovenia

Central European University, Budapest, Hungary

The paper investigates some mechanisms of thought-experimenting, and explores the role of perspective taking, in particular of mental simulation, in political thought-experiments, focusing for the most part on contractualist ones. It thus brings together two blossoming traditions: the study of perspective taking and methodology of thought-experiments. How do contractualist thought-experiments work? Our moderately inflationist mental modelling proposal is that they mobilize our imaginative capacity for perspective taking, most probably perspective taking through simulation. The framework suggests the answers to questions that are often raised for other kinds of thought-experiments as well, concerning their source of data, heuristic superiority to deduction, experiential, qualitative character and ease in eliminating alternatives. In the case of contractualist political thought-experiments, the data come from perspective taking and the capacity to simulate. Mental simulation is way more accessible to subjects than abstract political reasoning from principles and facts. There is a new experience for the subject, the one of simulating. Simulation normally is quick and effortless; the simulator does not go through alternatives, but is constrained in an unconscious way. We distinguish two kinds of political thought-experiments and two manners of imagining political arrangements, building third-person mental models, and first-person perspective taking. The two mechanisms, the first of inductive model building, the second for simulation, and their combination(s), exhaust the range of cognitive mechanism underlying political thought-experimenting.

Keywords: Thought experiment, simulation, social contract, veil of ignorance.

1. Introduction

A lot of thought experiments (TEs) requires the reader to take perspective on some morally, politically or legally relevant imagined situation; the Golden Rule TEs normally require one to take perspective on the victim's situation, the Veil-of-ignorance TE to take perspective on possible social arrangements under the supposition that one is ignorant of her own material situation, abilities and the like.¹ The topic of perspective-taking has become extremely popular in philosophy, psychology and related disciplines, in particular as far as its empathetic version is concerned.² In this paper I shall explore the role of perspective taking in political TEs (for short "PTEs"). What is the actual cognitive mechanism underlying the process? Here I shall opt for one particular, and rather popular view on perspective taking, namely that it crucially involves mental simulation (see Goldman 2006). Goldman has, of course noticed, the connection to various TEs, in particular to Golden Rule and the Veil-of-ignorance ones (2006: 294), but has not been developing it much. So, the goal here is accounting for cognitive mechanism underlying political thought-experiments (PTEs), more narrowly upon the presently most popular variant, namely the contractualist ones, in the widest sense of the term, with authors like Rawls, Scanlon, Habermas and Parfit (see References) at the forefront. These experiments typically address any given issue about the moral and political status of some arrangement (say, the status of the right to privacy) by inviting the reader to imagine a situation in which she is enabled to choose in the favor of it or against it, in her own name, and/or in the name of other people, under specified circumstances. She might be asked to imagine having to persuade other people to accept her choice, and reflect about ways of doing it, and so on. At the end of the experiment, the reader is supposed to have arrived at intuition(s) concerning the issue, for instance that she would choose the arrangement under such-and-such circumstances (say, under the Veil-of-ignorance), or that most people could not be persuaded to accept it, again under specified circumstances. These intuitions are not themselves normative, they are factual intuitions about possible choices. However, they serve as the basis for further theory-building, which then results in normative conclusions, usually of moral-cum-political character. The question this paper is addressing is simple to state: *Where do the intuitions come from? What is the possible psychological mechanism that produces the factual intuitions that serve as the basis for normative theory?*

The framework for the answer shall be my moderately optimistic, "deflationist" as David Davies (2018) calls it, mental modelling

¹ The paper originated from a presentation at a conference in Geneva 8–9 June 2017, on "Simulation and thought experiment". I would like to thank prof. Marcel Weber for inviting me, and the participants for interesting and helpful discussion.

² See chapters in Coplan and Goldie (2011) and Maibom (2017).

approach.³ I agree with him about the characterization, and I thank him. A variant of it has been developed in detail, namely the one that concerns building a mental model from the third-person perspective. I hope it can account for PTEs like Plato's *Republic*, where a group of young elite Athenians is supposed to imagine what life would be like for all sorts of people in a philosophers ruled state (Mišćević 2012a) In general, it is suitable for imagining political arrangements, primarily from the third-person perspective. However, it does leave open the accounting for a different kind of modelling, in which the experimenter is imagining social-political situations and arrangements, primarily from the first—person perspective—the social contract (SC) tradition and its present-day form, with star author like Rawls, Scanlon, and Habermas. For this kind of thought experiments I want to propose a solution within the general framework of mental modelling, but stressing a different kind of it: not building a model from the third person perspective, but trying to imagine how things would look to oneself, from the first person perspective. I will opt for one theory of such enactive imagining, namely the idea that we simulate perspective taking.⁴ Here is then the preview.

Section 2.1 summarizes the main idea of the SC tradition, also mentioning a simple forerunner of SC idea, namely the Golden Rule proposal. It proposes a division of SC theories, contrasting first the hypothetical ones, and the non-hypothetical (or partly non-hypothetical ones), and then, within the first group, those that rely on the picture of real, “normal” contractors, and those that propose idealization or other kinds of “retouch” of the parties participating. The SC PTE is built around the question for the would-be participants: what kind of arrangement would you accept, find just and liveable? The subject is supposed to arrive at an intuitional answer to the question.

Section 2.2 is the central part of the paper, dedicated to accounting for PTEs, in particular for the epistemic-psychological side and the question of how the relevant intuition gets formed. The first, very brief, subsection concerns the structure of a TE, and the second one turns to the role of simulation, that will be presented as the royal road to intuition. After a general brief mention of theories of simulation, it turns to its role in practical TEs. This is the central sub-section and the most important part of the paper, stressing the central role of simulation and showing how it fits well with independently established requirements of contractualist PTEs.

Section 2.3 mentions some difficulties with simulation that have been pointed to in the literature and offers some optimistic answer to them. In Conclusion we briefly sketch the bigger picture, namely our

³ I started defending this approach quarter a century ago, or, to put it more accurately a variant of it, in Mišćević (1992).

⁴ Of course, some kind of first person imagining might have been required in the *Republic* scenario: how would you feel if you had to lead the polis, and so on, but it is not central, as it will become in the SC tradition.

proposal for accounting for cognitive mechanism underlying PTEs in general, hoping that it might help understanding thought-experimenting in general and thus throw an additional light upon the foundations of methodology of philosophy.

Let me in the rest of this section introduce the kind of PTEs we shall be dealing with, namely the works in Social contract (SC) tradition that invite the reader to imagine social-political situations and arrangements that can come from the willingness of parties to come together and negotiate the best possibility. The first modern authors, Hobbes and Locke, are not completely clear about the factual status of the Contract, whether it is a historical event or merely imagined one; with Rousseau and most explicitly Kant, it becomes “hypothetical” contract, fit to be classified as a TE. At least since Rousseau it is discussed primarily from the first-person perspective; the typical leading question is *Would you sign a contract ...under such-and-such conditions?* We shall here set on one side contractarian version (due to authors like Hobbes and Gauthier (1986)) focused on maximization of the self-interest of each participant, since it poses less challenging problems to participant’s imagination, and focus on the more challenging contractualist line with authors like Rawls, Scanlon, Habermas and Parfit (see References). The typical demand here is to put oneself in another’s shoes while asking yourself: what demands cannot be rejected by my interlocutor, if she is rational?

We can describe the crucial imaginative exercises as moral and political TEs from the first-person perspective. This makes the SC tradition contrast with another famous tradition of thought-experimenting, starting at least with Plato’s *Republic*: building and understanding a complex social arrangement primarily from the third-person perspective (tell me, Glaucon, how would you judge the commonality of children? is it just or not? and so on...). It continues by building a mega-arrangement, and in more practically oriented cases ends as a utopia or dystopia, so to speak, with famous authors like Al Farabi, Thomas More and Fourier. Let me mention a contemporary proposal from the third-person perspective, a simple and fine example: the camping trip and equality among campers in G. A. Cohen *Why Not Socialism*. On camping people exercise solidarity, treat each other as equal, help altruistically and without reservations, so, Cohen concludes, we can use it as a model for a socialist society Cohen uses the understanding of equality provided by the trip-model to argue for extremely high level of equality in his socialistic society. (Some cases that are difficult to classify, say, prominently Dworkin’s anti-luck TEs).

Back to contractualism. Let me quickly propose a systematization of the main philosophical proposals within the hypothetical contract views since they are most relevant for discussion of thought-experimental methodologies, all this with apologies for brevity. One line does not propose, at least explicitly, any retouch of ordinary circumstances: the participants are real people, endowed just with ordinary rationality. Kant

and Parfit (in *On What Matters*) are prime examples of such approach.

With Rawls, a different line of experimenting started. The real subjects are replaced or supplemented by somewhat “retouched” model participants; in Rawls’ work the “parties” in the Original position, are famously placed behind the Veil-of-ignorance, and they just “represent” the real persons who make their contract on the basis of principles figured out by the “parties”.⁵ In the Original position the person decides to try the Veil-of-ignorance; she attempts to answer the crucial question: what arrangement would you choose if you were ignorant of some important aspects of your future situation? You ask yourself: shall I be male? Or female? And what is the best decision to take if I don’t know the answer? Shall I be intelligent? Or stupid? And so on.

Her counterpart, the “party” behind the Veil has to do the job:

The idea here is simply to make vivid to ourselves the restrictions that it seems reasonable to impose on arguments for principles of justice, and therefore on these principles themselves. Thus it seems reasonable and generally acceptable that no one should be advantaged or disadvantaged by natural fortune or social circumstances in the choice of principles. It also seems widely agreed that it should be impossible to tailor principles to the circumstances of one’s own case. We should insure further that particular inclinations and aspirations, and persons’ conceptions of their good do not affect the principles adopted. (Rawls 1999: 16)

The typical questions concern wealth, status, talents and the like. How would you decide if you knew you will be poor? Or, deprived of interesting and important talents? The person’s identity is preserved, and she simulates her reaction in a different situation than her actual one. The reason why in the Original position she has to deprive the participants of concrete knowledge of their actual standing in various relations in society is that participants have working models of social interaction. Therefore, if the person is choosing rationally, she will be partial to his (actual and future) self, and the promise of justice will be gone. Now, behind the Veil the participant does not know how rich she will be. She has to imagine herself being very rich (wow!), being moderately well off (not bad!) and being very poor (God forbid!).

Since parties have rich general information, she uses her default knowledge of how it feels being rich, well off and poor. She does not proceed to building a further model from the third-person perspective, but reasoning from the first-person perspective: let me imagine myself being poor, etc.! It is here that we shall introduce the idea of simulation. And the imagining will result in producing an answer, a particu-

⁵ Rawls in his *The Basic Liberties and Their Priority, The Tanner Lectures on Human Values April 10, 1981* stresses the following advice: “Two different parts of the original position must be carefully distinguished. These parts correspond to the two powers of moral personality, or to what I have called the capacity to be reasonable and the capacity to be rational. While the original position as a whole represents both moral powers, and therefore represents the full conception of the person, the parties as rationally autonomous representatives of persons in society represent only the Rational (...).” In McMurrin (1986: 19).

lar intuition: I should secure myself against the risks of ending up in a very bad situation.

Let me just mention the other famous retouch options. The main alternative to ignorance is idealization: how would my interlocutor react if she were made a bit more rational, and the discussion and decision situation were closer to an ideal one? Again, we are invited to put ourselves in another's shoes, this time in the shoes of a richly rational person, in the sense of rationality that also includes moral sensibility (in contrast to the means-end rationality of parties behind Rawls' Veil. What demands cannot be rejected by my interlocutor, if she is rational, Thomas Scanlon is asking:

My view ... holds that thinking about right and wrong is, at the most basic level, thinking about what could be justified to others on grounds that they, if appropriately motivated, could not reasonably reject. On this view the idea of justifiability to others is taken to be basic in two ways. First, it is by thinking about what could be justified to others on grounds that they could not reasonably reject that we determine the shape of more specific moral notions such as murder or betrayal. Second, the idea that we have reason to avoid actions that could not be justified in this way accounts for the distinctive normative force of moral wrongness. (Scanlon 1998 :5)

The procedure is presented as valid for political, institutional arrangements as well as for individual morality. Scanlon talks about “standards that institutions must meet if they are to be justifiable to those to whom they claim to apply” (Scanlon 2016: 5). So, suppose I want to propose a practice or institution *P* that is to apply to you. I have to get into your shoes: what kind of arguments would rationally persuade you to accept *P*? And a particular intuitional answer follows.

Finally, let me mention Habermas, who explicitly talks about his proposal as a TE (1989), and proposes to introduce idealized communicative situation as a whole, not just idealizations concerning the participants. Here is a brief characterization from the chapter “Remarks on Discourse Ethics”:

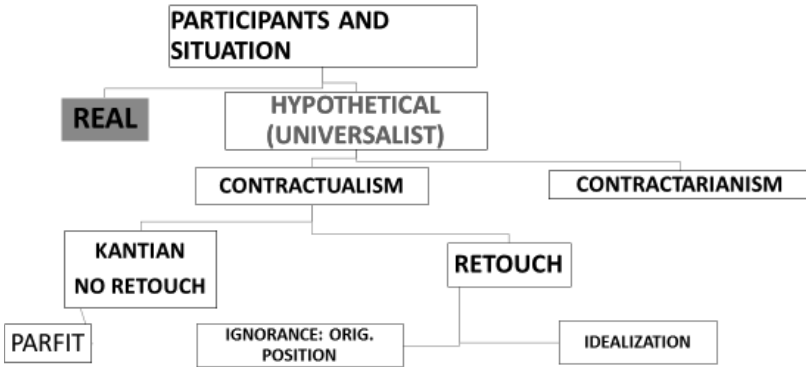
The notion that ideal role taking—that is, checking and reciprocally reversing interpretive perspectives under the general communicative presuppositions of the practice of argumentation—becomes both possible and necessary loses its strangeness when we reflect that the principle of universalization merely makes explicit what it means for a norm to be able to claim validity. Already in Kant the moral principle is designed to explicate the meaning of the validity of norms; it expresses, with specific reference to normative propositions, the *general* intuition that true or correct statements are not valid just for you or me alone. Valid statements must admit of justification by appeal to reasons that could convince anyone irrespective of time or place. In raising claims to validity, speakers and hearers transcend the provincial standards of a merely particular community of interpreters and their spatiotemporally localized communicative practice. (Habermas 1994: 52).

We are invited to see idealizations as those simultaneously unavoidable and trivial accomplishments that sustain communicative action and argumentation (Habermas 1994: 55). Commonsensical moves, like

attributing identical meanings to expressions, attaching “context-transcending significance to validity claims”, and ascription of rationality and accountability to speakers are pragmatic presupposition of communication that involve some idealization. The philosophical idealizations just continue where the ordinary ones stop.

In short, we thus have “retouched” (distorted) model participants and situations, changed basically in two directions. First, ignorance: what arrangement would you choose if you were ignorant of some important aspects of your future situation? Second, idealization what demands cannot be rejected by my interlocutor, if she is rational? And if we are placed in an ideal communicative situation? And here is the scheme of the division:

VARIETIES OF CONTRACT



Of course, the proponents of the ignorance strategy, Rawls and his many followers, have been confronting the defenders of idealization, like Habermas and Scanlon with their followers, and vice versa, and the debate has reached epic proportions. However, we have to leave it for another occasion, and pass to our main topic, the mechanism that produces the answers-intuitions.

2. Accounting for PTEs: The Role of Simulation

2.1 The task ahead

How do people understand imaginative scenarios essential for PTEs? Not much has been written about mechanism underlying PTEs. We need a more detailed look at TEs in general, and PTEs in particular.

Take first the simplest example, the Golden rule. Suppose I am bragging around with my knowledge of some area, and letting my colleague know how ignorant and incapable they are, when it comes to important issues. My wife asks: “Well, how would you feel if somebody were doing this to you?” I am supposed to imagine the reversed situation, go through the process of being humiliated, and feel what people normally

feel in such situations. This should make me sensitive to what I am doing. Here is Parfit's description of the typical process:

When we apply the Golden Rule, our thought-experiment is fairly simple. As when making many ordinary decisions, we ask what would happen in the actual world if we acted, on one occasion, in each of certain possible ways.

We don't even need to decide what are the morally relevant descriptions of these particular possible acts. But we try to think about these possibilities, not only from our own point of view, but also from the points of view of all of the other people whom our act might affect. We ask what we would rationally be willing to do, and have done to us, if we were going to be in all of these people's positions, and would be relevantly like them. (Parfit 2011: 328)

Let me propose a picture of the process of reasoning in a TE. We have two persons, the experimenter and the subject. At the preliminary stage, call it *stage 0*, the experimenter formulation her design: in our example, show to me that I should not humiliate my colleagues, and she wants to do this by asking me to imagine switching the role with a colleague, call him Jack.

At *stage one*, comes the presentation of the scenario thus constructed to the experimental subject, in this example to me: imagine you are Jack, the person that you have been humiliating!

At *stage two*, I, the experimental subject, come to understand the question. For instance, how would you feel if somebody were doing this to you?

At *stage three* comes the tentative production, "modeling" of the scenario at the conscious level. I imagine being humiliated. Then some unconscious processing might get in. The stage concerns the production of the answer, involving the generation of intuition; for instance, how I would feel in the shoes of the victim. This probably involves reasoning at the unconscious level; for instance, I might have to control my arrogance, and belief that yes, my colleagues are not as good as I am, and the like. This might result in an immediate, unconscious intuition, e.g. Yes, I would feel terribly...

At the next, *fourth stage*, the thinker comes out with explicit intuition at the conscious level, usually geared to the particular example and having little generality (again, I would feel terribly, etc.). This ends the core TE

Usually however, there is a *fifth* stage. The thinker often has to do some varying and generalizing, at the conscious and reflective level and, perhaps, at the unconscious one too. For instance, in the story I might be unimpressed by threat concerning my professional abilities. Imagine then, my wife might say, your young colleagues making deprecatory remarks about your age, suggesting it's time for you to retire, and let more energetic, younger people occupy the stage. And imagine that this is done by a younger colleague, what if it is done by a brilliant doctorate student, of someone else, over whom I have no power? Sometimes this process of going through related micro-TEs is called intui-

tive induction (Chisholm 1966). I end up with a general belief that my behavior is morally not acceptable. No matter what, such *kind of treatment* is awful, I would feel this for sure if someone did treat me thus.

If I am reflective enough, I might go one step further, to stage *six*. I first, consciously perform the aggregation of micro-TEs; second, I try to harmonize the results of these micro-TEs with each other, and finally, I arrive at a judgment of their coherence with other moral intuitions one might. In other words, this philosophical unification can be described in terms of reflective equilibrium, first narrow and then wide. Here, the general knowledge of more empirical kind is brought into play. I arrive at important and difficult task of comparing the result with all we know about life and politics, both personal experiential level and from history, social and natural sciences, reaching a wide reflective equilibrium as the final result.

A similar, but more demanding process goes on in the case of a contractualist TE. Take the Veil-of-ignorance situation and assume you are a male. Now you are asked to ponder the following Rawlsian question: what distributive arrangement would I choose if I didn't know whether I will be rich or poor? I basically go through same or analogous stages, and reach the final (non-moral) intuition, say I don't want to risk extreme forms of poverty, I want a decent life even if I am not rich. (Habermas similarly talks about "interlocking of perspectives", where everyone is required to take the perspective of "everyone else" (1995: 117)).

Here, we shall be mostly interested in stages three and four where this is supposed to occur. How does the thinker model the situation proposed in the scenario, and how is the resulting intuition produced? For this, we turn to cognitive investigations.

2.2. *Simulation, the royal road to intuition*

We have implicitly pointed to a promising answer: the thinker arrives to her intuition through mental modelling. I have been defending the role of mental modelling for more than two and a half decades (see Mišević 1992). David Davies mentions that I set "out clearly (1992: 24) how this approach solves the usual puzzles about TEs" (2018: 520). TEs enable us to produce new data by manipulating old data through the generation of a manipulable representation of a problem. In constructing and manipulating this model, we mobilize prior cognitive resources in new ways.

I would go further and claim that what cognitive science tells us about perspective taking, and more particularly about simulation, offers an interesting variety of this answer to our question. The idea that simulation produces the relevant intuition suggests the role of competences in TEs. I have been conjecturing that some of them might be quite general (the capacity to simulate other person's mental states which will occupy us in the sequel), some less general (folk-physics),

and some completely specialized (spatial, linguistic and mathematical skills). This suggests that the typical verdicts from TEs are in a way voice of competence, albeit a discreet one, often mixed with those from other sources (general intelligence, social skills, emotional life) (see Mišćević 2006, 2012). Here I want to introduce some new proposals, focusing on our capacity for perspective taking.

Let me distinguish two kinds of mental modelling that often get confused in the literature. One is the *third-person model-building*, for instance imagining a planet when reading a science-fiction story or Putnam's Twin earth description. The planet is an object that is imagined from a third-person perspective, relatively static, although allowing for some imagined movement. The other is the kind that will interest us here: *the first-person process of modelling, typically through mental simulation*, in which the subject imagines herself as the protagonist.

Let me first say a few words about the third-person model-building. Mental models, psychologists tell us, purport to represent concrete situations, with determinate objects and relations (precisely what is demanded in thought experiments) (Johnson-Laird 1983: 157). Their structure is not arbitrary, but "plays a direct representational role since it is analogous to the corresponding state of affairs in the world" (Johnson-Laird 1983: 157). One can distinguish simple static "frames" representing relations between a set of objects, crucially of human beings in the PTEs, temporal models consisting of sequences of such frames, kinematic models which is the temporal model with continuous time, and finally dynamic models which model causal relations. Reasoning in mental models demands rules for manipulation. Johnson-Laird hypothesizes the existence of general procedures which add new elements to the model, and 'a procedure that integrates two or more hitherto separated models if an assertion interrelates entities in them'. The integration of models is subject to consistency requirements—if the joint model is logically impossible, some change has to be made (Mišćević 1992: 220).⁶

Can this model-building from the third-person perspective help us with examples like Golden Rule and Veil-of-ignorance? Doesn't look very promising. The sinner in the Golden Rule experimenting is not supposed to imagine a neutral, distanced situation; she has to imagine *herself* in the reversed situation. Similarly, the thinker behind the Veil asks herself how *she herself* would choose, from the first-person perspective.

Here, a plausible alternative mechanism that would enable modelling from the first-person perspective is a mechanism of perspective taking. Psychologists talk about various ways of simulating and have interesting things to say about this. Consider the famous psychologists C. Daniel Batson. In his (2009) paper he writes:

Encounter a stranger in need and, sometimes, you will feel empathic con-

⁶ Other authors in the similar vein are Zwaan and Radvansky (1998), Hohwy (2013), and Frith (2007).

cern—an other-oriented emotional response evoked by and congruent with the perceived welfare of that person. What determines whether you will? Perhaps the most common answer among psychologists is that empathic concern is felt when you adopt the perspective of the person in need (...). But in this answer, what is meant by adopting the person’s perspective? First, it is an act of imagination. One does not literally take another person’s place or look through his or her eyes. One imagines how things look from the other’s point of view. Second, it is not the same as perspective taking in the symbolic-interaction tradition (...). In that tradition, one adopts the perspective of another—often a significant other—to imaginatively see oneself through the other’s eyes (and values). The perspective taking that evokes empathic concern involves imaginatively perceiving the other’s situation, not oneself. (Batson 2009: 267)

Philosophers and psychologists talk a lot about empathy, as a paradigmatic kind of perspective taking. However, there are several terminological problems connected with the term “empathy”; first, it often carries connotations of “sympathy”, so that empathetic understanding is the one that is accompanied by sympathetic feelings.⁷ Famous authors routinely connect the two (de Waal 2009, Clohesy 2013).⁸

Since we find mental Simulation Theory the best account of perspective taking, we shall turn to it and talk about “mental simulation” as the relevant activity. We hope, however, that what we have to say holds for perspective taking in general, and, if Simulation Theory turns out to be defective, can be connected to whatever account of perspective taking replaces it. We shall thus speak of perspective taking, and in particular of mental simulation as the second kind of modelling, besides the third-person one, that we need in order to answer the question of how we arrive at our responses and other people’s ones in cases like Golden Rule or the Veil. Indeed, some authors relying on cognitive science count ability to simulate as a part of general modelling ability. Thomas Metzinger mentions important traits of mental models, like being multimodal, mutually embeddable, often analog rather than digital, and then adds *ability to simulate (independently from input)* (2003: 109ff.)

So, let us turn to simulation. We are mental simulators, not in the sense that we merely simulate mentation, but in the sense that we understand others by using our own mentation in a process of simulation,

⁷ For relevant warnings see Amy Coplan (2011: 3). We shall heed the warnings and avoid unqualified use of the term.

⁸ Here is a statement by psychologist Chris Frith: “One obvious question is why have we put together empathy and fairness? In neuroscience there is not much overlap in the literature on these topics. Fairness tends to be studied within the realm of neuroeconomics, whereas empathy springs from the burgeoning studies that followed the discovery of mirror neurons. However, the two concepts are linked when we think of a possible basis for morality. We don’t like to be treated unfairly ourselves and we empathise with others who are treated unfairly. We will act to correct unfairness and to prevent it recurring” (Firth 2007: 1).

wrote Martin Davies (1994), and many colleagues, philosophers and cognitive scientists agree.⁹

We first have to clear a terminological mess. Some psychologists use the term “simulation” for any kind of imaginative enacting, so it ends up as meaning: model-building and model-activating:

The model can depict the system at some point of abstraction (...) A simulation is an applied methodology that can describe the behavior of that system using either a mathematical or a symbolic model (Sokolowski and Banks 2009: 5)

We shall use “simulation” in a narrower sense: the modelling through simulation does not primarily result in an object-depiction, but is primarily a first-person guided *process*, from which the subject can learn relevant first-person counterfactual matters (e.g. what would I do if I had to determine the price of my used car). Simulation thus involves the imagining subject (or his/her counterpart) as a part of scenario imagined. Remember: we understand others by using our own mentation in a process of simulation (Martin Davies). I shall be relying on a work already mentioned in the Introduction that is a synthesis of psychological and philosophical research on simulation (Goldman 2006). Goldman points out the existence of an alternative view of psychological understanding, namely Theory-Theory that postulates the existence of a cognitive module containing assumptions about “other minds” and ways they work.¹⁰ He allows for combination of the two (ST stands for “Simulation Theory”):

I shall call the act of assigning a state of one’s own to someone else projection. As we have just seen, projection is a standard part of the ST story of mindreading. It is the final stage of each mindreading act, a stage that involves no (further) simulation or pretense. Indeed, it typically involves an “exit” from the simulation mode that occupies the first stage of a two-stage routine. The simulation stage is followed by a projection stage. Thus, a more complete label for the so-called simulation routine might be “simulation-plus-projection”. (Goldman 2006: 40)

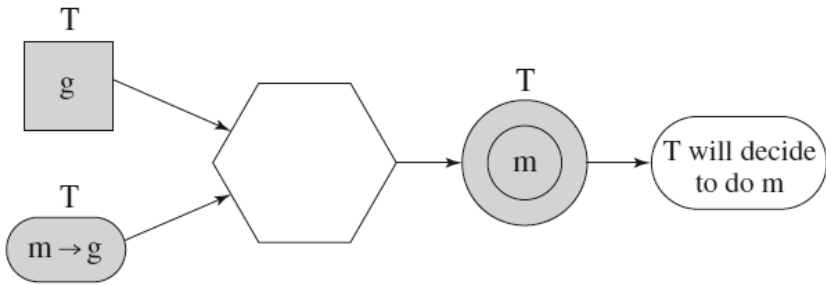
Similarly, Perner stresses simulation but allows for Theory-Theory episodes Perner and Kühberger (2005). I would also advocate a hybrid: a blend of Simulation Theory and Theory-Theory, with emphasis on simulation. I shall also borrow a term from Goldman, “enactive imagining”. He notes the following: “When I imagine feeling elated, I do not merely suppose that I am elated; rather, I enact, or try to enact, elation itself. Thus, we might call this type of imagination ‘enactment imagination’” (Goldman 2006: 47). He distinguishes more primitive type of simulation, mainly unconscious and related to mirror neurons, and more sophisticated, higher kind exemplified by enactive imagining, that is

⁹ See, for example Currie (2002) and the now classical text Gordon (1986). For early debate between the two kinds of theories and for important original contributions to it see *The mental simulation debate*, in Peacocke (ed.) (1994: 104).

¹⁰ For early debate between the two kinds of theories and for important original contributions to it see *The mental simulation debate* in Peacocke (ed.) (1994: 104).

relevant to us here. The latter is characterized by its target, namely “mental states of a relatively complex nature, such as “propositional attitudes”, by being partly “subject to voluntary control” and being to a high degree accessible to consciousness (Goldman 2006: 147). He proposes a nice flow chart to illustrate the simulation. Let me illustrate it with the famous Fat Man TE. I am supposed to decide whether I would push the Fat Man from the bridge in order to save five other people. Call me T I simulate my doing so; the final stage of the process looks somewhat like the following.:

desire



belief decision mechanism decision belief

Let **g** stand for “I don’t want to feel guilty”, **m** for “I am not pushing the Fat Man”. Then (**m**→**g**) says that if I don’t push the Fat Man, I shall not feel guilty. The decision is not to push, and the belief is my belief about myself, namely that I will not do it.

We shall use the flow chart in the sequel for most important PTEs to be discussed.

A distinction often drawn in the context of Simulation Theory is the one drawn by Robert Gordon in his (1995) and discussed by Gregory Currie in his (2002: 56ff). It concerns the contrast between two projects; in the first I “imagine is myself in your situation”, in the second I imagine being you in that situation. A philosopher is immediately reminded of a puzzle famously raised by Bernard Williams in his paper “Imagination and the Self” (1976):

It seems unproblematic for me to imagine that I am Napoleon; asked to do this, I know roughly how to comply. (Contrast this with the instruction to imagine that someone else—Abraham Lincoln, say—is Napoleon; here it is much less clear how to proceed.) But if imagining is a guide to possibility, my imagining may lead me to a further, more metaphysical thought: the thought that I could have been Napoleon. And it is this that Williams finds puzzling: “I do not understand, and could not possibly understand, what it would be for me to have been Napoleon” (Williams 1976: 45).

Indeed, how could I (or Williams or anyone other than Napoleon) have been Napoleon? Surely only Napoleon could have been Napoleon. The answer would demand a paper of its own.¹¹

¹¹ But see Vendler (1984) and Ninan (2016) for relevant discussion.

Can we trust Simulation Theory? How certain is it that people use simulation to understand each other? Here is a cautious formulation in a recent overview, “Folk Psychology as Mental Simulation”, offered by Gordon himself, together with his collaborator Luca Barlassina, in *Stanford encyclopedia*:

In particular, while the consensus view is now that both mental simulation and theorizing play important role in mindreading, the currently available evidence falls short of establishing what their respective roles are. In other words, it is likely that we shall end up adopting a hybrid model of mindreading that combines ST and TT, but, at the present stage, it is very difficult to predict what this hybrid model will look like. Hopefully, the joint work of philosophers and cognitive scientists will help to settle the matter. (Gordon and Barlassina 2017, ST stands for Simulation-theory, and TT for Theory-Theory)

The consensus “that both mental simulation and theorizing play important role in mindreading” is enough for our purposes. We shall thus assume that the Simulation Theory is the correct theory about *a* way, possible *the most important way*, in which a human being comes to find out and understand the thoughts of her conspecifics. (This allows for other ways, like the ones proposed by Theory-Theory or special module theory.) So, we shall assume that our thought-experimenter simulates the possible states, including feelings of oneself and others, and derives her judgments from the simulation. (The presence of additional, say Theory-of-mind elements would not change the basic situation, as long as simulation does play an important role). However, I hope that most of conclusions of this paper are valid for perspective-taking and imaginative enacting in general, independently of a particular mechanism in charge of it.

2.3. *Simulation in TEs, moral, political and legal*

It is now time to bring together the issues of perspective taking and our main topic, moral, political and legal TEs. Some TEs obviously involve empathetic perspective taking that ends in sympathy with the characters involved. The Trolley and Fat man TEs are a clear example, where the experimental results show a direct and strong involvement of subjects who have to imagine, presumably enactively to push by their own hands the Fat man, and kill him in this way. We normally have no problem in simulating to some degree the pain of other. Here is what neuroscientists tell us.

Seeing or imagining others in pain may activate both the sensory and affective components of the neural network (pain matrix) that is activated during the personal experience of pain. (Minio-Paluello, Avenanti, and Aglioti 2006: 320).

So, why people find pushing the Fat man way more problematic than just turning the switch? Apparently different parts of brain get involved. Simulation assumption might help a bit: when one simulates

turning the switch, the act itself looks neutral, apart from indirect consequences. When one simulates pushing a person from a bridge, it feels like actually doing it. The neurologists (Roth et al. 1996) tell us that in people simulating movement the primary motor cortex gets involved, as if they were themselves doing the hand movements. If this holds, there is a qualitative difference in feeling when one is simulating turning the switch, a neutral indirect causing of change in trolley's path, and when one is simulating the effortful pushing of a heavy object, the Fat man. If one feels imaginatively enacting the later as if it were one's actual effort, it is clear why it feels like killing the man with one's hands. Indeed, here simulation might explain the difference in feeling. (But more research has to be done before any definitive conclusion is taken.)¹²

The other pretty obvious kind of perspective taking are the Golden Rule cases. It is here that the very rich literature on empathy, often connected with sympathy, and sometimes distinguished, becomes relevant.¹³ And simulation normally generates empathy and sympathy (see Copman and Goldie 2011).¹⁴

Here, the issue of moral evaluation intervenes. Let me quote the philosopher who connects morality and empathetic simulation very directly and radically. It is Mark Johnson.

Moral imagination is our fundamental capacity to imagine how certain values and commitments are likely to play out in future experience, without actually performing those actions and having to deal with their lived consequences. The quality of our moral thinking therefore depends on (1) the depth and breadth of one's knowledge of the physical and social worlds he or she inhabits, (2) one's understanding of human motivation and cognitive/affective development, (3) one's perceptiveness of which factors are most relevant in a particular situation, and (4) one's ability to simulate the experiences and responses of other people with whom you are interacting. It is thus as much an affair of imagination as it is an appropriation of prior knowledge. (Johnson 2016: 363)

Passing to moral imagination Johnson characterizes it simply as simulation. It gives us "a deep sense of how others might experience a situation" and he connects it with empathy and talks about "empathetic imagination," which, in his view, makes it possible for us to appreciate and take up the part of others.

Let us now pass to social contract PTEs which make up one of the two most prominent kinds and traditions of macro-PTEs. (With apologies for very little space dedicated to each famous PTE in the tradition;

¹² The reader might like to consult the chapters on imagination and morals by Thomas Schramme, Antti Kauppinen, Alison E. Denham, David Shoemaker, Ishtiyaque Haji and Maurice Hamington in Maibom (2017).

¹³ On Golden Rule and empathy see Neusner and Chilton (2008), Pfaff (2007) and Wattles (1996: 144ff.).

¹⁴ Goldman has anticipated it in his (1992) again reprinted in Goldman (2013: 174–197).

I am looking for a general pattern). They are ideal for bringing simulation and political thought-experimenting together. Here, in contrast to Plato's tradition of building of the ideal state from the third person perspective, the interlocutor is asked to consider the possibility of living in some given arrangement and she is expected to imagine herself actually doing it. Here is a fine, relatively recent statement connecting the tradition to perspective taking:

Contract theorists hold that to judge whether an action or institutional arrangement is morally justified, one must determine whether it is in conformity with principles that would be the object of agreement. They thus assume that persons are able to discern the content of this hypothetical agreement. They thereby assume, I will argue, that persons are able to determine the acceptability of principles from other perspectives than their own present point of view. This is one out of two assumptions on which my investigation regarding the empirical plausibility of contract theory will concentrate. (Timmerman 2014: 2)

Now, behind the Veil the participant does not know how rich s/he will be. S/He has to imagine him/herself being very rich (wow!), being moderately well off (not bad!) and being very poor (God forbid!). According to my general proposal, s/he uses his/her default knowledge of being rich, well off and poor. How does the knowledge then get used? Not inbuilding a further model from the third-person perspective, but in simulating: let me imagine myself being poor, etc.! Let me remind you of Rawls' formulation from his *Theory of Justice*:

The aim is to rule out those principles that it would be rational to propose for acceptance, however little the chance of success, only if one knew certain things that are irrelevant from the standpoint of justice. For example, if a man knew that he was wealthy, he might find it rational to advance the principle that various taxes for welfare measures be counted unjust; if he knew that he was poor, he would most likely propose the contrary principle. To represent the desired restrictions one imagines a situation in which everyone is deprived of this sort of information. One excludes the knowledge of those contingencies which sets men at odds and allows them to be guided by their prejudices. In this manner the veil of ignorance is arrived at in a natural way. (Rawls 1999: 16)

As we mentioned, we shall concentrate on stages two to six, stressing the third stage. At *stage one*, the question is understood by you: you realize that you have to decide on purely rational grounds, in your own best interest. At *stage two* you start consciously building the model of the scenario suggested. You might be tempted to take a risk: why not special privileges for the rich ones, at the expense of the poor ones. But then you imagine yourself being poor, and people you know suddenly being very privileged rich ones. Here the simulation might set in.

The *third stage*, we propose, concerns the production of the answer, involving the generation of intuition as to whether the arrangement is acceptable to you. This probably involves decision making at the unconscious level. Your cognitive apparatus might revive some memories of poor people that you have suppressed from your consciousness, and

they might at the end motivate you not to risk. Richer simulation helps. You then come first with an immediate, unconscious intuition (I don't want to risk, I want an I arrangement that will be generous to the poor), at the stage *four*; other consideration intervene, and at the *fifth* stage, you come out with explicit intuition at the conscious level: I don't want to risk extreme forms of poverty, I want a decent life even if I am not rich.

Let us return to Goldman's schema. Call me T. Remember, I have a belief box (of oval shape), with the relevant belief: if I reject the privileges for the rich (**m**), I might end up having a decent life (**g**), even if I am relatively low on the social scale. I also have a desire box, square shaped in the drawing. The desire to have a decent life is sitting there. My simulating apparatus, of hexagonal form, puts together the two contents, **m**→**g** and **g**. But how can I, the imagined or simulated T, get to **g**? Well, by **m**—rejecting the privileges for the rich.

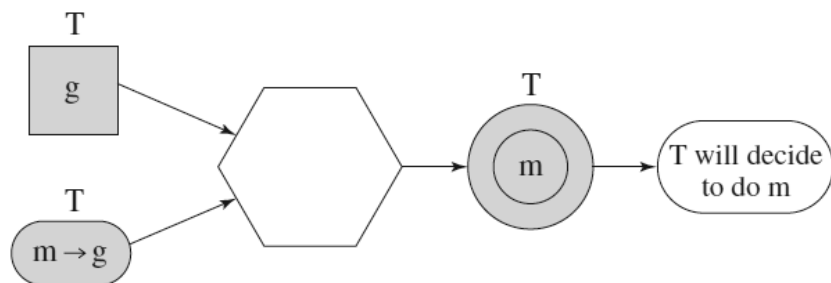


Figure 2.3. Decision attribution reached by simulation. (Adapted from Gallese and Goldman, 1998, with permission from Elsevier.)

So, T will decide to reject privileges for the rich. Rawls is vindicated.

But wait, I have assumed I shall be a well-surviving gentleman? But what if I turn out to be a female? And turn out to love children? I want more chances for myself and for them. This *sixth* stage of varying and generalizing, the intuitive induction, might make you even more egalitarian: I want children of relatively poor couple to have equal opportunities as children of relatively rich ones

What does cognitive study of simulation tell us about these processes and capacities involved? As we noted, there are two different ways of perceiving the other's situation, and these are often confused. First, you can imagine how another person sees his or her situation and feels as a result (an *imagine-other perspective*). Second, you can imagine how you would see the situation were you in the other person's position and how you would feel as a result (an *imagine-self perspective*).

Goldman notes that "egocentric" mindreading tendencies are found in both children and adults. Goldie described the imagine other perspective as "imagining the enactment of a narrative from that other person's point of view" (1999: 397). The result is not simply understand-

ing, but *sensitive* understanding. It is this form of perspective taking that has been claimed to evoke other-oriented empathic concern (Batson 1987, 1991). This imagine-self perspective connects self-recognition to other recognition (see Pfaff 2007: 65ff). A developed account of this kind appeals specifically to mental simulation or something sufficiently like it. (see Pfaff 2007: 69ff.). C. Daniel Batson, a famous author in cognitive study of perspective taking writes:

An imagine-self perspective involves, in Adam Smith's colorful phrase, "changing places in fancy." It has also been called "mental simulation" (Goldman 1992; Gordon 1992). Especially when the other's situation is unfamiliar or unclear, imagining how you would feel in that situation may provide a useful, possibly essential, basis for sensitive understanding of the other's plight. It may provide a stepping-stone to imagining how the other is affected by his or her situation and so to empathic concern. But if the other differs from you, then although focusing on how you would think and feel in the other's situation may provide comparative context, it also may prove misleading (...). (Batson 2009: 268).

Back to Rawls: the easier task is for myself, a male with long life experience, to imagine myself being poor. It is the case of imagining myself, as I am in a different situation. The difficult task is to imagining myself being a relatively young women with with a strong attachment to my newborn child, who needs me 24 hours a day. Human beings can in principle do both Goldman's sketch of simulation offers an elegant way to depict the process (he mentions the connection (2006: 294), unfortunately without developing it).

In the situation we are discussing, I am cognitively to "quarantine" my beliefs and desires that are irrelevant (2006: 30). Let me apply it to the reasoning under the Veil. Change the meaning of **g**, **h** and the rest. Suppose that **g** stands for "I want good circumstances for my child to develop and live in", and suppose that I am a relatively indifferent male. For me, then, $\sim\mathbf{g}$ holds. I might also have a belief **h** that my ability to struggle and achieve good conditions for myself are way more important than social care for children. Then I will never arrive at doing **m**, say accepting a very high degree of egalitarianism.

Well, what I should do is to quarantine $\sim\mathbf{g}$, **h**, and **my reservations about the m-g connection**, $\sim(\mathbf{m}\rightarrow\mathbf{g})$ belief. Rejecting $\sim\mathbf{g}$ makes me want good circumstances for my imagined child to develop and live in, rejecting **h**, helps me to avoid unreasonable self-confidence. I realize that accepting a very high degree of egalitarianism (our **m**) would provide the right circumstances for my would-be child ($\mathbf{m}\rightarrow\mathbf{g}$):

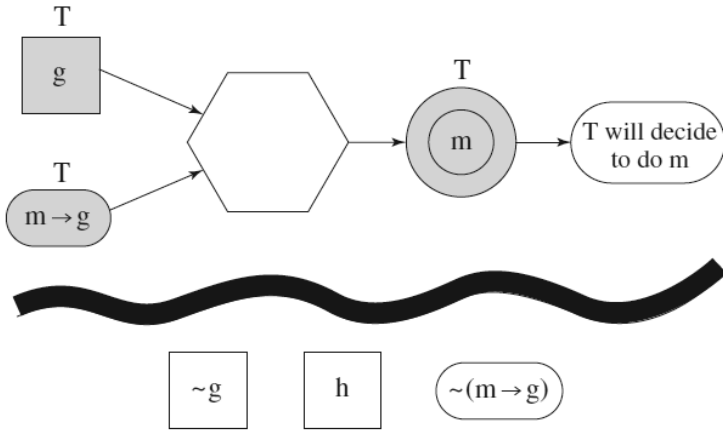


Figure 2.4. Decision attribution reached by simulation, showing quarantine.

I opt for **m**, and end up with the Rawlsian choice. All in all, I have to abstract from (‘quarantine’) my other interests and perhaps some relevant (male-centered) beliefs. This brings us to a frequent objection to Rawls’s Original position:

... the parties are deprived of so much information that they are incapable of making any choice at all. How can we make any rational choice without knowledge of our fundamental values? To begin with, the parties do know of their need for the primary goods and their higher-order interests in the moral powers. (Freeman 2007: 160).

Translated into cognitive terminology, how much information, in particular information about myself, can I quarantine, and still competently decide about the right choice for *m*? Let me try a hunch of an answer to the question quoted. Supposed I am behind some science-fictional veil of ignorance, let us say abducted by aliens from an inhabitable exoplanet, say Kepler 62 f, and I know two things. First, there are different societies co-existing there, some more tolerant, some less, and second, the aliens might tamper with my brain, and change the values I shall wake up with after the tempering. What society should I choose? It seems obvious to me that I should opt for the most tolerant one. In the worst case, if I am to wake up with a lot of crazy ‘values’ ruling in my brain, I want now to be tolerated once this happens; the more tolerant the society, the better for me. The Minimax gives the right answer: choose the society which will offer most even in the worst case. A lot more should be said, but I believe that the quarantining interpretation offers a good first step in direction of an answer.

Let me come to the end of my short list of illustration of famous TEs that seem to demand simulation in their implementation, with a brief, all too brief pointing to the work of Thomas Scanlon. One of his many examples is the right to privacy (1998: 204), but he does not give any

detailed recipe; so let me try to provide one. How would one argue for the right, discussing matters with a somewhat voyeuristic neighbor? The first move is like the Golden Rule one: one can ask the neighbor to imagine that he is being peeped upon. Imagination will involve simulation. If the neighbor sees the point, one can try to offer a more general proposal. Imagine other people, how would they feel if deprived of right of privacy? More simulation might be required. Here is Scanlon's general statement:

Some of the most common forms of moral bias involve failing to think of various points of view which we have not occupied, underestimating the reasons associated with them, and overestimating the costs to us of accepting principles that recognize the force of those reasons. (Scanlon 1998: 206)

The simplest way to recognize the reasons associated with "points of view which we have not occupied" is by trying to simulate them. We "quarantine" our own point of view, and replace it with the target one, and then enter simulation *n*. (Habermas discusses "taking the attitude on the other" commenting G. H. Mead in the fundamentally important chapter on Mead and Durkheim of his *The Theory of Communicative Action* v. 2, from 1981. He then incorporates it into his own theory as its basic assumption; for a brief, principled statement see his (1995: 117)).

Let me note that Scanlon's most famous work on the topic has been done in the area of ethics; however, like other contractualists, he connects it with political philosophy and talks about morality of institutions (2016) along the same lines he proposed for individual morality.

So, back to the stages of TE, armed with a sketch of Simulation Theory. We have located the perspective taking at the stage four, the one in which the scenario proposed is being worked out. At the next, *fifth stage*, thinker comes out with explicit intuition at the conscious level, usually geared to the particular example and having little generality.

In our example, the male thinker has imagined being a female, and has arrived to the decision that the best course for him would be to opt for gender equality in the future society.

Sixth stage: since the typical job in previous stage is consideration of some particular scenario, the thinker will next have to do some varying and generalizing (deploying both moral and rational competences) at the conscious and reflective level and, perhaps, at the unconscious one too. Sometimes this process is called intuitive induction (Chisholm 1966). In our example, the thinker imagines himself as being poor, and then as being not very talented for well-paid jobs, and so on.

In the Veil-of-ignorance kind of TEs the experimenting yields a series of prudential answers-intuitions. What about the moral judgment? It is the result of more theoretical reflection, after the descriptive information gained by simulation has been systematized. (Rawls sometimes talks about a wider framework of entering social contract, with "strains of commitment" securing the moral side, but we cannot enter it now; for a fine analysis see Waldron's "Strains of Commitment" in Hinton (2015)

This kind of combined strategy is a must for the classical contractualist tradition. In Kant (and Parfit) you decide about the moral status of a maxim after you have calculated the consequences of its becoming the universal rule. In Scanlon, you decide about the moral status of your proposal after you have gone through imagining other people's reactions to it, and your attempts at persuading them. In Rawls, you decide about the normative status of your proposal after you have tested it under the Veil, possibly comparing it to other alternatives, and calculating which of them will assure the maximin result. Call the first task the descriptive-factual exercise, and the second the normative derivation. Note that Habermas and Scanlon build more normativity into the decision phase. Habermas, for example, derives it from the regulative use of speech: "The social reality that we address in our regulative speech acts has by its very nature an *intrinsic* link to normative validity claims" (Habermas 1990: 61).

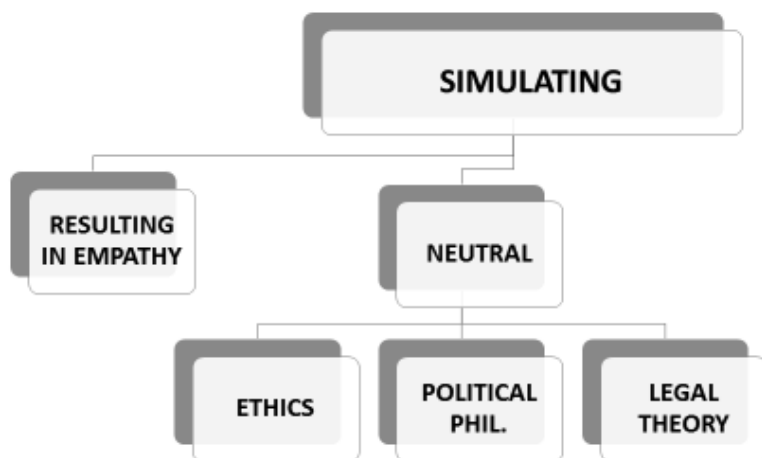
In simpler kinds of practical TEs, like the Fat Man and Golden Rule, the initial moral judgment is the direct result of empathy generated by simulation. It offers an account of how morality enters the picture, congenial to sentimentalists ethics.¹⁵

Let me conclude by mentioning an example from philosophy of law, the area that has not been much discussed until now in terms of thought-experimenting (but see in this issue the paper of Miomir Matulović, to whom I also owe the example that follows). Friedrich Carl von Savigny talks about the interpreter reconstructing the thought (*Gedanke*) "enclosed in a law". Good interpreters should put themselves in the same starting point of the legislator, and "artificially repeat in themselves his way of proceeding, so that the law may come to be born again in their mind" (1867: 171). Here, we are explicitly asked to step into ancient legislator's shoes, and pretend we are legislating in his place. This putting oneself in the place of the author ("*Gleichsetzung mit dem Verfasser*", in the original German) seems to be a common strategy suggested in early nineteenth century German hermeneutics. Schleiermacher asks: "But how can we understand the inner process of the writer from this? By observation. But this is based on self-observation" (1998: 135; the original manuscript dates from 1828). And, he claims that ".../ one must put oneself in the place of the author on the objective and the subjective side" (1998: 24); the "objective" here is "linguistic" and the "subjective" is the psychological.¹⁶

Here is then the overview of the areas where simulation plays a central role:

¹⁵ See for instance Slote (2007). The darker side of this connection, recently intensely discussed in connection with the Fat Man TE, is the possibility that empathetic gut reactions usurp the place of rational consideration. See for discussion and references Cushman, Young, and Greene (2010).

¹⁶ For parallels with Savigny's contemporaries writing about understanding and empathy in general see the historical overview in the Introduction to Coplan and Goldie (2011). See also Girard (2017) and Leyh (1992).



2.4. *Should we be pessimistic about simulation?*

A quick glance at the literature

What about difficulties connected with simulation? Let me briefly mention a few problems raised by some authors, then an optimistic counterproposal, and conclude with moderate optimism. The simple schemas we reproduced seem to suggest that quarantining and putting oneself in another's shoes is an easy matter, but of course, neither Goldman nor cognitive psychologists think it is. Here is a characterization of some difficulties.

Epley and Caruso (2009: 297ff) list three “critical barriers” as they call it: activating the ability to simulate, adjusting “an egocentric default” (in their parlance), and accessing information about others. For this, they have to do several things. First, they must actively think about another person's mental state, thus activating the process of perspective taking. Second, they must abstract from their own characteristics, which is normally not easy. Third, they must deploy non-egocentric information about other people in a skilled manner. (Ibid.). None of this is particularly easy.

Some authors go much further in pointing to problems. Let me choose a recent warning due to Shannon Spaulding. In her paper on “Simulation Theory” (2016a), and even more in “Imagination Through Knowledge” (2016b) she comes up with interesting challenges brings further challenges. But before doing this, she offers clarificatory and classificatory information that is extremely useful, given that the term “simulation” is used in many senses, and there is clear need to distinguish them to avoid very bad confusions; let me summarize the information quickly. Spaulding starts from Goldman's proposal according to which a process P simulates process P' if and only if first P duplicates, replicates, or resembles P' in some significant respect and two,

in its duplication of P', P fulfills one of its purposes or functions. In the case of mindreading simulation, the purpose or function of is to understand target's mental states. She then introduces the crucial distinction between abstract and concrete simulation, the first including activities like computer simulation and the like, and the second being the psychological simulation that involves "sameness of system and fine-grained process" (2016a: 264). The distinction is very helpful, and could save writers from confusions that mark the scene of present-day investigation of simulation.

Spaulding next distinguishes high-level from low-level simulation. She lists three characteristics of the former. First, it "involves imagination in the conventional sense" (267) Second, it explains our engagement with fiction, where we put ourselves "in the fictional character's position and imagine what we would think, feel, and do in that situation. Third, it explains how one can get knowledge through simulation, so that it could be "co-opted to explain how some thought experiments work" (267). In contrast, in low level simulation, "imagination operates unconsciously and automatically."

Now, on the skeptical side, Spaulding's most direct challenge to the project of finding constraints that would rehabilitate imagination is to be found here, "Imagination Through Knowledge" (2016b). On her view, the puzzle of how we arrive to knowledge through imagination suggests that imagination is "not sufficient for new knowledge" (2016b: 222). The argument seems to be the following: if imagination is to be constrained by extra-imaginative pieces of information and by other abilities, then imagination does not bring new knowledge. But this is too severe a demand. Compare physical constraints. I commute from my home town to my working place about hundred miles distance. For the car to bring me to my work there should be a well-established and well-kept road, constraining the travel, there should be red lights helping to prevent crashes, and so on. Imagine someone arguing that therefore "car is not sufficient" for commuting, and is not doing any real work! Well, the fact that an item needs constraints to function properly does not entail that it never performs any function.

Spaulding has an auxiliary argument: "I have argued that the cognitive capacity to imagine scenarios is distinct from the cognitive capacities that underlie our ability to judge the accuracy of our imaginings" (2016b: 222) and ".../there is nothing in the capacity of imagination itself that could evaluate the accuracy of the possibilities we imagine." (2016b: 222). Indeed, there is nothing in the car itself that recognizes red/green light. This does not show that the car will not take me from home to work, only that car *alone* will not do the work. So much about Spaulding's direct challenge to the instructive use of imagination.

Let me mention, however, that in her text the challenge is preceded by a rich and very provocative analysis of one particular kind of imaginal enactment, namely simulation. Her argument resembles the general one we just summarized. Her example is the following: I watch

John tease Mary, and try to figure out why he is doing this. I simulate his activity, and end up concluding that John likes Mary and is trying to get her attention. Fine, but how do I choose this option rather than some other, equally plausible in itself, for instance that he is just humiliating her? I need additional information, and my simulation tells me nothing about these matters. Again, to me it looks like simulation has done the main job, like the car in our example; the fact that the main job cannot be fully accomplished by the main agency in question, tells little against it.

So much about criticism;¹⁷ we had to be very brief. For balance, let me conclude this section by mentioning a very helpful and more optimistic book, Peter Timmerman's 2014 *Moral Contract Theory and Social Cognition* who comes very close to our topic with an important difference—his is moral contractualism rather than the political one (and he says nothing about the psychological mechanism that makes accessible to people „other perspectives than their own present point of view.”) But he has a lot to say in defense of the view that simulating oneself in various situations and simulating others are in principle within one's power.

He notes several differences between the kind of perspective-taking that normally interest psychologists, and the kind relevant for contract theorist. For example, there is the difference in the target (Timmerman 2014: 36). In contrast to psychologists who are interested in factual agreement, “we need to find out not whether others would in fact agree to principles that permit it but whether *they have reason* to do so. We are thus not first and foremost interested in what they *would* think or feel about a principle. We are, however, interested in a closely related question. As we need to determine whether others have reason to agree to a principle, we are interested in *what they would reasonably think or feel* with regard to the principle. The second difference “concerns the sort of perspectives that are considered”. Philosophers are interested in general, abstract viewpoints, psychologists in our ability to recognize perspectives of “particular others” (Timmerman 2014: 36).

He further distinguishes several variables relevant for moral contract, and his picture can be easily applied apply it to the political one (Timmerman 2014: 26ff.). The first variable, he writes, “concerns which agents can use the procedure adequately to form moral judgments.” A second variable, concerns *the circumstances* under which agents can apply the procedure, and the third the extent of their capacities. For all three cases, he comes close to contrasting idealizations and realistic proposals. He has some fine ideas about measures that could help normal agents to face the daunting task(s). He assumes, (...) that potential interaction partners can detect whether one can be trusted to comply or not, and as such will refrain from interacting cooperatively with persons who are not disposed to comply.) Also, he argues that

¹⁷ See Klampfer (2018) in this issue.

we may assume that “persons are able to determine the acceptability of principles from other perspectives than their own present point of view”. He mentions two important means, information gathering and “the internalization of moral principles” to which the biggest part of the Chapter Three of the book is dedicated.

Of course, the discussion between PTE-defenders and PTE-skeptics is going on, but I think we have no reason to be pessimistic about the basic abilities involved in political thought-experimenting. Let me then conclude.

3. *Conclusion*

In our investigation of cognitive mechanisms of PTEs, we have tried to bring together two blossoming traditions: the study of perspective taking and methodology of thought-experiments. Both are extremely rich, but we have narrowed our topics down to PTEs on the side of experiments, and to mental simulation on the side of perspective taking.

We have discussed the kind of PTEs that has marked the central tendency in a tradition of political philosophy, active at least from Kant on, but especially since and including Rawls’ *Theory of Justice*. It is the tendency to view political justice and moral value in terms of a hypothetical contract. Political and moral TEs presented within the contractualist tradition (in the widest sense) typically ask the thought-experimenter to imagine how other people would take the experimenter’s moral and political proposals, how they would feel about them, and whether and how they could be persuaded to accept them. A somewhat special but perhaps most famous case is imagining what one would propose as political arrangement if one were ignorant about one’s abilities and material situation in the future situation. Again, I apologize for cramming together all the famous PTE in the tradition, each of which deserving at least a long paper attempting to account for its mechanism; but this is the price of arriving at a general pattern, if all goes well.

We have concentrated upon contractualist methodology, where imagining is supposed to yield factual intuitions about whether the subject(s) in question would accept proposed arrangements, and the normative work is done by theory. However, we have noted that there is another, more direct route to normative judgements, directly from empathy provoked by simulation, explored by a number of cognitive psychologists and stressed by Goldman on the side of philosophers; it is a matter relevant for sentimentalist ethicists, but also for understanding some very popular TEs, like the Trolley and Fat man ones. Here, the appeal to simulation yields a fine by-product, a more direct route to moral judgment.

How do all these TEs work? Our moderately inflationist mental modelling proposal is that they mobilize our imaginative capacity for perspective taking, most probably perspective taking through simulation. The framework proposed is moderately optimistic; it suggests the

answers to questions that are often raised for other kinds of TEs as well. To quote James Robert Brown, one wonders how one can learn new things without new observational data? (Brown 1991: 111ff.). In the case of our PTEs, the data come from perspective taking: the information producing capacity is either the capacity to simulate or some closely related ability. His second worry, why are thought experiments superior to deduction in terms of heuristic power, obviousness and ease, can be alleviated or even discarded by appeal to the fact that mental simulation is way more accessible to subjects than abstract political reasoning from principles and facts, and its output is usually quite obvious to the subject. The third question is: where does the “experiential” element in thought experiments come from? Are there any new experiences or quasi-experiences present in thought experiment, and of what nature are they? Yes, it is a new experience, namely the experience of simulating. In the case of empathetic simulation, the qualitative, emotional character of experience is highly prominent, and in the case of less emotional simulation, it still has experiential character (“Let me imagine that I am a generally incapable person; how would I feel in a strongly competitive society, at the bottom of its pecking hierarchy?”). Finally, as Brown puts it “if the reasoning in thought experiment is broadly inductive, how can it eliminate alternatives and reach its conclusion so quickly and effortlessly, and assert it with such force?” (Brown 1991: 111ff.). Simulation normally is quick and effortless; the simulator does not go through alternatives, but is constrained in an unconscious way.

Let us conclude by placing the account within a bigger picture, returning for the moment to our starting point. We have distinguished two kinds of PTEs and two manners of imagining political arrangements. The first consists in building third-person mental models, based on our inductive knowledge, and on default assumptions about people, about practices and institutional arrangements. The second consists in perspective taking, imagining oneself (as oneself or even as someone else) and asking about condition one would accept. Golden Rule and social contract are prime examples, either in realistic or somewhat unrealistic scenarios of ignorance and/or ideal rationality and the like.

We have proposed first-person mental simulation as the basic mechanism, although we did not insist on the “purity” of mechanism. (Goldman himself proposes the idea of a „hybrid theory” according to which the simulation and the reasoning on the bases of theoretical knowledge about human minds (theory-of-mind, can interact, for example ‘cooperate’ (ch. 2.7 of his 2006 book); this might be an interesting option, to discuss at some other occasion. And of course, simulation might make occasional appearance in the first, predominantly first-person model building; the author might ask the reader how she would feel in such and such an arrangement, something that happens all the time in *The Republic*. But, from a wider perspective, the two mechanisms, model building and simulation, and their combination(s) exhaust the range of psychological

mechanism underlying political thought-experimenting. This is the ambitious proposal to which the present paper is a tentative contribution.

References

- Batson D. C. 2009. "Two Forms of Perspective Taking: Imagining How Another Feels and Imagining How You Would Feel." In Markman, K. D., Klein, W. M. P., and Suhr J. A. (eds.). *Handbook of imagination and mental simulation*. New York: Psychology Press Taylor and Francis Group.
- Brown, J. R. 1991. *The Laboratory of the Mind Thought Experiments in the Natural Sciences*. London: Routledge.
- Carruthers, P. and Smith, P. K. (eds.). 1996. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.
- Chisholm, R. 1966. *Theory of knowledge*. New Jersey: Prentice Hall.
- Cohen G. A. 2009. *Why Not Socialism?* Princeton: Princeton University Press.
- Clohesy, A. M. 2013. *Ethics, solidarity, recognition*. London: Routledge.
- Coplan, A. and Goldie, P. 2011. *Empathy Philosophical and Psychological Perspectives*. Oxford: Oxford University Press.
- Currie, G. and Ravenscroft, I. 1997. "Mental Simulation and Motor Imagery." *Philosophy of Science* 64 (1): 161–80.
- Currie, G. 2002. "The Simulation Programme." In Currie, G. and Ravenscroft, I. (eds.). *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Oxford University Press.
- Cushman, F., Young, L., and Greene, J. D. 2010. "Multi-system Moral Psychology." In Doris J. M. (ed.). *The Moral Psychology Handbook*. Oxford: Oxford University Press: 47–70.
- David D. 2018. "Art and thought experiments." In Stuart, M. T, Fehige, Y., and Brown, J. R. (eds.). *The Routledge Companion to Thought Experiments*. London: Routledge: 512–525.
- Davies, M. 1987. "Tacit Knowledge and Semantic Theory: Can a Five per Cent Difference Matter?" *Mind* 96 (384): 441–462.
- Davies, M. and Stone, T. (eds.). 1995a. *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell Publishers.
- Davies, M. and Stone, T. (eds.). 1995b. *Mental Simulation: Evaluations and Applications—Reading in Mind and Language*. Oxford: Blackwell Publishers.
- Davies, M. 2001. "Mental Simulation, Tacit Theory, and the Threat of Collapse." *Philosophical Topics* 29 (1/2): 127–173.
- de Waal, F. 2009. *The Age of Empathy: Nature's Lessons for a Kinder Society*. New York: Random House.
- Decety, J. and Grèzes, J. 2006. "The power of simulation: Imagining one's own and other's behavior." *Brain Research* 1079: 4–14.
- Epley, N. and Caruso, E. M. 2009. "Perspective Taking: Misstepping Into Others' Shoes." In Markman, K. D., Klein, W. M. P. and Suhr, J. A. (eds.). *Handbook of Imagination and Mental Mimulation*, (New York: Taylor and Francis Group).
- Freeman, S. 2007. *Rawls*. London: Routledge.

- Frith, C. 2007. *Making up the Mind: How the Brain Creates Our Mental World*. Oxford: Blackwell.
- Gallese, V. and Goldman, A. I. 1998. "Mirror Neurons and the Simulation Theory of Mindreading." *Trends in Cognitive Sciences* 2: 493–501.
- Gauthier, D. 1986. *Morals by Agreement*. Oxford: Clarendon Press.
- Goldie, P. 1999. "How we think of others' emotions". *Mind and Language* 14: 394–423.
- Goldman, A. I. 1992. "Empathy, mind, and morals". *Proceedings from the American Philosophical Association* 66: 17–41. Reprinted in Goldman 2013: 174–197.
- Goldman, A. I. 1995. "Empathy, Mind, and Morals". In Davies, M. and Stone, T. (eds.). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell: 185–208.
- Goldman, A. I. 2005. "Simulationist Models of Face-Based Emotion Recognition." *Cognition* 94: 193–213.
- Goldman, A. I. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.
- Goldman, A. I. 2008. "Mirroring, Mindreading, and Simulation". In Pineda, J. (ed.), *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition*. New York: Humana Press: 311–330.
- Goldman, A. I. 2012. "Theory of Mind." In Margolis, E., Samuels, R., and Stich, S. P. (eds.). *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford: Oxford University Press: 402–424.
- Goldman, A. I. 2013. *Joint Ventures Mindreading, Mirroring, and Embodied Cognition*. Oxford: Oxford University Press.
- Goldman, A. I. Forthcoming. "Mindreading by Simulation: The Roles of Imagination and Mirroring". In Lombardo, M., Tager-Flusberg, H. and Baron-Cohen, S. (eds.). *Understanding Other Minds*. 3rd ed. Oxford: Oxford University Press.
- Gordon, R. M. 1986. "Folk Psychology as Simulation." *Mind and Language* 1 (2): 158–171. Reprinted in Davies and Stone 1995a: 60–73.
- Gordon, R. M. 1992. "The Simulation Theory: Objections and misconceptions." *Mind and Language* 7: 11–34.
- Gordon, R. M. 1995. "Simulation Without Introspection or Inference from Me to You." In Davies and Stone 1995b: 53–67.
- Gordon, R. M. 1996. "'Radical' Simulationism." In Carruthers and Smith 1996: 11–21.
- Gordon R. M. and Barlassina, L. 2017. "Folk Psychology as Mental Simulation." *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Zalta, E. N. (ed.), URL = <<https://plato.stanford.edu/archives/sum2017/entries/folkpsych-simulation/>>.
- Habermas, J. 1989b. "Towards a Communication-Concept of Rational Collective Will-Formation: A Thought-Experiment." *Ratio Juris* 2: 144–154.
- Habermas, J. 1990. "Discourse Ethics." In *Moral Consciousness and Communicative Action*, Cambridge: Polity Press: 43–115.
- Habermas, J. 1994. "Remarks on Discourse Ethics." In *Justification and Application. Remarks on Discourse Ethics*. Translated by Ciaran Cronin. Cambridge: The MIT Press: 19–112.
- Habermas, J. 1995. "Reconciliation Through the Public Use of Reason: Remarks on Rawls's Political Liberalism." *Journal of Philosophy* 92 (3): 109–131.

- Hinton, T. (ed.). 2015. *The Original Position*. Cambridge: Cambridge University Press.
- Hohwy, J. 2013. *The Predictive Mind*. Oxford: Oxford University Press.
- Johnson, M. 2016. "Moral imagination." In Kind, A. (ed.). *The Routledge Handbook of Philosophy of Imagination*. London: Routledge.
- Johnson-Laird, P. N. 1983. *Mental Models*. Cambridge: Cambridge University Press.
- Kind, A. and Kung, P. (eds.). 2016. *Knowledge Through Imagination*. Oxford: Oxford University Press.
- Leyh, G. 1992. *Legal Hermeneutics: History, Theory, and Practice*. Berkeley: University of California Press.
- Maibom, H. M. (ed.). 2017. *The Routledge Handbook of Philosophy of Empathy*. London: Routledge.
- Metzinger, T. 2003. *Being No One The Self-Model Theory of Subjectivity*. Cambridge: The MIT Press.
- Minio-Paluello, I., Avenanti, A., and Aglioti, S. M. 2006. "Left hemisphere dominance in reading the sensory qualities of others' pain?" *Social Neuroscience* 1 (3–4): 320–333
- Mišćević, N. 1992. "Mental models and thought experiments." *International Studies in the Philosophy of Science* 6 (3): 215–226.
- Mišćević, N. 2006. "Intuitions: the discreet voice of competence." *Croatian Journal of Philosophy* 16: 69–96.
- Mišćević, N. 2012a. "Plato's *Republic* as a Political Thought Experiment." *Croatian Journal of Philosophy* 12 (2): 153–165
- Mišćević, N. 2012b. "The competence view of intuitions—a short sketch." *Balkan Journal of Philosophy* 4 (2): 147–160.
- Mišćević, N. 2013a. "Political Thought Experiments from Plato to Rawls." In Frappier, M., Meynell, L., and Brown, J. R. (eds.). *Thought Experiments in Science, Philosophy, and the Arts*. London: Routledge.
- Mišćević, N. 2013b. "In Search of the Reason and the Right—Rousseau's Social Contract as a Thought Experiment." *Acta Analytica* 28 (4): 509–526.
- Mišćević, N. 2018. "Thought experiments in political philosophy." In Stuart, M. T, Fehige, Y. and Brown, J. R. (eds.). *The Routledge Companion to Thought Experiments*. London: Routledge: 153–170.
- Neusner, J. and Chilton, B. 2008. *The Golden Rule the Ethics of Reciprocity in World Religions*. London: Continuum.
- Ninan, D. 2016. "Imagination and the self." In Kind, A. (ed.). *The Routledge Handbook of Philosophy of Imagination*. London: Routledge: 274–285.
- Parfit, D. 2011. *On What Matters. Volume One*. Oxford: Oxford University Press.
- Peacocke, C. (ed.). 1994. *Objectivity, simulation and the unity of consciousness*. London and Oxford: British Academy and Oxford University Press.
- Perner, J. and Kühberger, A. 2005. "Mental Simulation Royal Road to Other Minds?" In Malle, B. F. and Hodges, S. D. (eds.). *Other Minds How Humans Bridge the Divide between Self and Others*. New York: Guilford Press: 174–189.
- Pfaff, D. W. 2007. *The neuroscience of fair play: Why We (Usually) Follow the Golden Rule*. New York: Dana Press.

- Rawls J. 1999. *A Theory of Justice*. Revised edition. Cambridge: Harvard University Press.
- Rawls, J. 1986. *The Basic Liberties and Their Priority, The Tanner Lectures on Human Values April 10, 1981*. In McMurrin, S. M. (ed.). *Liberty, Equality, and Law: Selected Tanner Lectures on Moral Philosophy, The Basic Liberties and Their Priority*. Salt Lake City: University of Utah Press.
- Roth, M. et al. 1996. "Possible involvement of primary motor cortex in mentally simulated movement: a functional magnetic resonance imaging study." *Neuroreport* 17 (7): 1280–1284.
- Savigny, F. C. von 1867. *System of the Modern Roman Law*. Madras: Higginbotham Publishers.
- Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge: The Belknap Press of Harvard University Press.
- Scanlon, T. M. 2016. "Individual Morality and the Morality of Institutions." *Filozofija i društvo* 27 (1): 1–36.
- Schleiermacher, F. 1998. *Hermeneutics and Criticism*. Cambridge: Cambridge University Press.
- Spaulding, S. 2016a. "Simulation Theory." In Kind, A. (ed.). *Handbook of Imagination*. London: Routledge: 262–273.
- Spaulding, S. 2016b. "Imagination Through Knowledge". In Kind and Kung 2016: 208–225.
- Slote, M. 2007. *The Ethics of Care and Empathy*. London: Routledge.
- Sokolowski, J. A. and Banks, C. M. 2009. *Modeling and Simulation For Analyzing Global Events*. New York: John Wiley and Sons.
- Timmerman, P. 2014. *Moral Contract Theory and Social Cognition*. Amsterdam: Springer.
- Vendler, Z. 1984. *The Matter of Minds*. Oxford: Clarendon Press.
- Wattles, J. 1996. *The Golden Rule*. Oxford: Oxford University Press.
- Williams, B. 1976. "Imagination and the Self." In *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge: Cambridge University Press.
- Zwaan, Z. A and Radvansky G. A. 1998. "Situation Models in Language Comprehension and Memory." *Psychological Bulletin* 123 (2): 162–185.

The 'Arguments Instead of Intuitions' Account of Thought Experiments: Discussion of The Myth of the Intuitive by Max Deutsch

ANA BUTKOVIĆ
University of Zagreb, Zagreb, Croatia

*After decades of receiving a lot of attention on the epistemological level, the so-called 'problem of intuitions' is now in the center of debates on the metaphilosophical level. One of the reasons for this lies in the unfruitfulness of the epistemological discussions that recently subsided without producing any significant or broadly accepted theory of intuitions. Consequently, the metaphilosophical level of discussion of the 'problem of intuitions' inherits the same difficulties of the epistemological level. The significance of Max Deutsch's book *The Myth of the Intuitive* is his effort to resolve these problems in a clear and persuasive way. He is not only trying to debunk problems behind the vagueness of the 'intuition-talk' by drawing important distinctions that usually go under the radar in the contemporary literature, but also develops his own account of philosophical methodology. In this paper I will present some of his arguments against the traditional view of intuitional methodology, as well as his own solutions to the presented problems. Regardless of Deutsch's insightful account of the 'problem of intuitions', I find that some difficulties in his own proposal are inherited from the unresolved issues of intuitions on the epistemological level.*

Keywords: Intuitions, evidence, thought experiments, arguments.

The 'problem of intuition' in recent years became the center of many epistemological and metaphilosophical discussions mainly because of the rise of experimental philosophy (xphi) and many criticisms raised against the method of cases, i.e. the method of appealing to intuitions elicited by thought experiments as evidence for or against some philosophical theory. The so-called negative program within the xphi got the most attention since their theses are the most challenging ones. In

a nutshell, negative program advances argumentation that intuitions, as used in philosophical thought experiments and hypothetical cases, should not be trusted nor relied on as evidence. This, rather pessimistic view of philosophical practice led to the increasing number of metaphilosophical papers and books as a response to this “restrictionist challenge” (Weinberg et al. 2010). Generally speaking, there are two most developed ways to respond to the restrictionist challenge. The first is to defend intuitions viewed as a source of evidence and the distinctive way of doing philosophy within the analytic tradition. The second is to claim that xphi misconstrue the target of surveys since philosophers are not appealing to intuitions as evidence in thought experiments. Deutsch devoted his book to defend the latter view. The central idea he develops is that it is a *myth* that philosophers rely on intuitions as evidence in thought experiments and that this myth needs to be debunked. Therefore, results of xphi’s surveys about untrustworthiness of intuitions as evidence are not troubling if the target of their surveys can be refuted. This is the strategy Deutsch advances as a part of his metaphilosophical account. In addition, he also elaborates his own view that it is arguments, rather than intuitions, that are the basis of any thought experiment and bearers of the evidential force. This is what I will call the *‘arguments instead of intuitions’* view.

I.

The first chapter of the book is devoted to xphi’s theoretical framework and distinction between its positive and negative program, as well as the analysis of results of recent xphi’s studies. Negative xphi program rises worry about the epistemic value of philosophical intuitions due to their susceptibility to the truth-irrelevant factors such as cultural background, gender, order effect, etc. Subsequently, they take a more negative stance toward the traditional philosophical method of appealing to intuitions and argue that intuitions cannot be trusted or relied on in philosophy as evidence. Positive xphi, on the other hand, is advancing more moderate conclusions that do not condemn the use of intuitions in philosophy.

In developing his ‘argument instead of intuition’ account as a way of responding to xphi criticism, Deutsch focuses on the two most discussed thought experiments, Gettier cases and Kripke’s Gödel case. For him, Gettier cases are somehow exceptional in a sense that if there is an appeal to intuitions anywhere in philosophy, then it is in Gettier cases. From xphi’s conclusion regarding intuitions, i.e. that not everyone shares Gettier’s intuition, it follows that, contrary to the established view in the last 50 years, Gettier has not refuted the JBT theory of knowledge since intuitions can not be trusted or relied on as evidence. Now, as Deutsch sees it, for this xphi’s argument to work experimentalists must assume not only that (i) philosophers are treating intuitions about cases as evidence, but also that (ii) intuitions are treated as es-

sentential or *only* evidence, which is a much stronger claim. Deutsch argues against both (i) and (ii) and concludes that negative χ phi critique fails to debunk traditional philosophical arguments that “do *not*, in any relevant sense, *depend on treating intuitions as evidence*” (20). In other words, χ phi fails to hit the target.

Deutsch substantiates his central thesis—the *myth of the intuitive*—with empirical and theoretical arguments. As he himself admits, he is doing this without any accepted theory of intuitions, the *no-theory theory of intuitions*, as he calls it. That way he “offers enough without offering too much” (29). The reason for this, according to Deutsch, is that accepting a theory of intuitions is not necessary for asking and answering questions about the role of intuitions in philosophy. Furthermore, any attempt to develop such a theory involves conceptual analysis of *intuition* in terms of necessary and sufficient conditions which ends up being a very difficult task because, ironically, every proposed analysis of intuition give rise to variety of counterexamples.

This looks like the right diagnosis of the current epistemological efforts to provide any plausible account of intuitions. So while epistemologists and metaphilosophers are endeavoring this futile, hard-to-settle task, Deutsch thinks that the best strategy for asking and answering some crucial questions about the role of intuitions in philosophy is to conduct empirical investigation of the actual practice via analyzing original texts where some of the most influential thought experiments were presented, with no-theory theory of intuitions. The rationale behind it is this:

It offers enough of an account because it allows for fruitful discussion of the argumentative role of intuitions. It offers not too much of an account because it does not invite the potentially endless cycle of counterexample-and-theory-revision endemic to many attempts at conceptual analysis. (29–30)

I am inclined to say that it is questionable whether this is a tenable move. Although I agree with everything Deutsch says regarding diversity of proposed accounts of intuitions, and unfruitfulness of the endeavor of analyzing the concept of intuition in terms of necessary and sufficient conditions only for it to become the target of endless counterexamples, there are some problems with no-theory theory approach. Particularly problematic is his claim that “a theory of the (psychological) nature of intuitions is not required for understanding the role of intuition in philosophical argument” (26). The natural questions arise: ‘How can one conduct an empirical analysis with no accepted theoretical framework of the analysandum?’, or ‘How can he or she “recognize an intuitive judgment when he or she encounters one” (29), if they do not have some general insight of what they are encountering?’.

Although Dutsch is calling his no-theory theory of intuitions the ‘*examples-plus-commonality*’ theory, it is far from clear how this would help to answer previous questions because he is never explicit about what those commonalities would be. The most precise he gets is say-

ing that examples he has in mind are those like Gettier cases, where certain judgments about certain hypothetical situation are made, and that philosophers agreed that they are intuitive judgments. And when analyzing sufficient number of these examples, he is able to abstract commonality from them that is “relatively uncontroversial” (31). It seems that for Deutsch to establish the no-theory theory of intuitions and to investigate whether thought experiments are about intuitive judgments, it is enough to find examples where the uncontroversial usage of certain judgments is present.

However, the no-theory theory does assume that we have examples of intuitive judgments about which those party to the debate over their role can agree; that is, these parties can agree that the examples are examples of intuitive judgments. (30)

So, the commonality among those examples is that “examples are all judgments about hypothetical cases and thought experiments” (32).

Difficulty with this approach is that there is so much diversity in what exactly philosophers find intuitive in original examples of thought experiments which results in diametrically different accounts of what intuitions are, and consequently results in diversity of the usage of the term ‘intuitions’. Wide arrays of views of what intuitions are lies between the views that they are *sui generis* states (e.g. Bealer 1998, Pust 2000), inclinations to believe (e.g. Sosa 2007) or simply beliefs (e.g. Lewis 1983, Jackson 1998). Or, if we have in mind views regarding the justificatory status of intuitions, some philosophers hold that such justification is a priori (e.g. Bealer 1998, BonJour 1998) and others argue that it is a posteriori (Devitt 2011, Kornblith 2007). Moreover even the most ‘uncontroversial’ features of intuitions, that of being spontaneous or noninferential, are controversial for Deutsch. He is appealing to Rawls’s method, which supposed to depend on intuitions, and yet Rawls explicitly says that the relevant judgments are our *considered* moral judgments and, therefore, cannot be spontaneous. Consequently, it is difficult to see what is the rationale for Deutsch’s thesis that philosophers are not using intuitions in thought experiments, as he has no clear description of what precisely is that thing that philosophers are not appealing to.

As I see it, Deutsch cannot proceed with his endeavor just by examining thought experiments with no accepted background theory. For instance, if he wants to argue, as he does, that in Gettier cases there is no appeal to intuition that is then used as evidence, it would have to be clear in what *sense* intuition is not appealed to and used as evidence. Is it *sui generis* state, inclination to believe or simply belief or all of the above? So, if he claims that those things some philosophers refer to as ‘intuitions’ are nowhere to be found in original texts, it has to be that he is implicitly assuming *some* theory of intuitions, or at least some characterization of them. And this I find to be one of the methodological weaknesses of Deutsch’s strategy. In all honesty, the attempt to pro-

vide an account of intuitions in order to show that philosophers are not using intuitions would not have better standing. It would be seriously undermined since there is no agreement of what intuitions are, so that would not be helpful either.

II.

For now I will set aside this methodological worry and explore the way Deutsch is arguing for his main thesis in the book, i.e. the evidence claim about intuitions. This is the central task of the chapter 2.

(EC) Many philosophical arguments treat intuitions as evidence.

According to Deutsch, the reason for this misconstrual of the philosophical methodology and the view that philosophers appeal to intuitions as evidence lies in the ambiguity of the term 'intuition' in (EC). To clarify this ambiguity he is advocating the distinction that corresponds to Lycan's (1988) *intuitings/intuiteds* distinction, which he formulates in the following way:

(EC1) Many philosophical arguments treat the fact that certain contents are intuitive as evidence for those very contents.

(EC2) Many philosophical arguments treat the contents of certain intuitions as evidence for or against other contents.

The difference consists in the following: either it is the *state* of having an intuition or the *content* of the intuition that is doing the justification of some proposition. Deutsch is claiming that a prevailing number of philosophers who are defending (EC) are defending it in (EC1) state-sense, while his stand is that the only sense in which (EC) can be true, is (EC2) content-sense.

When I deny that philosophical arguments treat intuitions as evidence, I mean to deny (EC1), not (EC2). According to me, very few philosophical arguments treat the fact that p is intuitive as evidence for p itself. (38)

In other words, it can be asserted that philosophers rely on intuitions as evidence, if it means that the content of the intuition is used as evidence, not the fact that one finds something intuitive. The step from this claim to the rejection of the xphi's results is very clear. Xphi mishit the sense in which it is taken that philosophers appeal to intuitions as evidence. Hence, their criticism does not affect philosophical method and in this misconception lies the myth of the intuitive, concludes Deutsch. He goes on and argues not only that philosophers who are endorsing (EC1) sense are mistaken when explicitly addressing the question of how philosophy should be done, they are also mistaken in characterizing their own methods. Now, Deutsch rightly emphasizes the vagueness among advocates of the method of appealing to intuitions regarding the sense of (EC) they are using. Some philosophers are not clear on that matter. And since this is an empirical question, i.e. whether philosophers use (EC1) or (EC2) sense of the evidence claim, we will

take a closer look at the Gettier case and see whether it can be said that it is a paradigm example of refutation by counterexample or, as some opposition to Deutsch would claim, appealing to intuitions as evidence.

The reader should bear in mind that Gettier cases are somehow specific, according to Deutsch, being an exceptional case where philosophers almost unanimously agreed that standard definition of knowledge as justified true belief is false (this too is an empirical question). Important question that Deutsch is addressing is how Gettier argues against the standard JTB theory of knowledge:

for every subject, S, and every proposition, p, if S justifiably and truly believes that p, then S knows that p.

Deutsch's answer is by "presenting (alleged) counterexamples in the form of hypothetical cases, to the generalization" (42). In other words, Gettier did not use or appealed to intuitions in (EC1) sense as evidence against the JTB theory of knowledge in his famous cases. Instead, "Gettier refuted the JTB theory, if he did (...) by presenting counterexamples, full stop. Whether these counterexamples are intuitive for anyone is a separate, and purely psychological, matter" (46). Deutsch further develops his 'arguments instead of intuitions' view by introducing the condition that counterexample has to fulfill in order to be regarded as successfully refuting some theory. Since, obviously, not any counterexample will do, the condition of *genuineness* of the counterexample has to be satisfied. So, the real question that we should be asking ourselves is not whether the counterexamples are intuitive, but rather are Gettier cases genuine counterexamples. Only the latter matters in settling the issue of refutation of the JTB theory of knowledge.

Deutsch's main argument in support of his 'arguments instead of intuitions' view consist of the two following thesis: (i) there is "*no mention of intuitions or intuitiveness of any proposition* in Gettier's presentation" (43)—the 'lack of explicit terminology' thesis as I will call it—and (ii) Gettier refuted JTB theory because his counterexample are genuine—the genuineness of counterexample' thesis. Since both theses require careful reading and precise analysis of original Gettier case, here is the crucial paragraph from his 1963 paper.

But it is equally clear that Smith does not *know* that (e) is true; for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job. (Gettier 1963: 122)

Regarding the 'lack of explicit terminology' thesis, it is true, as Deutsch remarks, that Gettier does not explicitly use the term 'intuition' or its cognates alongside his conclusion that 'it is equally clear that Smith does not *know*' and so makes no explicit appeal to the premise of the form "It is intuitive that there is an F that is not G" (45). Additionally, Deutsch claims that being 'obvious' or 'clear', terms Gettier does use, is different from being intuitive. Those terms usually presuppose

that something is true, not that they are evidence for the truth of some claim. Nevertheless, he admits that the lack of explicit terminology is not conclusive evidence to debunk the myth of the intuitive. However, he then shifts the burden of proof to the opposition to provide the evidence that Gettier does appeal to intuitions in his cases. I do not find this move to be very pervasive, for he must provide some rationale behind this shift of the burden of proof. Not only that in this context the lack of explicit terminology is not conclusive evidence in favor of his claim, it does not contain any reason why the opposition should bear the burden of proof at this point. He has a long tradition of analytic philosophers who, rightly or wrongly, beg to differ so the reason to shift the burden of proof must be more substantiated. It is not like traditional philosophers were not aware that Gettier did not use the term 'intuition' explicitly. They did, and nevertheless continued to argue that it was the intuition about cases that refuted the JTB theory of knowledge. So, in order to reverse the dialectical situation and shift the burden of proof, Deutsch must present some new reason to do so. Moreover, just because one is not explicitly saying "it is intuitive that there is an F that is not G" in order to be qualified as using intuition as evidence, it does not follow that one is not using it implicitly. Unfortunately, Deutsch does not discuss this possibility in any detail.

III.

Deutsch is devoting a substantial amount of space to account for the second thesis, namely to develop an account of how Gettier refuted traditional JTB theory of knowledge, i.e. what makes Gettier cases and all similar thought experiments genuine counterexamples. Deutsch's opposition would address this matter by appealing to the intuitiveness of Gettier's counterexample, arguing that intuitions provide evidence for the refutation of the JTB theory of knowledge. Deutsch thinks this view is mistaken and argues that Gettier presented an argument of why his Smith character does not know. So, Deutsch's answer to this "evidence-for-the-evidence" question, to use his own words, is *arguments*. The conclusion of the Gettier argument is stated first: "it is equally clear that Smith does not *know*" (Gettier 1963: 122), and premises are presented after the semicolon, "for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job" (122).

At this point, Deutsch is presenting Jennifer Nagel's (2012) view, as the representative example of the opposition to the claim that Gettier presented explicit arguments in his cases. She argues that Gettier did not offer any account of knowledge in terms of necessary conditions (including one that would exclude justified belief that is luckily true), which Smith fails to satisfy, and so he is not explicitly stating why

Smith does not know that (e) is true. It is important in this discussion to emphasize that this is not 'either-or' choice. Namely, not all philosophers would argue that Gettier cases are only about intuitions as evidence, and not at all about arguments, as Deutsch seems to indicate in several places. For instance, Malmgren (2011) explains the way that argument is based on intuitive judgment in the Gettier case:

Let the 'Gettier judgement' be the intuitive judgement that I (and many with me) would make about this case, if asked the appropriate question—a judgement that we might express by saying: 'Smith has a justified true belief, but does not know, that someone in his office owns a Ford.' And let the 'Gettier inference' be the inference by which we get from this judgement to the belief that the target theory—the theory that knowledge is justified true belief—is false. (272)

Now, both Deutsch and Nagel's views are results of careful reading, word by word, and analysis of the original text. And yet, they cannot agree on whether Gettier appeals to intuitions or to arguments as evidence against the JTB theory. Nonetheless, they both agree that neither is presented in explicit way. How can we solve this dispute? Since it is not explicitly obvious that Gettier presented only an argument, and it is not explicitly obvious that he appealed to intuition, is it possible that this matter comes down to what seems intuitive to whom? It seems to me that it does, although this is not something Deutsch would agree on. In other words, this dispute is a matter of whether it seems intuitive that Gettier presented an intuitive counterexample, or it seems intuitive that he presented only an argument.

This is something along the lines of what Deutsch considers as a possible problem for his own account, namely the possibility that arguing for his 'arguments instead of intuitions' view as evidence in hypothetical examples, simply delays the question of the real 'evidence-for-the-evidence', and that carrying out reasons or arguments has to stop at some point. And at some point intuitions would enter anyway as regress stoppers at the end of evidential chain. So, even if the Gettier's counterexamples are not presented in terms of intuition as evidence, at some point the end of the chain of evidence for why counterexamples refuted traditional JTB theory of knowledge, or why they are genuine, lies in intuition. Deutsch recognizes this as the relocation problem and devotes chapter 3, 4 and 5 to account for it.

Deutsch rejects the proposed possibility and claims that it is never about intuitions but, rather, about *more arguments*. If Deutsch's opposition would insist that "it cannot be arguments all the way down" (122), and that, as the answer to the 'evidence-for-the-evidence' question, intuitions must come in at some point, Deutsch replies that "arguments [are] further down than the myth of the intuitive would have us believe" (123) and that evidential levels or chains of reasons are finite, as well as philosophical texts, and at some point must come to an end. That is why every argument takes at least one premise for granted—which Deutsch calls *philosophical starting points*—and im-

portant methodological note is that they “need have no special phenomenological or epistemological features” (124). In a nutshell, for Deutsch, regress stoppers are not intuitions viewed as ‘rock bottom’ evidence, but rather philosophical starting points, which are taken for granted, are not unified, and vary among philosophers.

I find this argument somehow problematic since he seems to be advancing the double standard of what qualifies as regress stopper. First he accounts for philosophical starting points and claim that “nothing unifies the claims that get taken for granted”, that “different philosophers have different starting points, and the starting points are as heterogeneous as can be (124)”. But later argues that “‘judgments about philosophical cases’ [i.e. intuitions about cases] names too heterogeneous a class for every judgment in the class to qualify as foundational in the sense required by foundationalist solutions to the regress problem” (127). In other words intuitions are too heterogeneous to be regarded as rock bottom evidence. But on the other hand, vaguely described ‘philosophical starting points’, which are a matter of choice for philosophers according to Deutsch, and are also not unified in any substantial way beside the fact that they are the “un-argued-for premises in a philosophical argument” which “need have no special phenomenological or epistemological features”, can count as regress stopper (124). As I see it, either there is no difference between intuitions (or intuitive judgments) and Deutsch’s description of philosophical starting points concerning this matter, or the difference between the two is very sophisticated.

Deutsch analyses the possibility of intuitions being regress stopper via foundationalist criterion of what qualifies something to be a foundational judgment. In this regard he considers two possibilities, basic perceptual and self-verifying judgments, and dismisses the possibility that judgments about cases, i.e. intuitive judgments, could qualify as either of the two. First, I am puzzled as to why Deutsch dismisses the possibility of intuitive judgments being self-verifying judgments without any further explanation. Or the way that perceptual judgments are not heterogeneous in a sense that intuitions are. Second, Deutsch argues that intuitions cannot be unified on the ground of being spontaneous or noninferential judgments, but gives somewhat dubitable explanation of why this is to. Namely, Deutsch thinks that intuitive judgments might seem as nonreflective, or spontaneous, or noninferential because we are taking the wrong perspective on thought experiments. The inventor of any given thought experiment “took a considerable amount of ingenuity, careful thought, and inference” (98) to arrive at the judgments which are then often described as intuitive. This claim seems to be controversial on several levels, but I will focus just on one of them. I think it is wrong in this context to assume the correctness of the first person perspective, because the amount of work and careful thought one invested in constructing the thought experiments is beside the point. What is relevant here is whether such thought experiments

elicit nonreflective and noninferential judgments, which are then used as evidence. Whether thought experiments elicit intuitions and the amount of work philosopher has to do in order to construct them are two separate questions. And the amount of careful thought and effort that is putted in their construction does not say anything of whether they elicit intuitions. And to additionally claim that Gettier himself, for instance, did not intuit that Smith character does not know requires some empirical confirmation, which Deutsch does not provide.

Third, it is doubtful whether Deutsch's account of philosophical starting points would pass this foundational criterion that he imposes on intuitive judgments. And if intuitions, as heterogeneous group, must pass such criterion, so should Deutsch's heterogeneous group of philosophical starting points. It seems that Deutsch is willing to accept un-argued-for premises in philosophical arguments, and if he does not explain why the latter is acceptable while the former is not, I do not see any substantial difference to justify his rejection of intuitions. Moreover, it does not seem plausible to maintain that these un-argued-for premises in philosophical arguments need not to have epistemological features, as Deutsch argues. The chain of epistemic reasons of the given argument end in those premises, so they certainly have to have some epistemic merits. My point is that if intuitions are too heterogeneous group and cannot be unified in a way to qualify as evidential starting points, the same should apply to the Deutsch's proposal of philosophical starting points, which are also heterogeneous group. The difference should be elaborated in more details, especially since he does not discuss the way that intuitions are heterogeneous—which can be ascribed to the fact that he does not have a theory of intuitions. It could be that Deutsch is not evaluating philosophical starting points via foundational criterion, but instead appeals to the possibility of coherentist solution to the regress problem. He is proposing solution to the relocation problem in term of hypothetical claim:

(...) if some form of coherentism about inferential justification is true, then it is something other than resting on rock bottom evidence that justifies our inferences. Some premises are justified not by inference from further premises but instead by their coherence with other premises—if coherentism is true, that is. (127)

Deutsch is maintaining that if coherentism is true, then the demand of foundational justification could be avoided. Unfortunately, Deutsch is not arguing that coherentism is true, nor is he providing any reason why we should accept his coherentist solution rather than foundational one, so his 'arguments instead of intuitions' account of thought experiments is still facing the relocation problem. Furthermore, Deutsch is trying to make it immune from problems regarding the truth-irrelevant factors that (supposedly) affect philosophical intuitions.

Truth-irrelevant variability in the intuition that *p*, where this is understood as variability in whether different groups of people have or lack the intu-

ition that p , will not matter in the slightest. If there is a cogent and compelling argument for p , then p may perfectly well be regarded as true and taken as evidence for or against the truth of other, related propositions. (75)

If we agree with Deutsch that in thought experiments philosophers are not using intuitions, but rather arguments as evidence for p , then it follows that philosophers are not very proficient in argumentation, or that they simply refuse to accept a good argument for the truth of p . If cogency and compellingness of argument for p is all that is needed for p to count as true, then either we have very few such arguments—which would be very unfortunate since philosophers do that for a living—or there is something else that prevents philosophers to accept something to be the evidence for or against some theory. For if a philosopher sets forth an argument for p , and given the fact that philosophers do not agree about much else beside the fact that traditional JTB definition of knowledge is false, then our ability to construe a good argument is very poor. Of course, pervasiveness of arguments, or the absence of it, can lie at the philosophical starting points, which are very heterogeneous group, as Deutsch argues. But this would not be an accurate interpretation of argumentative practice among philosophers since more often than not these starting points are taken for granted among the opposition, and philosophers proceed evaluating and rebutting the arguments.

There is one more important thing that should be stressed regarding Deutsch's claim that arguments in thought experiments might elicit intuitive judgments, but that those do not have any evidential strength. Additional arguments, that support those intuitive judgments do.

Judgments about thought experiments can be given argumentative support, even if the judgment is intuitive. Arguments for the truth of some intuitive judgment are arguments that reveal that the content of the judgment may qualify as evidence (...). (75)

No defender of the intuitional methodology would deny that intuitive judgments elicited by thought experiments are not often reinforced by supplementary arguments. The disagreement is whether the former has any evidential force.

As I see it, we are faced with two horns of a dilemma: either philosophers have different intuitions or they are bad in argumentation. I would argue that the latter is less preferable option. For one thing, variability in intuitions existed in philosophical discussions even before the arrival of the $xphi$ and that did not present any problem. For instance, internalists and externalists regarding the problem of justification in epistemology, just to mention one example, engaged in their exchange of arguments in spite of having different intuitions as starting points. That did not present any problem for the ongoing discussions or exchange of arguments. If any of the thought experiments should count against the externalism and reliabilism, it should have been the Lehrer's Truetemp case (1990) and BonJour's Norman

case (1984). If those thought experiments really are arguments used as evidence for internalism and against reliabilism, as Deutsch is suggesting, why philosophers did not unanimously reject reliabilism as a false theory? Is it because those are not very good arguments or not well construed thought experiments? These would be the only viable options if we accept Deutsch's view that thought experiments are not about intuitions. However, this certainly is not the accurate verdict since no one would argue that those are not good thought experiments. So the plausible explanation of the continuance of the disagreement would be the difference in philosophers' starting intuitions regarding the concept of justification. I am puzzled as to why these variations did not present any problem until xphi conducted surveys which revealed that folks do not share philosophers' intuitions. In other words, philosophers were fully okay with not having same intuitions with each other, but fully concerned about their methodology when realizing that folks are having the same variation.

And although Lehrer's Truetemp case is a good thought experiment that received a lot of philosophical attention and is substantiated with additional arguments against externalism, there are still a vast number of externalists. This is a good indicator that essentially it all comes down to the initial intuitions philosophers have as a starting point. And philosophers are ok with diversity in that respect. But I do not think they would be ok with the other horn of a dilemma, namely that there are no good arguments in philosophy, or that they do not accept a good argument as evidence for p even if they see one.

As we can see, the trouble with intuitions is on both sides of the camp. Epistemological and metaphilosophical accounts of intuitions are, in one way or another, flawed and the level of obscurity and ambiguity in using the term 'intuition' is deeply troubling. Consequently, any attack on intuitions and intuitional methodology stands on equally troubling grounds. It is not enough simply to argue that philosophers are not using intuitions as evidence in their philosophical texts on the ground that one simply does not find any appealing on intuitions in thought experiments, or that philosophers do not explicitly use the term 'intuitive'. What is needed is some plausible empirical analysis of it. And it seems that empirical analysis comes down to intuitions, or what philosophers would say Gettier cases are about, intuitions or arguments.

References

- Bealer G. 1998. "Intuition and the autonomy of philosophy." In DePaul M. and Ramsey W. (eds.). *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. New York: Rowman & Littlefield: 201–240.
- BonJour, L. 1998. *In Defense of Pure Reason*. Cambridge: Cambridge University Press.

- Devitt, M. 2011. "Experimental semantics." *Philosophy and Phenomenological Research* 82 (2): 418–435.
- Gettier, E. L. 1963. "Is justified true belief knowledge?" *Analysis* 23: 121–123.
- Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Clarendon Press.
- Kornblith, H. 2007. "Naturalism and intuitions." *Grazer Philosophische Studien* 74: 27–49.
- Lehrer, K. 1990. *Theory of Knowledge*. Boulder: Westview Press.
- Lewis, D. 1983. *Philosophical Papers, vol. 1*. Oxford: Oxford University Press.
- Lycan, W. G. 1988. *Judgment and Justification*. Cambridge: Cambridge University Press.
- Malmgren, A. 2011. "Rationalism and the Content of Intuitive Judgments." *Mind* 120 (478): 263–327.
- Nagel, J. 2012. "Intuitions and experiments: A defense of the case method in epistemology." *Philosophy and Phenomenological Research* 85 (3): 495–527.
- Weinberg, J. M. et al. 2010. "Are philosophers expert intuiters?" *Philosophical Psychology* 23 (3): 331–355.

“The Brain in Vat” at the Intersection

DANILO ŠUSTER

University of Maribor, Maribor, Slovenia

Goldberg 2016 is a collection of papers dedicated to Putnam’s (1981) brain in a vat (‘BIV’) scenario. The collection divides into three parts, though the issues are inter-connected. Putnam uses conceptual tools from philosophy of language in order to establish theses in epistemology and metaphysics. Putnam’s BIV is considered a contemporary version of Descartes’s skeptical argument of the Evil Genius, but I argue that deception (the possibility of having massively false belief) is not essential, externalism does all the anti-skeptical work. The largest section in the collection covers Putnam’s model-theoretic argument (MTA) against metaphysical realism (MR) and its connections with the brain in vat argument (BVA). There are two camps—unifiers (there is a deep connection in Putnam’s thoughts on BVA, MTA and MR) and patchwork theorists and I try to provide some support for the second camp. All of the papers in the collection are discussed and the anti-skeptical potential of BVA is critically assessed.

Keywords: Putnam, brain-in-a-vat scenario, skepticism, realism, model-theoretic argument.

It is not easy to track the provenance of the *brain in a vat* (‘BIV’ for short) scenario. The contemporary empirical source seems to be the work of Canadian neurosurgeon Wilder Graves Penfield on neural stimulations (in the 1930s) and experiments in waking human subjects undergoing epilepsy surgery. Penfield observed quite complex memories being switched on by electrical stimulation of the appropriate parts of the cerebral cortex (Tallis 2011: 36). Its philosophical use is (first?) registered in the work of Gilbert Harman (1973)—a playful brain surgeon might be giving you “normal” experiences by stimulating your cortex in a special way, but in reality “you might really be stretched out on a table in his laboratory with wires running into your head from a large computer. Perhaps you have always been on that table. ... Or perhaps you do not even have a body. Maybe you were in an accident and all that could be saved was your brain, which is kept alive in the laboratory” (Harman 1973: 5). This type of scenario leads to

familiar philosophical problems of other minds and the external world skepticism, evoked, famously by Descartes. Recall: "... some evil spirit, supremely powerful and cunning, has devoted all his efforts to deceiving me. ... What truth then is left? Perhaps this alone, that nothing is certain" (Descartes 2008: 16).

Nowadays the scenario is almost automatically associated with Hilary Putnam (the first chapter of his 1981). An entire new collection (Goldberg 2016 in the series on *Classic Philosophical Arguments*) is now dedicated solely to philosophical applications and ramifications of the version proposed by Putnam. Descartes is still in the background, thus Goldberg in *Introduction* (2016: 2) "Putnam's reflections on the BIV scenario have a familiar historical precedent, of course, in Descartes's reflections on the Evil Demon scenario." The connection with the Cartesian deceiver is not entirely accurate and I find the proper role of deception to be controversial. Putnam actually writes: "Perhaps there is *no* evil scientist, perhaps (though this is absurd) the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems" (Putnam 1981: 6). In Putnam's BIV world everyone is raised as brains in vats, but their perceptual input is qualitatively just like ours. Could this be our predicament? Putnam argues from some plausible assumptions about the nature of reference to the conclusion that it is *not* possible that all sentient creatures are brains in a vat. A deceptively simple and enormously influential argument ('BVA' for short) in various fields of philosophy. The collection divides into three parts, though the issues are inter-connected. Putnam uses conceptual tools from philosophy of language in order to establish theses in epistemology and metaphysics.

The first part, "Intentionality and the philosophy of mind and language" opens with an essay by Anthony Brueckner, one of the earliest commentators who wrote several papers on the argument. His seminal paper reconstructed the argument in terms of a disjunctive dilemma suggested by Putnam (Brueckner 1986: 154; more or less reproduced by Pritchard and Ranalli in Goldberg 2016: 78):

- (1) Either I am a BIV (speaking vat-English) or I am a non-BIV (speaking English).
- (2) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.
- (3) If I am a BIV (speaking vat-English), then I do not have sense impressions as of being a BIV.
- (4) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are false. [(2), (3)]
- (5) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are true iff I am a BIV.
- (6) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are false. [(5)]
- (7) My utterances of 'I am a BIV' are false. [(1), (4), (6)]

Whatever proposition is expressed by my utterances of ‘I am a BIV’ is a false proposition. The anti-skeptical conclusion seems to be that I therefore know that I am not a BIV. The argument is based on an analysis of the truth conditions for the sentences uttered (or thought) by a BIV. These conditions depend on the assignments of references which one would make in evaluating the truth value of BIV’s utterances. According to semantic *externalism* when S uses a referring term, she refers to whatever typically causes her uses of that term (in the case of BIV—sense impressions as of being a BIV, according to Brueckner and many other commentators, but not real “brains” and “vats”).

The exact role and type of *externalism* used in the argument has been disputed, however. *Kallestrup* (Goldberg 2016: 53) argues that the causal constraint on reference needed in Putnam’s proof is actually quite weak and consistent with semantic internalism: “semantic externalists are no better placed than semantic internalists in terms of being able to appeal to Putnam’s proof as a semantic response to epistemological skepticism.” *Grundmann* (Goldberg 2016: 90–110) on the other hand compares the New Evil Demon (NED) intuition—one can have justified beliefs about the world even if one is living in a demon world with the Old Evil Demon (OED) intuition (BIV, dream). According to the latter one cannot possess justified beliefs about the world unless one is able to rule out relevant skeptical hypotheses. There was always a strong tendency to regard the NED intuition as evidence for the internalism, but Grundmann argues that the NED intuition does not provide a compelling argument for mentalism but is in fact compatible with the view that justification requires reliability. The BVA assumes the view that the individuation conditions of mental content depend, in part, on external or relational properties of the subject’s environment. If these connections are constructed reliabilistically and reliability is a necessary condition for justification this would vindicate the crucial role of externalism in Putnam’s argument, or so it seems.

An interesting new development in this area is explored by *Bernecker* (Goldberg 2016: 54–72). Whereas content externalism locates mental states inside the head or body of an individual, the hypothesis of *extended mind* claims that the role of the physical or social environment is not restricted to the determination of mental content. Mental states are not only externally individuated but also externally located states. Just as the brain in a vat forms a coupled system with the supercomputer that feeds it all of its sensory-input signals, the supercomputer forms a coupled system with the evil scientist who programs it (Goldberg, ed. 2016: 64). But the scientist presumably speaks a “normally” referring language, and since the brain in a vat should count as an extension of the evil scientist’s mind it too, can, after all refer to trees and vats and so on. When content externalism is combined with the extended mind hypothesis it is robbed of its anti-skeptical power according to Bernecker.

The topic of externalism, self-knowledge and reliabilism in the form of sensitivity principle is also discussed by *Becker* (Goldberg 2016: 111–127). The crucial belief “I am a not BIV” is *sensitive* (and thus fulfills a necessary condition for knowledge), for if it were false I would not believe that I am. “I would have some other belief, such as that I am not some specific state type of some particular automated machinery” (Goldberg 2016: 116). But unless I *know* that my terms are referring and my thoughts are about brains and vats, I don’t know whether the belief that I express by ‘I am not a BIV’ is that I am not a BIV. The appeal to sensitivity has not explained how I could know that the skeptical hypothesis is false. Becker’s result is largely negative—sensitivity adds nothing to the standard view and standard discussion.

Standard discussion views the BIV scenario primarily as a vehicle for Cartesian angst (cf. *Button* in Goldberg 2016: 142). The worry that it generates is that appearances might be radically *deceptive*, so that (almost) all of our beliefs are *false*. In particular, my utterances of ‘I am a BIV’ are false if I am a BIV (speaking vat-English), according to Brueckner (recall step 4 in the disjunctive argument above). The vat-English truth conditions of ‘I am a BIV’ are not satisfied because of *deception* (I am not fed experiences about my “reality”, representing me to be a disembodied BIV). But I think that deception, implying *false* beliefs, is, strictly speaking, not essential at all. On the assumption of externalism BIVs lack conceptual resources to even think about the reality of their situation. The Evil Demon scenario has undergone an important historical transformation.

We should follow the suggestion by Mišćević (Mišćević 2016) and explore the diachronic developments in a long-term life of a thought experiment. The BIV scenario lies at the intersection of “trails” of two thought experiments, the Cartesian Evil Demon scenario and Putnam’s *Twin Earth* scenario (Oscar on the Twin Earth, not being in causal contact with Earthly H₂O, does not refer to water). Deception is of course crucial in the Cartesian scenario, but when the two scenarios are combined all the anti-skeptical work is done by semantic externalism—in order for our words to refer to a particular kind of thing, it is necessary for our uses of the term to be connected in an appropriate way with things of that kind. Recall Putnam’s initial analogy: an ant is crawling on a patch of sand and as it crawls, it traces a line in the sand which ends up looking like a caricature of Winston Churchill (Putnam 1981: 1). The Putnamian intuition is that the caricature does not refer to or represent Churchill, because the *presuppositions* of successful reference are not fulfilled. This suggests that the main problem with BIV mental states is not a cruel deception, but lack of proper connection.

Suppose we take seriously the parenthetical part of Putnam’s own comment (Putnam 1981: 15): “the sentence ‘we are brains-in-a-vat’ says something false (if it says anything).” We should then reconsider the anti-skeptical argument not on the assumption that “We are not brains in a vat” is false, rather, the preconditions for its being true or false

are not fulfilled (I try to do this in Šuster 2016). To repeat, I think that Putnam’s externalism is the basis of his reply to BIV skepticism: no false beliefs because no real beliefs (thoughts) at all (and not because some demonic machinery is feeding us *false* impressions). Still, a vast majority of authors in the collection take the crucial role of massively false beliefs for granted (with *Folina* as an exception).

I will return to the assessment of the Putnam-style refutation of radical skepticism later (Part II: “Epistemology”). Let me jump to the third and the largest section, “Metaphysics”, covering Putnam’s model-theoretic argument (*MTA*) against metaphysical realism (*MR*) and its connections with the brain in vat argument (*BVA*). It is a vexed issue how to reconstruct interrelations between *MTA*, *BVA* and *MR*. Even Putnam himself is (characteristically) ambiguous. According to his own report (Putnam 1992: 362):

I gave a seminar at Princeton in the late seventies at which I presented and defended my model-theoretic arguments. David Lewis, who was present, commented that “there must be something wrong somewhere”—because, if my arguments were right, it followed that we could not be brains in a vat!

So there is a direct connection between the BIV scenario and the model-theoretic argument, *MTA* implies *BVA*? But there are other reports, for instance by Brueckner, who thinks that *BVA* should be sharply distinguished from the model-theoretic argument against metaphysical realism (1986: 149, footnote 2):

Putnam has indicated (in conversation) that it was in fact his intention to construct an argument in chapter 1 [of Putnam 1981, i.e. *BVA*, D.Š.] quite different from the model-theoretic argument of the later chapters.

Guyer (1992: 100) noticed that some commentators are committed to the assumption that the views of a great philosopher like Kant must possess a profound unity that can be brought out by a sympathetic interpretation. A different interpretation is defended by Guyer himself and the so called “patchwork” theorists: Kant’s greatness lies more in some of his particular analyses and arguments and in his recognition of the complexity of the connections among them than in his pretensions to systematicity. I think that something similar is true of Putnam and his interpreters. *Button* and *Sundell* belong to the camp of *unifiers*, *Sher* is clearly a *patchwork* theorist, *Douven* and *Marino* are less explicit, but probably also accept just a juxtaposition, not an amalgamation of *BVA* and *MTA*.

Let me start with Putnam himself. The first chapter of *Reason, Truth and History* is dedicated to the BIV scenario, and model theoretic results are briefly mentioned (Putnam 1981: 7), when he says about the *BVA* argument: “It first occurred to me when I was thinking about a theorem in modern logic, the ‘Skolem-Löwenheim Theorem’, and I suddenly saw a connection between this theorem and some arguments in Wittgenstein’s *Philosophical Investigations*.” The prime locus of Wittgensteinian themes seems to be the private language argument: mental representations are not *magically* connected with what they

represent. On the other hand, when discussing the problem of (anti) realism later in the book, the possibility of a BIV scenario is one of the dividing issues between the camps. According to the perspective of metaphysical realism:

... the world consists of some fixed totality of mind-independent objects. There is exactly one true and complete description of ‘the way the world is’. Truth involves some sort of correspondence relation between words or thought-signs and external things and sets of things. I shall call this perspective the externalist perspective, because its favorite point of view is a God’s Eye point of view (Putnam 1981: 49).

On the internalist perspective, defended by Putnam, the question of what objects does the world consist of is a question that it only makes sense to ask within a theory or description. ‘Truth’, in an internalist view, is some sort of (idealized) rational acceptability. A ‘Brain in a Vat World’ is then only a *story* and not a possible world at all (Putnam 1981: 50):

For from whose point of view is the story being told? Evidently not from the point of view of any of the sentient creatures in the world. Nor from the point of view of any observer in another world who interacts with this world; for a ‘world’ by definition includes everything that interacts in any way with the things it contains. ... So the supposition that there could be a world in which all sentient beings are Brains in a Vat presupposes from the outset a God’s Eye view of truth, or, more accurately, a No Eye view of truth — truth as independent of observers altogether.

For a metaphysical realist the truth of a theory consists in its corresponding to the world as it is *in itself*, so the BIV scenario cannot be dismissed. This establishes an elegant connection between MR and BIV in the form of *modus tollens*, in the version of *Sundell* (Goldberg 2016: 229):

- 1) If metaphysical realism is true, then pervasive error is a coherent possibility.
- 2) But pervasive error is not a coherent possibility.
- 3) So metaphysical realism is false.

The first premise is based on the non-epistemic notion of truth inherent to MR: even an empirically adequate theory—a theory that is predictively accurate and that satisfies any theoretical virtue one may like—may still be false (cf. *Douven* in Goldberg 2016: 175). In Putnam’s *earlier* writings the BIV scenario sometimes really figured as an illustration of the possibility of pervasive error. According to MR (Putnam 1977: 485)

THE WORLD is supposed to be independent of any particular representation we have of it—indeed, it is held that we might be unable to represent THE WORLD correctly at all (e.g., we might all be “brains in a vat”, the metaphysical realist tells us).

The most important consequence of metaphysical realism is that truth is supposed to be radically non-epistemic—we might be “brains in a vat” and so the theory that is “ideal” from the point of view of operational utility,

inner beauty and elegance, “plausibility”, simplicity, “conservatism”, etc., might be false.

But Putnam (1981) does not justify premise (2) above with the *impossibility* of BIV demonstrated by BVA. The main work of justifying the impossibility of pervasive error is done by MTA, an epistemically ideal theory is guaranteed to be true, according to Putnam. As noted by *Sundell*:

For an ideal theory to be false, it must be the case that the theory fails to correspond to what the world is like on *the correct interpretation of that theory*. But the MTA shows that there is no way to privilege such an interpretation as correct. The theory is guaranteed to be true on some interpretation, and nothing from inside or outside of the theory can show that that interpretation is the wrong one (Goldberg 2016: 229).

But what I find much more doubtful is that for Sundell “... the anti-realist application of the BVA is the same as the anti-realist application of the MTA. Both arguments attack the coherence of pervasive error” (Goldberg 2016: 234). Putnam’s aim in his 1981 was to refute three “solutions” to the puzzle of what it is that determines reference and metaphysical realism is not the main target (cf. De Gaynesford 2011: 579). The main problem is the relation of correspondence on which truth and reference depend for MR. Putnam argues that MR cannot offer a satisfactory account of *determinate* referential relations between the words and the things. If one is in BIV the relation of independent correspondence characteristic for MR is not available, so, given MR commitments, the scenario is paradoxical, a puzzler (Putnam 1981: 51). As he notes in his earlier writings, “Suppose we (and all other sentient beings) are and always were “brains in a vat”. Then how does it come about that our word ‘vat’ refers to *noumenal* vats and not to vats in the image?” (Putnam 1977: 487).

We can agree with *Sher* (Goldberg 2016: 208) that the MTA argument shows that (i) we cannot theoretically determine the reference of our words, and that, as a result, (ii) we must renounce the correspondence theory of truth and robust realism. The BVA argument, on the other hand, shows, that (iii) we cannot truly believe that we are BIVs, and that (iv) Cartesian skepticism is thus undermined. MTA is the main weapon against MR and BVA seems to be a different, *juxtaposed* issue. Sher is also critical with respect to Putnam’s results—she thinks that the meta-logical considerations that lead Putnam to conclude (i) are irrelevant to a robust realist/correspondence account of reference (I tend to agree).

The other two “patchwork” theorists, *Douven* and *Marino*, do not have much to say about BVA, but they are also critical with respect to the prospects of MTA. According to *Douven* MTA against realism is based on two assumptions:

- (CT) Truth is a matter of correspondence to the facts.
- (SN) Semantics is an empirical science like any other.

At the time when the MTA was conceived, it was common to think that a semantics could not be scientifically acceptable if its key concepts could not be accounted for in strictly physicalist terms. But (CT) is no longer the only game in the town, specialists working on truth are nowadays more inclined toward some version of *deflationism*. Douven argues, convincingly, that semantics can be pursued in a scientific spirit without necessarily being part of a reductionist–physicalist research program. Thus MTA is no longer supported (Goldberg, ed. 2016: 189).

Marino discusses the question how does the model-theoretic argument look from the point of view of contemporary *naturalism*. She also stresses that naturalistic forms of disquotationalism diverge from or challenge Putnam's own understanding of reference and truth. Her prime example of a contemporary naturalistic philosopher is "the Second Philosopher", from Maddy (2007). The whole idea of metaphysical realism is somehow misguided from the perspective of modern naturalism and, at least from the contemporary perspective, Putnam seems to be fighting a straw man:

... the rejection of metaphysical realism seems significant to Putnam only because of his desire for an account that will, from outside the use of our methods, support and justify those methods—a desire the Second Philosopher does not share (*Marino* in Goldberg 2016: 200).

On the other pole of interpretation the main defender of unification is *Button*. He sees a deep connection between Putnam's thoughts on BIVs, on Skolem's Paradox, and on permutations (also called the "cats and cherries" argument from the *Appendix* in Putnam 1981: 217–218). The last two are based on model-theoretical results but Button unites them all in the form of the BIV-style argument. All types of skepticism—permutation-skepticism (the worry is that our words do not refer as they are intuitively supposed to), skolemism (the worry here is that we cannot tell whether there really are uncountable sets, or merely seem to be¹) and BIV skepticism are self-refuting when considered as types of *internal* skepticism. Internal skepticism is based on assumptions which we ourselves hold, the skeptic raises an antinomy from within our own worldview. The lynchpin of all of the anti-skeptical arguments is self-refutation, if the skeptical scenario were actual, then we would be unable to articulate this (Goldberg 2016: 153).

Button elegantly develops the template in the form of the BIV-style argument, where the core principle is the principle of *disquotation*. According to Brueckner's original assessment (cf. *Pritchard* and *Ranalli* in Goldberg 2016: 78) one can get the proper anti-skeptical conclusion

¹ Let me note a disturbing typo, the argument against the skolemist is stated as (Goldberg 2016: 143):

(1S) A smallworlder's word 'countable' applies only to countable (H) sets.

(2S) My word 'countable' applies only to countable (H) sets.

(3S) So: I am not a smallworlder.

But surely, (2S) should be "My word 'countable' does not apply only to countable (H) sets."

“It is *not* the case that I am a BIV” from “My utterances of ‘I am a BIV’ are false” only with the help of the additional *disquotation* principle:

(T) My utterances of ‘I am a BIV’ are true iff I am a BIV.

This looks question-begging. I am entitled to (T) only if I am entitled to assume that I am a normal human being speaking English rather than a BIV speaking referentially defective vat-English. Since I do not know whether I am speaking English or vat-English, I do not know whether the truth conditions of my utterances of ‘I am a BIV’ are disquotational ones or not. Still, *Button*, *Ebbs*, *Sundell* and in this collection also *Brueckner* (Goldberg 2016: 21–22), they all defend our knowledge of the semantics of our own language (i.e. our language disquotes and we are entitled to (T)). According to *Ebbs* (Goldberg 2016: 27–36) the goal of the argument is not to show, by strictly a priori methods, that we are *not* always brains in vats. Rather, we always start “relying on already established beliefs and inferences, and applying our best methods for re-evaluating particular beliefs and inferences and arriving at new ones” (Goldberg 2016: 31). The point of the BVA is to transform our understanding of the statement that we are not always brains in vats. If we presuppose substantive beliefs that suffice for minimal competence in the use of the words, we may infer that the disquotational premise (T) is true.

This is still very cautious. In the opening article of the collection *Brueckner* now *defends* Putnam’s reasoning in the form of the Simple Argument (Goldberg 2016: 21–22):

- (1) If I am a BIV, then my tokens of ‘tree’ do not refer to trees.
- (2) My word ‘tree’ refers to trees.
- (3) So, I am not a BIV.

How does he refute his own earlier criticism? How is (2) justified? *Brueckner* now claims that whichever language is the one that I am speaking (English or vat-English), my language disquotes. This is licensed by my knowledge of the semantics of my own language (Goldberg 2016: 24).

Button is the most resolute of the three—for him the falsity of disquotation is genuinely *unrepresentable*. He considers the following version of BVA (Goldberg 2016: 135):

- (1B) A BIV’s word ‘brain’ does not refer to brains.
- (2B) My word ‘brain’ refers to brains.
- (3B) So: I am not a BIV.

Premise (1B) is justified by semantic externalism and premise (2B) by defending disquotation in the mother-tongue. To understand, talk or even just present the BIV scenario, we need to rely on disquotation, so the skeptic cannot even raise doubts about (2B)—“premise (2B) is implicitly required by the BIV skeptic herself in the very *formulation* of her skeptical challenge ..., to deny (2B) is self-refuting” (*Button* in Goldberg 2016: 137). Without relying upon disquotation the skeptic cannot even present her worry that everyone is a BIV.

For Button a simple argumentative template, based on self-refutation (as exemplified by the BVA), shows us how to defeat skolemism, permutation-skepticism and BIV skepticism and, in so doing, how to overthrow certain philosophical pictures. The process that unifies MTA and BVA is the following (Goldberg 2016: 153):

- Step 1. Isolate a particular philosophical picture.
- Step 2. Observe that some skeptical challenge is unanswerable, given this picture.
- Step 3. Show that the skepticism in question is actually self-refuting (or reliant on magic).
- Step 4. Conclude by rejecting the original picture as incoherent (or reliant on magic).

Let me start by noting that this unifying process is very *general*, it could easily fit, for instance, Berkeley's critique of materialism as a particular philosophical picture (given materialism the skeptical challenge is unanswerable, but skepticism is self-refuting, because in order to conceive of mind-independent objects, we must ourselves be conceiving of them.) A road to *idealism* as is often suspected by Devitt in his comments on Putnam? Not necessarily, the process could perhaps also fit some of Wittgenstein's strategies, the point is, rather, that there need not be any *specific* unity in Putnam's discussions of brains in vats, of Skolem's paradox, and of cats and cherries (that all and only those three arguments fit the procedure diagnosed by Button). My sympathies remain with the *patchwork* theorists but as it is clear from the quotes above, in the late seventies there were several lines of thoughts in Putnam's writings, sometimes separate, sometimes intersecting and Button does a great job in his attempt to provide a unified picture (also in his very elegant and "user-friendly" presentation of skolemism and the permutation argument).

Next, is it really impossible to make sense of the statement that we are not always brains in vats being false? It seems to me that our knowledge of semantic features (disquotation) of our own language cannot be *a priori*—it is an established semantic fact that even in plain vernacular English containing empty names (and perhaps vague expressions) disquotation fails. Suppose we take seriously the idea that sentences uttered by BIVs are neither true nor false, because the preconditions for their having a truth value are not fulfilled. The disquotation scheme for sentences is just the Tarski's schema:

(T) "P" is true if and only if P

If truth-value gaps are admitted, then this principle is no longer valid. Sentences with empty terms ('this dagger' when used by someone under a hallucination), lack the disquotational properties. Yet we still seem to be linguistically competent and possess some level of understanding of our words even if disquotation fails. "Quasi-understanding" perhaps, so that BIV's mental states lacking normal referential properties do not count as real thoughts but "quasi-thoughts" only. Still, BIV's are not

like ants, the scenario makes sense only if they are relevantly similar to us—capable of engaging in cognitive mental activities. In the vat I cannot *really* think “I am a brain in a vat” since I cannot think about real world brains and real world vats. But, as *Folina* (Goldberg 2016: 172) rightly notices, it does not follow that I cannot have thoughts that are epistemically identical to the BIV thought or nearly so. Just recall the discussions about *narrow* content—no matter how different the individual’s environment were, the belief would have the same content it actually does. *Horgan, Tienson and Graham*, for instance, defend the notion of narrow phenomenology—according to Cartesian intuitions, as they name them, one intuitively judges that the BIV’s mental life exactly matches one’s own, the BIV has numerous beliefs, both perceptual and non-perceptual, that exactly match one’s own “normal” beliefs (Horgan et al. 2004: 297). Can we really exclude this possibility on the grounds of self-refutation? Contrary to *Ebbs* I find the worry of the question-begging nature of the BVA quite persuasive (Brueckner 1986: 160, quoted by *Ebbs* in Goldberg 2016: 36):

I can conclude from this [argument] that I am a normal human being rather than a BIV—and thereby lay the skeptical problem to rest—only if I can assume that I mean by “I may be a BIV” what normal human beings mean by it. But I am entitled to that assumption only if I am entitled to assume that I am a normal human being speaking English rather than a BIV speaking vat-English. This must be shown by an anti-skeptical argument, not assumed in advance.

The challenge has now really changed—the original worry was the Cartesian possibility of having massively false beliefs, the “new” skeptical worry is how do we know that our terms refer, that the preconditions of our having real thoughts are fulfilled. Or, in words of *Folina*, our inability to think of or about the exact conditions under which we may be deluded implies that the skeptical thought lacks *specificity*, it does not make it incoherent (Goldberg 2016: 172–173). Similar critical voices are represented by *Pritchard and Ranalli* (Goldberg 2016: 75–89). They provide a list of critiques of the anti-skeptical potential of BVA, ending on a pessimistic note—the BIV hypothesis is simply a template for making vivid what might be our actual epistemic predicament. “*Prima facie* it’s hard to see why some of those possible truths [truths we cannot conceive] are not skeptical, representing our epistemic predicament in ways that we cannot conceive” (Goldberg 2016: 89). And *Sher* adds (Goldberg 2016: 225): “... if there are conditions under which BIVs could figure out some things about the world, are we as different from them as Putnam thinks we are? Is it absolutely irrational to entertain the possibility that we are them, that we are at least a little bit like them?”

Let me try to summarize the problem of the relationships between BVA, MTA and MR from the perspective of the BIV scenario. Skepticism was traditionally a road to anti-realism (a total denial of knowledge is difficult to sustain, so the “reality” cannot be something that transcends our cognitive abilities) and externalism, in general, was

supposed to be realistic in spirit. One would therefore expect the anti-skeptical argument such as BVA to support realism, but Putnam is more subtle. According to his intersecting lines of thinking only *internal* realism can deliver the anti-skeptical goods. Metaphysical realism is always in the grip of the "mind the gap" warning: even a rationally optimal or 'ideal' theory of the world could be mistaken. Putnam argues that this is not possible, but his main weapon against MR is the model-theoretic argument. Metaphysical realism commits itself to claim that uniquely determinate referential relations exist between what we say (and think) and the world, and MTA challenges *this* claim. This suggests that we should interpret the BIV scenario as a *referential* puzzle for MR and not as a way of showing that pervasive error is incoherent and in this way opposing the view that a theory which gives every appearance of being true might really be radically false.

BVA, on the other hand, is primarily an anti-skeptical argument, but a Putnam-style refutation of radical skepticism looks like a small-pox vaccine which prevents the severest and the rarest form of small-pox only. The BVA excludes just those bad scenarios "cooked up to be vulnerable to the semantical reply" (Christensen 1993: 302), but one remaining is enough to "kill" your knowledge (DeRose 2000: 128). Even on its own terms Putnam's reasoning remains unconvincing as an antidote for skepticism—most of the vast literature has been critical and my presentation might be biased in this respect since the collection is quite balanced between those who assess the anti-skeptical potential of the argument positively (Brueckner, Ebbs, Button, Sundell) and those who are more doubtful (Pritchard and Ranalli, Folina, Sher). The connections between MTA and BVA might be tenuous (to loose to justify six articles out of fourteen altogether in any case), and perhaps some space should be dedicated to the historical dimension of BIV instead (this type of thought experiment did not start with Putnam in 1981). Still this is an excellent collection of papers provoking and extending discussion in various directions, the long-term life of the *brain in a vat* thought experiment seems to be guaranteed.

References

- Brueckner, A. 1986. "Brains in a Vat." *The Journal of Philosophy* 83: 148–16.
- Christensen, D. 1993. "Skeptical Problems, Semantical Solutions." *Philosophy and Phenomenological Research* 53: 301–321.
- Goldberg, S. C. (ed.). 2016. *The Brain in a Vat*. Cambridge: Cambridge University Press.
- De Gaynesford, M. 2011. "Putnam's Model—Theoretic Argument." In Hales, D. (ed.). *A Companion to Relativism*. Oxford: Wiley-Blackwell: 569–587.
- DeRose, K. 2000. "How Can We Know that We're Not Brains in Vats?" *The Southern Journal of Philosophy*, Spindel Conference Supplement 38: 121–148.

- Descartes, R. 2008. *Meditations on First Philosophy*. New York: Oxford University Press.
- Guyer, P. 1992. "Kant's Theory of Freedom by Henry E. Allison." *The Journal of Philosophy* 89: 99–110.
- Harman, G. 1973. *Thought*. Princeton: Princeton University Press.
- Horgan, T., Tienson, J., Graham, G. 2004. "Phenomenal Intentionality and the Brain in a Vat." In Schanz, R. (ed.). *The Externalist Challenge*. Berlin: Walter de Gruyter.
- Maddy, P. 2007. *Second Philosophy: A Naturalistic Method*. Oxford: Oxford University Press.
- Miščević, N. 2016. "In Defense of the Twin Earth—The Star Wars Continue." *European Journal of Analytic Philosophy* 12 (2).
- Putnam, H. 1977. "Realism and Reason." *Proceedings and Addresses of the American Philosophical Association* 50 (6): 483–498.
- Putnam, H. 1981. *Reason, Truth and History*. New York: Cambridge University Press.
- Putnam, H. 1992. "Replies." *Philosophical Topics* 20 (1): 347–408.
- Putnam, H. 1994. "Comments and Replies." In Clark, P., Hale, B. (eds.). *Reading Putnam*. Oxford: Blackwell: 242–295.
- Šuster, D. 2016. "Dreams in a Vat." *European Journal of Analytic Philosophy* 12 (2).
- Tallis, R. 2011. *Aping Mankind: Neuromania, Darwinitis and the Misrepresentation of Humanity*. London: Routledge.

Book Reviews

Michael Stuart, Yiftach Fehige and James Robert Brown (eds.), *The Routledge Companion to Thought Experiments*, London: Routledge, 2018, xiii+567 pp.

The Routledge Companion to Thought Experiments is a comprehensive and unprecedented collection of works meticulously compiled by Stuart, Fehige and Brown, the pioneers on the topic of thought experiments. The magnitude of the volume is nothing short of impressive as it draws together contributors dispersed across numerous spheres of philosophical inquiry. It is divided into four major parts, taking four different perspectives in approaching the discussion.

The first part is a selection of papers covering the topic of thought experiments from a historical perspective. It opens with a piece entitled “The triple life of thought experiments” by Katarina Ierodiakonou. In the beginning, she presents a couple of thought experiments from the antiquity including the one found in Aristotle’s *Physics* of a man standing on the edge of the universe trying to extend his hand, the famous Ring of Gyges from Plato’s *Republic* and the Sextus Empiricus’ in *Against the Physicists* dealing with the possibility of motion with regards to the existence of atoms, all of them serving the function of either confirming or refuting a particular theory. The purpose of her article is twofold; she explores the notion of thought experiments in ancient philosophy as a concept compared to its use in contemporary philosophy while also introducing a novel, somewhat uncommon role of thought experiments which was characteristic of the ancient Sceptics. In discussing the former she emphasizes that the term itself is a novel concoction and as such it has not been used by the ancient Greeks. Furthermore, she argues that they did not think of thought experiments as a special category of philosophical endeavor as they are thought of in contemporary philosophy but rather they were considered to be examples, corresponding to the Greek word *paradeigmata*. Nonetheless, she does not consider that to be an obstacle in applying the term thought experiments to their ‘examples’ as they share some of the core properties with what we call thought experiments.

After she has laid the ground for discussing the ancient ‘examples’ as thought experiments she delves into the function and usage of TE’s by the ancients offering an additional role to the confirmation or refutation of a theory, namely the suspension of belief, which can be found in the works of ancient Sceptics. Looking past refutation and confirmation as their func-

tion in discussions, she takes a step back specifying a general characteristic of ancient thought experiments: “the imaginary assumption initiates a process of thinking without a previously settled or determined conclusion, namely a series of arguments that should be clearly spelt out, compelling us to make up our mind on a particular subject” (35) which she considers to be the controversial nature of thought experiments. In support of that claim, she outlines the discussion between the Stoics and the Sceptics on several thought experiments, two of which are Plutarch’s *The Ship of Theseus* and Chrysippus’s *Dion and Theon*. Both the Stoics and the Sceptics agreed on the aforementioned controversial nature of thought experiments although they reached opposing conclusions; the Stoics used them to confirm or refute a thesis while Sceptics aimed at inducing a suspension of belief by allowing the possibility of reaching different conclusions.

Thus, what we can take home from her article is not just a piece of the historical puzzle of the ancient thought experiments but a lesson from the Sceptics as to the suspension of judgment which, in the contemporary setting is not advisable to be used with relentlessness and vigor of the Sceptics, but could at least make us more wary and less eager to settle for a conclusion which is controversial and ambiguous. Our skepticism should be rationed in healthy doses but employed nonetheless for it keeps us on our philosophical toes.

The second part of the collection is dedicated to the thought experiments with regards to specific branches of philosophy. Georg Brun’s “Thought experiments in ethics” is a compact and systematic analysis of thought experiments in the domain of ethics. After briefly outlining several thought experiments of the contemporary discussion including the ‘Trolley’, ‘Pond’, ‘Violinist’, ‘Ticking Bomb’, and the ‘Original Position’ he engages in a reconstruction of the thought experiments by explicating three key elements: “(1) A scenario and a question are introduced. (2) The experimenter goes through (imagines, thinks about, etc.) the scenario and arrives at some result. (3) A conclusion is drawn with respect to some target (e.g., an ethically relevant claim or distinction)” (196). Consequently, he makes a distinction between ‘core’ thought experiments which rely on the first two conditions and the extended ones which involve all of the three aforementioned properties thereupon dedicating the rest of the article to the analysis of the extended thought experiments. Firstly, his efforts are directed towards ‘epistemic’ thought experiments where he differentiates between constructive and destructive ones which are certainly the most prevalent functions of thought experiments together with it being one of the more commonplace classifications, inspired by James R. Brown. Constructive ones can either argue for the possibility of certain scenarios or provide support for a particular claim or a theory, while destructive are used as counterexamples to some claims emphasizing the problems with certain ideas. Subsequently, he turns to illustrative and rhetorical thought experiments. Illustrative, as the name says, are intended to illustrate or make the problem more vivid and relatable thus increasing the understanding of the experimenter. Rhetorical ones are similar to illustrative, however, they are employed when proving a particular point or arguing for a certain position. Pond experiment can be used as both of those. Another type are heuristic thought experiments

whose function resembles an ‘exploratory mission’ where the core experiment is run in experimenter’s mind in order to analyze the consequences and where it takes the experimenter. Sometimes they are used in determining which factors are relevant for evoking certain intuitions. As an example, Foot’s Trolley case has several variations which entice different intuitions about the problem. Their function is to extract the information relevant for making moral judgments.

He emphasizes that although epistemic thought experiments are the locus of the discussion on thought experiments, according to some accounts illustrative and heuristic ones do not fall behind in relevance. Specifically, it has been argued that understanding could be an important epistemic goal of thought experiments, no less potent than generating novel knowledge, to which illustrative and heuristic experiments majorly contribute. In discussing the functions of thought experiments, he narrows the scope to the ones grounded on reflective equilibrium since the functions vary with respect to meta-ethical theoretical framework. He discusses ‘wide’ reflective equilibrium which contains two components; one being that “judgments and principles are justified if judgments, principles and background theories are in equilibrium” (202) and the other that “this state is reached through a process that starts from judgments and background theories, proposes systematic principles and then mutually adjusts judgments and principles” (202). Under the assumption of cognitive equilibrium, thought experiments can be constructive in which an experimenter can produce a commitment to an option at any stage in the process, either in core experiments or in the extended ones, while deconstructive thought experiments use as a premise the result of a core experiment to point out the flaws in a theory or in the background assumptions which are challenged in the extended version.

There are several issues with the thought experiments in ethics, which are outlined in this paper. On the one hand, concerning those aiming at the result of core thought experiments, it has been argued that they reveal explicit commitments which appear in experimenter’s mind which is not necessarily how they would act were they faced in real situations. Furthermore, there is an issue with regards to intuitions since core thought experiments elicit ‘raw’ intuitions which can be revised in the extended ones during the process of cognitive equilibrium. The person could conclude the opposite of the content of his intuition in cognitive equilibrium, and some would argue that defeats the purpose of finding out what really is morally relevant. On the other hand, concerning the problems of extended thought experiments, destructive thought experiments do not always succeed in refuting the theory and it can point to the need for rethinking some assumptions, however, it does not pinpoint which information, in particular, has to be revised. Moreover, some thought experiments are analogies constructed based on a theory in support of it which is problematic since in order for transferring assumptions they need to be explicated.

Challenges to thought experiments are numerous and are directed either to a certain function of thought experiment or to a specific thought experiment. The author briefly outlines various ways in which thought experiments are put on spot, for example, the issue of intuitions generated by them, the possibility or lack thereof to be carried out in the real world,

deriving to conclusions etc. Naturally, Brun pays more attention to some well-known objections directed to ethical thought experiments, namely the ones questioning how realistic should thought experiments be and the others that argue for them being misleading or generating faulty results.

Turning to challenges which address the problem of thought experiments being unrealistic, it is argued that they do not justify moral principles which are developed to govern our actions in real life situations to which the author replies that some thought experiments deal with more fundamental principles that lead moral judgments to which thought experiments still hold relevance. Another challenge argues for the unreliability of core experiments of unrealistic scenarios by either questioning the reliability of intuitions or inability to discern what is morally relevant because of our own beliefs.

A distinct set of challenges assert that thought experiments are misleading on several accounts; one being that they pose dubious questions not encountered in our day-to-day lives or questions which limit the scope of answer. As an example he uses the "Should you pull the lever?", one which is not a plausible real life situation and which can only be answered with 'yes' or 'no' thus 'leading the witness', so to speak. Additionally, it is argued that they implicitly contain problematic assumptions while side tracking the additional information which might prove to be essential. Lastly, there is one more challenge to thought experiments, addressing the fact that some thought experiments are constructed in the form of analogies so that they lead the experimenter to draw conclusions about a situation different than what has been depicted in the experiment, examples of which are Pond and Ticking Time-bomb thought experiments. The author replies to two such objections to using analogies.

The author concludes with the warning that the discussion on thought experiments in ethics should not be taken lightly as inadequately constructed thought experiments may be used in public discourse for promoting immoral and problematic agendas. This paper is instructional both for novices in the exploration of ethics as a branch of philosophy as well as for the students tinkering with the subject of thought experimentation. It would prove to be no less useful for the experts of both fields as it compresses a masterfully elegant compendium of ethical intricacies which could prove to be a valuable reference text.

Nancy Nersessian's article "Cognitive science, mental modelling, and thought" experiments explores the underlying cognitive mechanisms which are employed in the process of thought experimenting. Her efforts are directed to accounting for the psychological frameworks which make such inquiries possible and which ultimately generate the knowledge that is novel in our everyday lives as well as in the work of science. Her hypotheses are supplemented by an overview of the body of work she offers from the fields of psychology, cognitive and neurosciences, and philosophy. After briefly introducing some basic notions and problems of thought experimentation, she outlines the 'story so far' concerning the mental model framework of which she has been the architect alongside Nenad Mišćević in this vast edifice that is the discussion on thought experiments. From the introduction of the term 'mental model' by Kenneth Craik in 1943 who hypothesized them as a

modus operandi of people's reasoning about physical situations by means of employing internal models in exploring them, to the no less influential work of Johnson-Laird whose *Mental Models* (1983) exploring the notion of logical reasoning, working memory and mental models. Although their views of mental models differ in some respects it undoubtedly casts a shadow over the investigation and discussions of them in years to come which is enormous and beyond the scope of Nersessian's paper.

Consequently, her attention is directed to interpreting literature on discourse and situational models in dealing with the issue of how mental models are constructed, the prevalent view being that thought experiments are revealed through narrative. However, the importance of narrative does not lie in the "system of propositions representing the content of the text" (313) to which we apply rules of inferences but rather that the model being manipulated is that of the situation represented by narratives as "discourse models make explicit the structure not of sentences but of situations as we perceive or imagine them (Johnson-Laird 1989: 471)" (313) In support of that claim, she mentions several experiments all pointing to the aforementioned hypothesis.

According to Nersessian, another key cognitive faculty which partakes in thought experimenting is mental spatial simulation which means that humans have the ability to mentally transform and manipulate objects in space that is akin to the physical transformation. After giving a couple of examples in over-viewing the literature exploring such capacities she concludes with the words of Kosslyn that: "psychological research provides evidence of rotating, translating, bending, scaling folding, zooming, and flipping of images" (314). It is hypothesized that such abilities are due to 'internalized constraints assimilated during perception'. Additionally, she cites the research which points to physical knowledge taking part in imaginary transformation noticing the subtle connection of imagination, perception and action emphasis that mental spatial simulation can be employed in manipulating both representational and non-representation content. Supplementing that notion with the literature on mental imagery and spatial simulation she concludes that perceptual and motor mechanisms do in fact largely contribute to construction and manipulation of mental images.

Together with mental simulation she explores the subject of mental animation. Even though they are closely related, mental simulation deals with spatial and temporal transformation, while mental animation includes causal and behavioral knowledge. In other words, mental animation is about mentally bringing static representation to life by inferring motion. To illustrate this, she uses prominent research done by Mary Hegarty's Pulley systems and Daniel Swartz's gear rotation studies which supply evidence for the human ability to perform "simulative causal transformations of static figures" (316). She highlights several findings, some of which are that participants animate the objects in a sequence which is dissimilar to how they would be manipulated in the physical world, they often use gestures while performing such mental actions, etc. together with the findings from the interference paradigm which imply that performing physically incongruent action to the mental animation prolongs the participant's response time. Additionally, she provides insight into neuroimaging studies which show that

the same brain areas involved in carrying out motor actions are employed in mental simulation, not to mention the fact that observing an action engages the brain in a similar fashion to actually performing the action.

In efforts to ground her theory in the long-term memory representation as the paper so far outlines compelling evidence just in the domain of working memory she includes the research done on embodied mental representation. The research on embodied mental representation aims to show that perception and action are integral to numerous cognitive processes such as “memory, conceptual processing, and language comprehension” (317). She outlines two strands of research in the domain of embodied mental simulation. One deals with the representation of spatial information in mental models the results of which indicate that spatial representation is not ‘3D Euclidian’ in relation to one’s body and gravity. In other words, representation of spatial information is ‘egotistical’ linked to the person’s body as a frame of reference.

Another line of research she lays out tackles the representation of concepts. In support of Barsalou’s view that mental representations maintain perceptual features which are reenacted during cognitive processes, which is his interpretation of current research in cognitive and neurosciences, she also outlines his distinction between modal and amodal features of concepts, introducing the idea of perceptual symbols as fundamental representations of both conceptual and sensimotor processing.

The aim of the research disclosed to this point aimed at setting the stage and being constituent of thought experiments as simulative model based reasoning. The cornerstone of such view is that people in their reasoning take advantage of mental models which they manipulate through simulation. Thus, mental models can be described as organized representations which are determined by the constraints of experience and current understanding like the knowledge of spatio-temporal relations and properties of entities, processes etc. The aforementioned constraints are as Nersessian enumerates them: “tacit and explicit knowledge of spatio-temporal relations, the represented situations, entities, processes, and other pertinent information such as causal structure” (319). In manipulating mental models we draw from linguistic, auditory, visual, kinesthetic and many other cognitive faculties. She sees thought experiments as fundamental to human reasoning and as such its application to scientific reasoning is all the more reasonable. Even though certainly more complex in nature, they are also accessed through narratives, which, as we have already seen, entice the experimenter to manipulate the mental model of the situation depicted, rather than draw inferences from proposition-like statements. Further, she distinguishes between fictitious imaginings and thought experiments with real life consequences in human day-to-day reasoning deeming the latter as far more significant. On that account, she argues that thought experiments in science exploit the same capacities she outlined so far in the paper. Her hypothesis being:

that the carefully crafted thought-experimental narrative leads to the construction of a mental model of a kind of situation and that simulating the consequences of the situation as it unfolds in time affords epistemic access to specific aspects of a way of representing the world. (320)

Lastly, she tackles Norton's view of thought experiments as arguments which enforces the notion that thought experiments produce truths about the nature. Nersessian, seeing such a view as too "epistemically potent" (320), offers two arguments to oppose it. The first being that thought experiments refer to the kind of phenomena being explored, not to the particular situation, thus making them generic. Second, she argues that science uses many devices and practices which do not always generate truths about the phenomena but, nonetheless, tell us something about the nature of things.

In the introduction of their paper "Intuition and its critics", Steven Stich and Kevin Tobia draw a parallel between linguistics and philosophy with regards to intuition. In Chomskian terms, intuition drives the spontaneous application of grammatical properties and rules to novel sentences. The speaker does not have to be consciously aware of the rules when they make grammaticality judgments and sometimes it is possible to make errors in judgments because of various factors that might impede on speaker's attention, memory etc. Similarly, philosophers have posed questions about the world and its characteristics in the form of hypothetical situations, evoking the intuitions which present themselves instantly in minds of participants of such discussions without explicit appeal to the rules of reasoning. On that note, their paper is based on the use of term intuition "for the spontaneous judgments that people make about philosophical thought experiments" (370).

After defining their use of the term "intuition", they set out to explore the usage of intuition as evidence in philosophy which brings them to the pre-Chomskian years of logical positivists whose view on the purpose of philosophy was conceptual analysis. Alongside this view, one of the methods of conceptual analysis were thought experiments and compiling intuitions evoked by them was the means of acquiring evidential significance. Justification for their use is similar to the aforementioned Chomskian take on intuitions about grammar shared by philosophers such as Alvin Goldman who maintains that intuitions can bear relevance in exploration of the content or extension of the concept. Another view makes use of intuitions as evidence for or against theories about phenomena in philosophical discussions for example truth, justice, good etc. different from conceptual analysis in that they do not seek to pinpoint the people's concept of these things. Conjointly, these two stances correspond to two ways of dealing with philosophical problems depending on their goal as outlined by Goldman and Pust:

Broadly speaking, views about philosophical analysis may be divided into those that take the targets of such analysis to be in-the-head psychological entities versus outside-the-head non-psychological entities. We shall call the first type of position *mentalism* and the second *extra-mentalism* (1998, 183). (370)

Accordingly, mentalist analysis deals with investigation of concepts or in-the-head psychological entities sometimes aided by implicit or tacit theories in their explanation of intuition generation. Conversely, extra-mentalism's analytic aim is harder to discern thus Goldman and Pust in efforts of narrowing down the scope of its inquiry emphasize three domains of their exploration: universals or Platonic forms, modal truths and natural kinds, taxonomy to which Stich adds moral facts. Their common denominator is that: "the correctness or incorrectness of an extra-mentalist theory does not depend on what is in the head of a person whose intuitions are used

as evidence" (371). They consider people's intuitions to be the truth about the extra-mental entities they explore. The problem with such account is the ambiguity and inexplicability of the connections between intuitions and aforementioned domains since it is not clear how we intuitively access for example Platonic forms. However, more problematic claim of extra mentalism is the previously mentioned stance that intuitions derived from thought experiments indiscriminately illicit the truth about these extra mental entities. Further, intuitions are challenged by another strong and budding philosophical branch: the experimental philosophy. Contrary to extra-mentalistic position, evidence from experimental philosophy indicate that intuitions vary among people depending on a number of factors which they briefly outline in the followings sections including the variation of intuition with regards to demographic groups, language and order in which the experiments are presented. Furthermore, findings from experimental philosophy also indicate that intuitions are not immune to framing effects and that they are affected by the physical and social environment in which they are evoked. (Thus, intuition one person has in the Trolley case of pulling the lever does not mean that pulling the lever is morally permissible since another person has the intuition of not pulling the lever) Besides the fact that studies show that intuitions vary across groups and conditions in which they are elicited, an additional problem is that people of the same groups and under the same conditions still seem to report having differing intuitions.

As a side note, most of these studies also endanger the mentalist stance on concepts with the exception of evidence that suggests that people of different demographic groups have in fact distinct concepts. As an example, people's concepts vary with respect to the academic field of their interest. Though such evidence do not pose problems for mentalist position on concepts per se, it should be specified beforehand whose concepts and why they are investigating. The findings brought forth by experimental philosophy undoubtedly pose problems for mentalist and extra-mentalistic analyses. In rising to their challenge, Stich and Tobia propose two ways of overcoming them.

The first appeals to professional ineptitude of the participants in the studies, also known as the expertise defense, which argues that the studies do not offer valuable insight for philosophy since the participants themselves are not professional philosophers. Analogous to other professions, we seem to deem the intuitions of doctors or chess players of more relevance than those of amateurs in those fields. There are several positions one can assume in taking the expertise defense; one asserts that philosophers are less likely to be seduced by the aforesaid factors which interfere with generating intuitions such as the order of presentation, framing or "ambient odors" while the other relies on the notion that intuitions of philosophers are more accurate than those of non-philosophers.

Stich does not hold the former approach in high-esteem as evidence, although scarce, does not seem to point to philosophers' immunity to such hindrances. The latter approach, enforced by Daniel Devitt in the domain of philosophy of language leave much to be desired. He engaged in an extensive theoretical exploration of the subject which regrettably has not barren fruit in the empirical, experimental examination so far. Granted, it is extremely difficult to empirically test whether philosopher's intuitions are

in fact more accurate than regular folk's intuitions so Devitt's efforts are nothing if not commendable. Still, one can take an alternative approach to what has been outlined so far, known as the restrictive accounts of philosophical intuitions. By defining intuitions more narrowly, their incentives are to explain why intuitions might be reliable enough to count as evidence and to fend off the attacks of experimental philosophy. One of the authors who endorse the restrictive position with respect to intuitions is Ludwig who proposed that only the intuitions derived from conceptual competence are the ones which are valid. Conversely, those influenced by factors mentioned earlier like framing or order of presentation which do not fall under the conceptual competence should not be regarded as intuitions. Such view, however, is not without its problems since it is almost impossible to tease apart conceptual competence from those interfering factors since the experimenter herself is not consciously aware of them. Authors like Cappelen even go a step further in their restriction of what intuitions entail narrowing their scope so profusely that even philosophical discussions do not seem to include them. In that sense, experimental philosophy does not endanger the philosophical practice but consequently, his proposal has not gained much momentum among philosophers. The paper ends on an optimistic note that even though intuitions are highly problematic they should not be discarded but rather they should be thoroughly explored further in which experimental philosophy should play a key role.

Let me pass on to Michael Stuart's "How thought experiments increase understanding". As the title indicates, this paper belongs to the domain of epistemology, its aim being the capability of thought experiments to increase understanding. The answers to why that function of thought experiments should be analyzed, are brought forth in the very beginning of the paper. Upon noticing that a great deal of discussion on thought experiments from the epistemological perspective is concerned with the question of how thought experiments generate new knowledge without experience, the author has directed his efforts to an important epistemological aspect which does not receive as much attention as it should, namely the contribution of thought experiments to understanding. As he points out, there are numerous roles thought experiments can assume to contribute to understanding the world among which are illustration of a theory, exemplification of properties and relations, provision of hypotheses and many others. Their sole function need not be increasing the experimenter's knowledge to be epistemologically significant. In order to see how thought experiments increase understanding, the author first tackles what understanding is and what it entails. He highlights Catherine Elgin's view on the subject which does not limit understanding to propositional knowledge but widens the scope to include work, actions, passions, situations etc.

Along the lines of her claim, there have been many classifications and subtypes of understanding; transitive and intransitive, propositional and non-propositional, interrogative and noninterrogative, to name a few. However, the focus of this paper is on three types of understanding: explanatory understanding (EU), objectual understanding (OU) and practical understanding (PU). Explanatory understanding is based on explaining, as the name says, of why some state of the matter is the way it is and it often but

not always, takes the form of propositions. Objectual understanding is the understanding of a thing, or an object itself and in relation to the context and subject matter it is immersed in. Finally, practical understanding is akin to tacit or implicit knowledge, basically knowledge “how” for example “Jimi understands how to play the guitar” (529) and it is contrasted with explanatory as it is not run-of-the-mill propositional knowledge.

It is mentioned that there is a debate about whether some kinds of understanding previously outlined can be reduced to just two or even one subtype of understanding. Stuart insists on their separation arguing that each type is obtained differently and we have distinct ways of pinning down their realization. Naturally, while explanatory understanding should strive for providing a better explanation of a phenomena and practical understanding should foster some abilities, objectual understanding is not as easy to pin down as its purpose is the understanding of the relations between things such as entities, events or experiences, objects and background knowledge. The authors opts for understanding the semantic content.

In subsequent sections the possibility of each of these types of understanding as a result of thought experimentation are given a closer look. In support of the hypothesis that thought experiments contribute to explanation, several arguments and studies are offered; one being an online survey which showed that people (some of which professional philosophers) strongly favor thought experiments as a method of explanation, another was a study on thought experiments in textbooks which reported that many thought experiments are employed because of their explanatory power even though they may be outdated. Further, they are prevalent in literature for explaining a variety of phenomena, for example, Darwin’s vertebrate eye, Newton’s cannonball etc. Explanation can also be viewed as consolidating phenomena that are in opposition to each other “why does x happen as opposed to y?” (531). In such case thought experiments also do not fall short. It is also stated that they provide explanation in situations where causal relation is sought, for example, in counterfactuals, causal chains etc. What these examples tell us is that thought experiments do increase understanding since explanation and understanding seem to be inextricably linked. Furthermore, thought experiments, as it is argued, seem to increase meaningfulness by enhancing the semantic connections between objects, entities, experiences and so on in contribution to objectual understanding (OU). The scientific thought experiments often assume such roles as they make the problematic and sometimes unfathomable concepts or theories more accessible to the laymen as well as to the students on their way to becoming experts. The history of science is abundant with such examples and two of them are briefly outlined in this paper, namely Darwin’s vertebrate eye and Maxwell’s demon.

Thus, the author asserts that thought experiments help us make semantic connections between concepts, theories, entities and between our past and present experiences, abilities etc.

Consequently, several remarks are disclosed in arguing for their fruitfulness. What is meant by that is the property of some thought experiments which makes us able to do something we had not been able to do before engaging in thought experiment for example “manipulate a model, make

a successful prediction, produce a good explanation for a phenomena, derive to a particular conclusion" (533). To support his claim, he mentions a couple of examples such as using thought experiments in therapy in order for clients to confront their fears, in education for incapacitating students to make further predictions and inferences about phenomena but also in the history of science as Darwin's vertebrate eye nudged the scientists in the following years to acquire the mechanisms by which evolution functions thus generating new hypotheses. As to how thought experiments increase understanding, the author focuses on objectual and practical understanding since explanatory is beyond the scope of this paper.

In explaining objectual understanding he references the work of Elizabeth Camp and her notions of perspective, characterization and frames. Perspective is the position we assume with respect to the world described in the narrative "as if it were the way the narrative presents it" (534). The application of perspective to a particular instance or a situation is labeled characterization (534). Framed is described as: "a representational vehicle that crystalizes a perspective by suggesting a characterization" (534). With the aid of these terms, the author further explicates how they contribute to understanding. Thought experiments provide frames with which we tap into characterizations. Although they are non-propositional they can be transcribed in forms of propositions but that is not where their potency lies. As Stuart asserts, they are "tools for thinking" (535) and good thought experiments are those which provide good frames by which we can assume a certain perspective which will be of epistemological significance.

In dealing with practical understanding, he highlights Alison Hill's explanation in terms of 'grasping' and 'cognitive control'. Having cognitive control means having the ability to manipulate propositions, i.e. to explain propositions and what can be deduced from them, to form analogies of propositions etc. Even though her account focuses primarily on propositions, Stuart argues that it can be applied in cases of gaining practical understanding by thought experimenting. That can be achieved by exposure to questions and analogies which have to be worked through to get a certain result. In that sense, they thought experiments should be formed in a way that they provide the necessary information and some guidelines to point the experimenter in the right direction, however, the result should be gained independently by working out a certain conundrum thus gaining 'cognitive control'. Gaining practical understanding can also be achieved through various tasks and puzzles, the important element being, as he asserts, the open-endedness of a particular problem without giving away possible solutions before going through an experiment in one's mind. Finally, he proposes some ways of exploring whether thought experiments increase understanding by introducing and testing them in educational academic settings.

MIA BITURAJAC
University of Rijeka, Rijeka, Croatia

Harris Wiseman, *The Myth of the Moral Brain. The Limits of Moral Enhancement*, Cambridge: The MIT Press, 2016, 352 pp.

For the past decade, the academic debate on the possibility of human enhancement¹ has produced quite a substantial record (Agar 2007). The youngest entry to the enhancement debate is the theme of moral enhancement. This rising new field of research, both scientific and philosophical, is “concerned not so much with the improvement of physical or cognitive capacities, but improvements in the way in which we act or reflect morally,” (Raus et. al. 2014) and is very much fueled by the rapid progress in fields of neuro and cognitive sciences. Of the many possible approaches to moral enhancement, the biomedical approach has become the focal point upon which many spears have been shattered in the still ongoing debate’s two opposing camps. These are the proponents of the traditional moral enhancement which include approaches such as moral education, advancement of moral reasoning etc., and those, on the other side, who argue for a direct use of biomedical procedures in trying to advance human morality. As the possibility to enhance morality through biomedical procedures in the past years has become entrenched within the neuropharmacological capacity to facilitate these desired modifications, the debate, unfortunately, hasn’t quite moved on with novel explorations.

In this regard, Harry Wiseman’s most recent work *Myth of the moral brain*, in which he systematically and thoroughly engages the predominant approaches to moral bioenhancement, is more than a welcomed refreshment and, for some, quite a realistic sobering. Wiseman’s work comes out just in the right time when other engaged scholars (including neuroscientists) are also pointing out that the neuropharmacologically based proposals, which have received the biggest impetus, hold serious and somewhere even irreparable flaws. For instance, Dubljević and Racine (2017) in their most recent contribution create a thorough assessment of currently predominant neuropharmacological options for the biomedical approach and find them all wanting,² Wiseman directly contributes to these findings with

¹ “Human enhancement... aims to develop technologies and techniques for overcoming current limitations of human cognitive and physical abilities...rely on advances in genetic engineering, pharmacology, bioengineering, cybernetics, and nanotechnology. The envisioned applications are limitless, and include the enhancement of human traits like muscular strength, endurance, vision, intelligence, mood, and personality” (Brey 2009: 169).

² To name just a few of their important findings, Oxytocin was found to promote trust, but only in the in-group, while with the out-group members of society it can decrease cooperation and selectively promote ethnocentrism, favoritism, and parochialism. Beta blockers were found to decrease racism but also blunt all emotional response which puts their effective usefulness in general doubt. SSRIs (Selective Serotonin Reactive Inhibitors) reduce (reactive) aggression, but have serious side-effects, including an increased risk of suicide. Deep brain stimulation was found to have no effect whatsoever on moral behavior. And so they conclude that biomedical and especially neuropharmacological „techniques are all blunt instruments, rather than finely tuned technologies that could be helpful” (Dubljević and Racine 2017).

a refreshing and, at many places, consummate entry as he aims to offer a realistic critique of the biomedical approach both in its philosophical musings and scientific underpinnings. The main point, and a general motif, of the book is that we require a more realistic approach which Wiseman terms the “bio-psycho-social” (245) model inside which he aims to “base our rationales for moral enhancement upon this foundation of what is realistically possible” (53). The book thus, in general, should be recommended as a good entry point for anyone interested in the moral enhancement debate (at least in its analysis of the ongoing debate) as it aims to dissect the bloated vision of some moral enhancement scenarios as well as trying to show where does exactly real science stand on issues pertaining to it. Following, the book is divided into four main parts: Philosophy, Science, Faith, and Praxis. We will explore them in order.

In Philosophy, Wiseman focuses on the works of Ingmar Persson and Julian Savulescu, James Hughes, and Tom Douglas. The hardest hit of the three gets the Persson Savulescu duo. And this is not surprising since in general Persson and Savulescu’s approach has generated the broadest amount of critiques. Wiseman aims to deliver the killing blow as he constantly engages their proposal throughout the entire book seeing it as a “hideous visage” a hypothesis that puts forward a “literally, morally enhance or die” (263). He believes that the Persson-Savulescu thesis has “really made a joke of this domain” (263) and hopes that this approach may “be abandoned by commentators completely, leaving nothing over and that it never be spoken of again” (263). The second, James Hughes, Wiseman credits as the “arch-transhumanist, perhaps the most intellectually credible of all transhumanists” (34) and engages his account of “voluntary virtue engineering” which is all about how “you are free to morally enhance yourself in any way which encourages free society” (44). This ought to be done by linking neurochemical changes with achieving the desired liberal personality as the morally superior option. Even though he puts aside the notion of liberal moral superiority, Wiseman is not impressed with Hughes’ approach which he sees as a “clumsy way of conceptualizing the operations of moral enhancement” (46) since it cannot guarantee to attain its specific moral character results and at the same time ignores unexpected side effects. As the “arch-transhumanist” Hughes should be strong on science but this is exactly what Wiseman points out he lacks the most and through him aims to show the focal mistakes of “enthusiast” enhancement proponents in general. Namely, that they are building up a “poorly evidenced and massively overoptimistic account of moral enhancement possibilities based on highly provisional and contested research” (46). Conclusively, Wiseman deems Hughes’ approach as “simply unrealistic” (46). The last one to be tackled, Tom Douglas receives the least critique given and even modest accolades as although, “Douglas’s approach should not be taken as a complete package, Douglas has managed to carve out a very limited but more realistic prospect for moral enhancement” (57). Douglas is not found to be guilty of enhancement enthusiasm but rather, according to Wiseman, offers a sober and precise outlook on the matter and from the looks of it could be taken as a proper example in evaluating the biomedical vision for moral enhancement. Still, his approach is seen only to best function with those

“moral problems that are predominantly or totally impulse-based rather than those requiring moral reflection and discernment” (52).

The second part, *Science*, aims to establish how realistic are the conjectures between regulating different neurobiological substances such as hormones or neurotransmitters with a personal disposal to behave and think morally. Wiseman focuses on the central and most predominantly present themes: Oxytocin which is connected with empathy, trustworthiness and generosity (Paul Zak), Serotonin which is connected with harm, fairness, and aggression (Molly Crockett) and Dopamine which is connected with rewarding behavior (Ed Boyden). The general conclusion Wiseman comes to is that none of these neurochemicals is powerful enough to fulfill the full scope required of the moral enhancement goal. Notwithstanding the many and possibly permanent undesired side-effects, the current state of neuropharmacology is simply inadequate to create the desired effect of moral enhancement. For instance, Oxytocin has a “nudge” potential but only with those who are already disposed towards prosocial behavior or empathy. Serotonin, especially through the SSRI (Selective Serotonin Reuptake Inhibitors)—a broadly available neuropharmacological substance has received a substantial appraisal. But Wiseman shows that not only is it true that what SSRI might improve with respect to one kind of aggression, namely reactive aggression, they may worsen with respect to another, namely premeditated aggression but that the complexity of Serotonin dependent systems (immune system for instance) is highly sensitively calibrated and purposely manipulating with Serotonin levels in the organism could lead to devastating side-effects (Therbeck-Chesterman 2013, Crockett 2014). Thus, in summary, Wiseman tackles not only the inability of neuropharmacology (he does applaud Boyden’s optogenetics approach with whom he shares a disbelief in neuropharm) to address the issue at hand but also tackles the incorrect emotional frameworks inside which certain emotional states (aggression for instance) are seen as being almost necessarily morally inhibitory or unwanted. He also warns of those frameworks which place a sharp distinction between emotions and reasoning and thus espouse an incorrect view of human moral cognition and its underlying sub cognitive processes (Helion and Ochsner 2016).

After dealing with neuropharmacology Wiseman confronts another and perhaps even more important problem—that of conceptual and methodological frameworks found in moral enhancement research. He uses the example of the recently given SSRI research (Molly Crockett) which has been viewed and consequently used by many researchers (Wiseman focuses on DeGrazia) as a very promising scientific result to reaffirm the moral bioenhancement approach. Unfortunately for the enthusiasts, Wiseman confirms another sober awakening (for the entire enterprise) by pointing out to a critical problem—that of external and ecological validity. He humorously (and almost ironically) remarks on the inadequate validity and thus usability of these scientific findings since the experimental frameworks in place are neither contextualized nor embodied—a hallmark of real-life human morality. As he poignantly remarks: “Indeed, it does seem as if most of the science upon which moral enhancement enthusiasts draw is conducted either using Ivy League students, or mice” (117). Wiseman in this regard calls in for a

much-needed refinement of methodological and conceptual paradigms and for a case by case approach in dealing with issues of moral enhancement especially in evaluating certain moral traits since the scientific experiments made and philosophical frameworks built upon these findings are detached from a real-life instantiation of expressing these traits. Additionally, he warns, scientists themselves sometimes publish their work with ingrained “seductive claims” which draw enthusiasts to infer conclusions that are, unfortunately in the end deemed incongruous. Additionally, even the best cognitive science frameworks such as the “Crockett’s Jekyll and Hyde, Greene’s dual-process theory of moral functioning...” (101) are inadequate to be used as a clear-cut extrapolation for philosophical conclusions. As he humorously remarks that, for instance, the trolley problem dilemma cannot be viewed as a realistic scenario “unless one is Oedipus standing before the sphinx” (120) and concludes that “these reductive approaches which rip moral functioning out of its meaningful contexts, strangle it through excessive control...and distort beyond all recognition and meaning the moral phenomena being investigated ... are simply not fit for purpose” (126).

Finally, in the third and the fourth part aptly named Faith and Praxis, Wiseman’s proposal for moral enhancement is, it could be said, not far from that ancient Benedictine motto *Ora et Labora*. Thus, in the first part, faith, Wiseman tries to offer a distilled number of core Christian theological points that portray a “Christian virtue ethical theology” (297) in putting forward a realistic attitude (the *leitmotif* of the book) which espouses that “moral enhancement cannot exist as a free-floating entity (as if apolitical, or here as a-religious), but rather needs to recognize the nature of the ground upon which it is to stand and build” (142). Since a big percentage of the human population, at least declaratively, are professing a certain religious stand, he takes that any “strong vision of moral enhancement will and must be understood in a way that can cater to the billions of persons who self-identify with one faith tradition or another ... and who will not be satisfied by a generic account of moral enhancement which attempts to simply ignore crucial tenets of their faith” (142). In this regard, he echoes some of the growing concern for urgency in that it is better for religious thought to engage the debate on moral enhancement sooner rather than later since “faith communities are not going to be neutral on moral questions, nor upon questions regarding moral formation” (141).

But why does he pick Christianity? The reason, it is said, is purely practical as he believes that “Simply put, there is a familiarity with the Western audience with matters of Christian faith, much of which is absorbed by osmosis, and often unconsciously and anonymously” (140). He doesn’t aim to put the Christian approach as the supreme approach but merely as one with which many thinkers are acquainted. Still, just a bit later he introduces the notion of Christian generosity, “the outward-facing” focus as an antidote to the “self-obsession and tremendous anxiety” (145) which results from the self-absorbed contemporary culture’s way of life. So perhaps the Christian approach is not here as just *the most practical option* but also serves as a critique of the contemporary culture and resembles previous Christian critiques of transhumanist philosophy and enhancement in general. Additionally, as is presented later on in the text, it seems that the Christian

community has the greatest generative power and overall functionality to foster a virtue-based remedy-like approach to moral enhancement which Wiseman espouses and thus, it seems, Wiseman's reason of choice goes beyond a practical "secular familiarity" with Christianity.

Still, before presenting us with his main proposal Wiseman once again reiterates his already well-established critique of unrealistic ME scenarios. And he wishes to deal them a final blow by confirming the inadequacy of biomedical approaches to solve those dimensions of morality which are completely out of their scope or category such as the "context, ambiguity, moral scaffolding, the predisposition of will" (189). For instance, concerning the context or ambiguity of moral goods, he cleverly remarks that "none of the enthusiasts in the philosophical literature want to get inside the heads of those who are to be enhanced. Yet people are motivated by different things, they understand their moral goods in different ways, and they need to be spoken to in different ways" (180). And on the issue of scaffolding: "moral enhancement might help augment a given vision of the good, but it cannot itself create a vision of the good" (185).

This conclusion is in line with his priory emphasized anti-reductionist stand but this time it is reinforced with regards to religious moral beliefs: "the empirical work conducted on 'the moral brain', makes no reference at all to the manner in which a person's religious faith may or may not be modulating their responses to the various tests that are applied" (147). This also serves as foreshadowing the socio-political acceptance of moral enhancement within the religious landscape: "a strong vision of moral enhancement must by implication propose some rationale for...contributing to the salvatory structures idiosyncratic to the faith traditions of those upon whom such strong visions of moral enhancement are to be impressed" (147). This ties in directly with the discussion and the distinction on the voluntary/compulsory enhancement since one could presume that a religious person would also seek religious (in between other) reasons when deciding to voluntarily pursue moral enhancement. In addition, policymakers would have to take into consideration religious sensitivities when engaging enhancement possibilities. Wiseman believes that what is important to have in mind in both of these cases is that we cannot devise a general like solution applicable to each and all but that "we need to be asking which *particular* intervention is best understood in voluntary terms and why the particular facts on the ground make things so" (203). This conclusion is especially important if we recall that the true problem with a compulsory general-like moral enhancement of the population is not only in the inability of the neuropharmacology to achieve such a precise level and intricacy of interaction with our biological systems—for instance by providing to the entire population an "empathy pill" but that even if we could do so (and we cannot) we must remember that certain emotional states which humans exhibit are there for an (evolutionary) reason—more often than not as a fail-safe survival mechanism. As such, if one would follow the idea to its end we might come to see, as I call it, the birth of an Eloi society. As the famous Eloi, the surface dwellers of the far future Earth depicted in the H. G. Wells *Time Machine* show, a being completely lacking the capacity to express anger or aggressiveness even if just to defend itself is a sitting duck in a world of evolutionary sur-

vival and predation (Prinz, 2011). And although Eloi, as well as Morlocks, were engaged in literal survival and predation, where Morlocks used Eloi as food, our own world is fraught with survival and predation with the difference that we, true enough, don't literally each other. As Wiseman remarks: „What use is an intervention to generate empathy in a society which rewards and valorizes cruel, self-serving, aggressively competitive behavior?“ (187).

And maybe this is the main reason why Wiseman wants to incorporate the biomedical procedures within a virtue-based character development inside a communal Christian narrative. If I am interpreting his intentions right, it seems that if we cannot opt for a global compulsory approach (due to its obvious problems) and neither can we rest our hopes upon the voluntary approach since the majority is not interested in moral enhancement at all—the only viable solution we are left with is the arduous “renewal” of society from within. And this *moral enhancement renewal*, it is presumed, could be achieved by the Christian community since it could create the cohesive and generative power to venture forward towards a realistic goal of a morally enhanced humanity. Be it as it is, his vision of the remedial moral enhancement proposal “one in which a biomedical intervention takes place in a mental health context, in a person-centered and fully bio-psycho-social fashion, one which respects the value and influence of personal agency, cultural scaffolding, and quality relationships” (220) should be applauded. Still, as he is fully aware, problems remain. As is the case with all interpersonal and group dynamics, the ones in charge are the ones who have the greatest influence in determining the outcome of the procedure. Since Wiseman requests a communitarian approach, with healthcare experts and counselors aiding or guiding the process of moral formation and providing the necessary scaffolding or moral motivation to the individual—a natural question arises: “*Who then watches the watchers?*”.

Wiseman is aware of the problem but is not able to provide a direct solution instead of pointing out that “we need responsible institutions in place, along with healthcare professionals who are not swayed by ... inappropriate shortcuts and easy remedies to complex problems” (223). And this leads me to a concluding remark in which I am left to wonder is Wiseman's proposal fulfilling or limiting the vision offered to us by moral enhancement? Surely, Wiseman gives his best in giving thorough argumentation why exactly currently present neuropharmacological means to moral enhancement will not be able to do the trick. And he does it successfully. Still, one is left to wonder if in pointing out all the faults and lacks the neuropharmacological approach holds both in its science and philosophical interpretation Wiseman doesn't leave us with much in striving for and achieving the grand vision of moral enhancement. According to Wiseman, it seems that the sobering reality of human biology, the complexity of the socio-political landscape and intricacy of even our everyday human morality calls us to reconsider our moral enhancement proposals to “sacrifice fantasy for something that might actually be of use, here and now” (226). But I cannot shake the idea that this approach no matter how much it works to be as realistic as possible, in its fervor for realism loses the hope beyond the horizon. The vision of moral enhancement has to be able to provide us with more than simply putting

it all, like so many times in the past, on the back of the individual. Unfortunately, for Wiseman the project of moral enhancement “is absolutely dependent upon the efforts and will of the person so enhancing” (279). Even if provided with a community to support the incentive and the lack of motivation by providing a safe guidance of a counselor, or a *moral doctor*, and if necessary administering remedy-like pharmacological means (Nalmefene and alcoholism example (233)) it all rests once again on the individual will, on the individual openness to attain or not to attain moral enhancement. So, as it seems, everything within the process has an instrumental role while the will of the individual determines the success of the procedure. And this conclusion is not something I can agree a moral enhancement vision should be built upon for the simple reason it lacks the capacity to enhance that what needs enhancing the most—human will. Surely neuropharmacology alone cannot be deemed as a “one size fits all” solution or even an efficiently applicable solution but the lacking’s of the neuropharmacological approach do not entail our incapacity to accomplish a grand vision of moral enhancement. And although the complexity of the moral life far exceeds the narrative neuroscience and neuropharmacology can currently provide us with, that doesn’t mean we are doomed to remain at the level of the individual effort while trying to accomplish this most noteworthy goal. And what could help us achieve such a goal—technology and science? Yes, but not biomedical.

TOMISLAV MILETIĆ

University of Rijeka, Rijeka, Croatia

References

- Agar, N. 2007. “Whereto Transhumanism? *The Literature Reaches a Critical Mass.*” *Hastings Center Report* 37 (3): 12–17.
- Brey, P. 2009. “Human Enhancement and Personal Identity.” In J. Kyrre, B. Olsen, E. Selinger, S. Riss.(eds.). *New Waves in Philosophy of Technology*. New York: Springer: 169–185.
- Crockett, M. J. 2014. “Moral bioenhancement: a neuroscientific perspective.” *Journal of Medical Ethics* 40 (6): 370–371.
- Dubljević, V. and Racine E. 2017. “Moral Enhancement Meets Normative and Empirical Reality: Assessing the Practical Feasibility of Moral Enhancement.” *Bioethics* 31 (5): 338–348.
- Helion, C. and Ochsner K. N. 2016. “The Role of Emotion Regulation in Moral Judgment”, *Neuroethics*. <https://doi.org/10.1007/s12152-016-9261-z>
- Prinz J. 2011. “Is Empathy necessary for Morality?” In A. Coplan and P. Goldie (eds.). *Empathy: Philosophical and Psychological Perspectives*. Oxford: Oxford University Press.
- Terbeck, S. and Chesterman, L. P. 2014. “Will There Ever Be a Drug with No or Negligible Side Effects? Evidence from Neuroscience.” *Neuroethics* 7 (2): 189–194.

Amy Kind and Peter Kung (eds.), *Knowledge Through Imagination*, Oxford: Oxford University Press, 2016, 251 pp.

Imagination has become a fashionable topic, and its role in procuring knowledge has become a central challenge in the analytical debate on imagination (see, for instance, the 2006 issue of *Metaphilosophy* under the same title as the present collection, *Knowledge through imagination*). The present collection offers a well-organized range of interesting and challenging contributions. They are divided into three groups, the first encompassing taxonomical and architectural issues (featuring papers by M. Balcerak Jackson, P. Langland-Hassan and N. Van Leeuwen), and the second offering “optimistic approaches” (T. Williamson, J. Jenkins Ichikawa, the co-editor A. Kind herself, and J. Church). The optimism is balanced in the third part, featuring “skeptical approaches” by H. Maibom, Sh. Spaulding and by the co-editor P. Kung. I shall choose a paper or two from each group, with apologies to the rest of the authors. (For quotations, I put page number in brackets.)

Let me start with the “Introduction” by the editors. They note that “the puzzle of imaginative use concerns two distinct and seemingly incompatible uses to which imagination is often put (1). Sometimes it is an escape *from* reality, and sometimes it is “used to enable us to learn about the world as it is, as when we plan or make decisions or make predictions about the future. But how can the same mental activity that allows us to fly completely free of reality also teach us something about it?” (1). How is the “instructive use” of imagination possible? The editors optimistically hope that a closer analysis will explain the joint possibility of the two uses, in particular the instructive one, and see the key to the explanation in constraints that thinkers-imaginers put upon their activity. The constraints come in two kinds. First, they “may be architectural; that is, they may result from our cognitive psychological architecture” (22). Second, the constraints may derive from more spontaneous sources, such as limitations that we voluntarily impose upon our imaginative projects (22).

Amy Kind develops these ideas further in her paper “Imagining Under Constraints”. She offers a characterization of imagining that involves a more active effort of mind than does supposition or entertaining a proposition (148), and quotes Kendall L. Walton’s (1990) classic *Mimesis as Make-Believe: On the Foundations of the Representational Arts*, Harvard University Press, suggesting that imagining “is doing something with a proposition one has in mind” Walton, p. 20 (148). She then proposes a conception of “ideal imagination” modelled on an entertaining science fictional story in which highly developed computing machines predict things in a cold, perfectly calculated way, marching step by step, with “irresistible steps”. They obey the ‘reality constraint’ in representing things, and the “change constraint”: “when their imaginative projects do require them to imagine a change to the world as they believe it to be, they are guided by the logical consequences of that change” (151). She then mentions Tesla and Temple Grandin as human quasi-ideal imaginers. Her conclusion is optimistic: “in modeling our imagination on the ideal imagination of the machines, we are able to make epistemic progress the way they do, by steady, irresistible steps” (159).

Other authors on the optimistic side take similar steps, specifying the constraints imposed upon imagination. Peter Langland-Hassan in his rich paper “On Choosing What to Imagine,” concentrates on imaginings that are voluntarily and suitable for guiding action and inference. He lists three essential components that guarantee the guiding power, first, the availability of (top-down) intentions to start imagining, second, of lateral constraints that govern the development of the imagining, and third, the possibility of cyclical interventions by subject and her intentions, in particular during a given imaginative episode (81).

In his contribution “Knowing by Imagining” Williamson joins the optimistic crew and proposes a cognitive view of imagination, without forgetting its practical value i.e. the importance of practical matters (124); he talks about “a wide range of possible ends” and possible practical evolutionary origin of imagination. Also, in his view fiction is not central for imagination, as he pointedly remarks in the concluding sentence of his paper: “... if we try to understand the imagination while taking for granted that fiction is its central or typical business, we go as badly wrong as we would if we tried to understand arms and legs while taking for granted that dancing is their central or typical business” (131).

Among cognitive function the prominent ones are raising possibilities and assessing the truth-values of propositions (115). This requires cognitive qualities, like rational responsiveness to evidence (116) and capacity to develop adequate scenarios: the imagination develops the scenario in a reality-oriented way, by default (116). Williamson does not call them epistemic virtues, but this is how a friend of virtue epistemology would describe them. They offer reliability: “...under suitable conditions, the method constitutes a reliable way of forming a true belief as to what would happen in hypothetical circumstances” (117).

Williamson wisely stresses similarities between various exercises of imagination, using them to suggest that most sophisticated among them, like thought experiments, are nothing special and mysterious. What about science? Williamson has a fine optimistic argument in favor of the serious epistemic status of imagination in it: “One might suppose that, as science progresses, the role of the imagination will increasingly be confined to the context of discovery, and that in the context of justification it will gradually be replaced by more rigorous methods. But there is evidence to the contrary. For rigorous science relies on mathematics, and so indirectly on the axioms or first principles of mathematics. But when one examines the justifications mathematicians give of their first principles, such as axioms of set theory, one finds unashamed appeals to the imagination” (123). He also stresses that thought experiments are part and parcel of the normal functioning of imagination: “We simply reserve the term ‘thought experiment’ for the more elaborate and eye-catching members of the kind.” So much for Williamson’s cognitive view of imagination in general.

The first issue that arises for the project is the classical philosophical one: what is imagination and what is the role of image in it? How close is it to belief? The term „cognitive” seems to suggest a very high degree of closeness; what about the differences? Take imagining a golden mountain: many people will stress the image in such an imagining, but how important is it

exactly. Williamson notes that many of his examples “appear to involve an essential role for mental imagery, in some sense” (118), but he quickly adds that “... we should not over-generalize to the conclusion that all imagining involves imagery” (118). And in fact, he presents the imaginative exercise differently, more as a matter of logic and even almost exclusively as a matter of logic and possibly quite sophisticated and complicated, with the full range of tableau methods in the foreground, continuing the venerable tradition of Jaako Hintikka interpreting Kant’s notion of *Anschauung* (in his 1969, “*On Kant’s Notion of Intuition (Anschauung)*”, in T. Penelhum and J. J. MacIntosh (eds.), *The First Critique*, Wadsworth Publishing).

On the other hand, here is how in his central example he presents the way people imagine. He invites us to think of a hunter who finds his way obstructed by a mountain stream rushing between the rocks (117). The hunter “imagines himself trying to jump the stream” (119) and presumably asks himself *If I try here, what is it going to be like?* Williamson notes that “he also has to look carefully at its banks in front of him, to tailor his imaginative exercise as exactly as he can to their actual contours” (119) But this tailoring of one’s imaginative exercise to the contours perceived sounds a bit like creating a visual-kinesthetic moving picture, a video: *it will be like this*. (This is what is often called a *mental model* of the situation, and here imagistic, video-like properties might help a lot.) So, even if we accept that image-producing is not a necessary feature of imagination, it could be a centrally important one, and the not image-involving cases might be a bit marginal. In general, judgments are easy to elicit with concrete examples. With naive subjects it is the *only* way. However, Williamson stresses the importance of deductive logic and the “tendency of imagination to use something like rules of deductive logic...” (123). He notes “the role of the imagination as a standard means for evaluating conditionals and modal claims (123). This raises the important issue of the role of logic in relation to imagistic cognition. Like Peter Langland-Hassan, Williamson wants to combine the two, and it will be interesting to see what the results in his subsequent work will be. So much about the optimists.

On the skeptical side, the most direct challenge to the project of finding constraints that would rehabilitate imagination is to be found in the paper by Shannon Spaulding: “Imagination Through Knowledge”. On her view, the puzzle of how we arrive to knowledge through imagination suggests that imagination is “not sufficient for new knowledge” (222). The argument seems to be the following: if imagination is to be constrained by extra-imaginative pieces of information and by other abilities, then imagination does not bring new knowledge. But this is too severe a demand. Compare physical constraints. I commute from my home town to my working place about hundred miles distance. For the car to bring me to my work there should be a well-established and well-kept road, constraining the travel, there should be red lights helping to prevent crashes, and so on. Imagine someone arguing that therefore “car is not sufficient” for commuting, and is not doing any real work! Well, the fact that an item needs constraints to function properly does not entail that it never performs any function.

Spaulding has an auxiliary argument: “I have argued that the cognitive capacity to imagine scenarios is distinct from the cognitive capacities that

underlie our ability to judge the accuracy of our imaginings” (222) and “... there is nothing in the capacity of imagination itself that could evaluate the accuracy of the possibilities we imagine” (222). Indeed, there is nothing in the car itself that recognizes red/green light. This does not show that the car will not take me from home to work, only that car *alone* will not do the work. So much about Spaulding’s direct challenge to the instructive use of imagination.

Let me mention, however, that in her text the challenge is preceded by a rich and very provocative analysis of one particular kind of imaginational enactment, namely simulation. Her argument resembles the general one we just summarized. Her example is the following: I watch John tease Mary, and try to figure out why he is doing this. I simulate his activity, and end up concluding that John likes Mary and is trying to get her attention. Fine, but how do I choose this option rather than some other, equally plausible in itself, for instance that he is just humiliating her? I need additional information, and my simulation tells me nothing about these matters. Again, to me it looks like simulation has done the main job, like the car in our example; the fact that the main job cannot be fully accomplished by the main agency in question, tells little against it.

Heidi Maibom’s paper “Knowing Me, Knowing You: Failure to Forecast and the Empathic Imagination” joins in with bad news about people’s abilities to recognize their own characteristics and attitudes, and abilities to project items of self-knowledge onto their neighbors.

Peter Kung’s “Thought Experiments in Ethics” is not so generally pessimistic as the papers by the two preceding authors. He just warns us that typical ethical thought experiments, especially ones that are meant to produce counterexamples to crucial ethical claims, CTEs for short, are organized around sharp, binary division, offering “forced choices fixed outcomes”: would you pull the lever, and kill three people, but save five, or not? He develops his criticism in a reach and subtle way, connecting it with issues of imagistic (he calls it “pictorial”) vs. non-imagistic representations, with topics of modality and so on. He claims that “imagining CTEs gives us *no reason* to believe that forced choices with fixed outcomes are genuine possibilities” (228, italics mine). We should use more realistic scenarios in our thought-experiments.

Let me note that real life often does offer “forced choices with fixed outcomes”: “Would you marry the person you are so passionately attracted to, but whom you realize to be a very dangerous partner, or not?”, “Would you vote for Trump, for Clinton or for Saunders, or not vote at all?” So ethicists might hope to offer some answers to people facing such choices, and they might prepare themselves by going through imaginary exercises featuring them.

Let me conclude that the optimistic side might have chances to survive. And let me add the following: if we accept that imagination follows real-world (or quasi-real-world) constraints, the question arises where the representations of the constraints come from. One possible unitary answer is that thinkers have mental models of reality, and that, when they ask themselves an instruction-oriented question, the models available to them constrain their subsequent imagination. If the result is worth remembering

and taking into account, it can be integrated back into one of the models, so that in the future it will provide a relevant “lateral constraint” to some exercise of imagination. If we assume that imagination is typically imagistic, and that mental models are typically concrete and “iconic”, but that both allow for thought processes that range from more iconic-pictorial to more digital deductive ones, then we shall notice that the two media, imaginative and model-sustaining one, nicely fit together and can interact in a non-problematic way.

NENAD MIŠČEVIĆ

University of Maribor, Maribor, Slovenia
Central European University, Budapest, Hungary

Bojan Borstner and Smiljana Gartner (eds.), *Thought Experiments between Nature and Society: A Festschrift for Nenad Mišćević*, Newcastle upon Tyne: Cambridge Scholars Publishing, 2017, xxxviii + 437 pp.

This volume is a festschrift dedicated to Nenad Mišćević, well-known Croatian philosopher, for the occasion of his 65th birthday. During his years in philosophy, Mišćević engaged almost all areas of philosophy. So, since thought experiments, according to some people, lie in the foundation of all the disciplines and subdisciplines of philosophy as an indispensable foundational reflective tool, and could be, at the same time, a philosophical problem of their own (well, everything, “everything”, “everything” can be a philosophical problem), it seemed appropriate to take them as the central theme of this celebration volume.

The book consists, beside Introduction by the editors, the personal account of Mišćević by Bojan Borstner and Tadej Todorović, and the Mišćević’s own account of his views on thought experiments, of 22 chapters and each chapter has Mišćević’s reply. Contributors to the volume are (in order of appearance): Timothy Williamson “From Anti-Metaphysics to Metaphysics”, Howard Robinson “Intuitions and Thought Experiments”, Maja Malec and Olga Markič “Mišćević on Intuitions and Thought Experiments”, Nenad Smokrović “Curiosity and the Argumentative Process”, Peter Gärdenfors “Sematic Transformations”, Danilo Šuster “Lucky Math: Anti-luck Epistemology and Necessary Truth”, Guido Melchior “Epistemic Luck and Logical Necessities: Armchair Luck Revisited”, Smiljana Gartner “Did a Particularist Kill the Thought Experiment?”, Marian David “Experimental Philosophy, Gettier-Cases and Pragmatic Projection”, Peter Simons “Concepts in a World of Particulars”, İlhan Inan “Is the Speed of Light Knowable A Priori?”, Andrej Ule “Mental Models in Scientific Work”, Ferenc Huoranszki “Natural Kinds and Conceptual Truth”, Majda Trobok “Grasping the Basic Arithmetical Concepts: the Role of Imaginative Intuitions”, Andraž Stožer and Janez Bregant “The Colour Dilemma: A Subjectivist Answer”, Matjaž Potrč “Dasain in a Vat”, Pierre Jacob “Knowing One’s Own Mind” (some real history instead of thought experiment: Balkan wars were fought 1912–1913 and Mišćević was not born then, so he could not be a victim of these wars.), Friderik Klampfer “The False Promise of Thought-Experimentation

in *Moral and Political Philosophy*”, Miomir Matulović “*Miščević, Mental Models, and Thought Experiments in Political Philosophy*”, Boran Berčić “*Are Nations Social Constructs? Nenad on Nations*”, Rudi Kotnik “*Thought Experiments in Teaching: TE as a Suppositional Real Story*”, and Boris Vezjak “*The Ring of Gyges and the Philosophical Imagination*”.

The articles are grouped under three main headings—the first deals with general problems about thought experiments, the second deals mostly with the relation of the thought experiments and the (science and metaphysical structure of the) world; the third concentrates on thought experiments in the philosophy of mind, philosophy of politics, morality and society. But subdisciplines of philosophy emerge in each of the three parts. Some of the articles deal more about some other particular problem which Miščević discusses in his numerous works, rather than exactly the thought experiments or intuitions.

Of course, it is not possible, in a short review, to give even an elementary justice to such a volume which contains many good and new ideas, arguments and well-supported theories; and to each chapter, so I have chosen just several chapters for more detailed exposition (so it is a subjective choice).

Miščević, in his overview “*Accounting for Thought Experiments—25 Years Later*” characterises thought experiment (13) as an “*armchair*” reflexion which involves “*experimental design*” for a theory which is to be tested, the construction of a counterfactual scenario and its careful presentation, thinking and reflecting carefully about the presented scenario and, finally, “*the decision*” about the theory that is tested. This “*decision*” is intuition of the experimental subject (it can be the author of the thought experiment himself, or an interlocutor), and it is usually compared with some relevant similar other thought experiments. So, thought experiments are performed only cognitively, “*in the laboratory of mind*,” to use James Brown’s characterisation (17). They often include visual imagination, but what is important in the end—to confirm or disconfirm the theory which is tested—should be careful reasoning about the scenario and the theory, though intuitions elicited are more scenario-based than inference based (26). Miščević further develops some details about where to place thought experiments in the wider theoretical picture and then develops some specifics of thought experiments—their phenomenology, the characterisations of mental models building and engages experimental philosophy which challenged the use of thought experiments. Miščević calls his proposal, which aims at characterisation and explanation of the structure and role of thought experiments and intuitions, “*Moderate Voice of Competence View*” (26). Briefly, according to this model, distinct group of phenomena is made by intuitions-dispositions and judgements; there is a psychological capacity to use imaginative and judgemental competencies so we get intuitional data which do not involve theory and contain only just a small amount of proto-theory. For Miščević, concepts are not the proper objects of intuitions; they are only subordinated in their role to the main function of intuition which is aimed toward external objects, items and facts (26).

Howard Robinson in his article expresses scepticism about the closely related notions of “*thought experiment*” and “*intuition*”—about their usefulness in philosophy. He uses the term “*revolution*” to illustrate the point.

Many various events are called revolutions, but only one property is common to them says Robinson—they are radical changes. Beside this, each particular case (of revolution) is for a discussion of its own, if we would like to say really important and significant matters about each of them. Precisely this is transferred to “intuitions” and “thought experiments”. Robinson (51) gives this definition for “intuition”: “A belief is intuitive when the grounds for holding it are either not dependent on the kind of reasoning, or publically available evidence, which are normally regarded as necessary for a rational belief, or go beyond what available evidential considerations of a more public kind would strictly justify”; and for “thought experiment” (53): “A thought experiment envisages a situation meant to throw light on a philosophical problem where, whether that situation actually obtains or not, is held not to be relevant to its ability to illuminate the issue.” Nothing else is generally important for these two notions—each case is on its own, with its content and details, for relevant discussion. So, after exposing a certain number of examples of “intuitively plausible or implausible cases” and thought experiments across semantics, problem of personal identity, philosophy of mind, epistemology and ethics, Robinson concludes that we should be very sceptical about discussing “intuitions” and “thought experiments” as that they are themselves a philosophical problem.

Smiljana Gartner questions the applicability of thought experiments in ethical contexts. It is possible to conceive a thought experiment as a certain ethically relevant situation and then to change only slightly the properties of that situation, but changes in attitudes toward the thought experiment, adding just these slight changes, could be, and sometimes are, dramatic; sometimes we can go back and forth even with contrary or contradictory attitudes what should we do in such situations. It seems that the condition of stability is not often satisfied concerning thought experiments in ethics. Gartner concludes that if we use thought experiments in ethics, we should be extremely careful and precise.

Peter Simons argues in his contribution, that there are no concepts and meanings as abstract objects. For Simons, there are only particulars and collections of them. Moreover, general concepts as well as singular concepts, fall to the same constraints if we explain them nominalistically. To have such nominalist explanation of the concepts, their use and understanding, we have to identify the collection of particulars that revolve around them (the main concrete example is the concept “horse”). These are: users, words, other external representations, acts, activities, capacities, compliants. Though interrelations between them are complex and sometimes very complicated, still we can find them and all these are, according to Simons, identifiable as concrete entities.

Boris Vezjak, in his article, challenges the idea that Plato offers a “thought experiment” in his *Republic*, as is claimed by Mišević, in the story of the myth of Gyges, and his objections are fourfold, so there are: general methodological objection, motivational objection, structural objection, and interpretative objection. Vezjak attempts to show by these considerations that Plato’s telling myth does not have relevant properties to be classified as a thought experiment as we today conceive what thought experiment is.

This Festschrift presents many different pieces of excellent philosophical work for further study and discussion. So, take a real experiment—take this book and read whatever interests you and find out Mišćević's answers to articles particularly mentioned here and, as well, for all the others. We can praise editors for their immense work done.

DAVOR PEĆNJAK
Institute of Philosophy, Zagreb, Croatia

Boran Berčić (ed.), *Perspectives on the Self*, Rijeka: Faculty of Humanities and Social Sciences, 2017, 375 pp.

The collection *Perspectives on the Self* brings together seventeen essays which explore the notion of the Self. Employing both historical and conceptual analyses of the Self, the authors cover a variety of topics from research areas that include metaphysics, philosophy of mind, philosophy of science, philosophy of language, ethics and history of philosophy. The book, published by the University of Rijeka, is a result of a conference, *The Self*, which took place at the Faculty of Humanities and Social Sciences in Rijeka (Croatia) on March 31 and April 1, 2016. As noted in the preface by the editor Boran Berčić, Full Professor at the Rijeka Department of Philosophy, the participants of the conference, whose essays make up the volume, are in different ways involved in the research project *Identity* at the University of Rijeka. Those include both Croatian and foreign philosophers along with the reviewers of the book, Nenad Smokrović and Dušan Dožudić.

The book is divided into six chapters. The first chapter, titled “Self and Body”, starts with “The Central Dogma of Transhumanism” by Eric T. Olson (University of Sheffield). Olson argues against the transhumanist claim that it is metaphysically possible to upload our psychological selves into a digital computer. He identifies the transhumanist claim as resting on a metaphysical assumption that we are essentially patterns (the *pattern view*) which can be transmitted as information. He then confronts the claim by insisting that we are essentially material things (more specifically—biological organisms), not patterns, and as such cannot be “detached” from our biological substrate and transferred into a computer. He also considers the so-called *constitutional view* and the *temporal parts view* but concludes that they cannot serve the transhumanist's purposes.

Miljana Milojević (University of Belgrade) in “Embodied and Extended Self” combines a functionalist ontology of the self with an embodied and extended view on the mind. She starts by accepting the psychological-continuity criterion of personal identity (Parfit). She then casts it in a realizier-functionalist ontology which, Milojević believes, allows for an embodied view on the mind for which she finds justification in the works of Gallagher, Shapiro and others. Finally, she uses multiple realizability of the mental to extend the self beyond the boundaries of the organism.

Zdenka Brzović (University of Rijeka) in “The Immunological Self” surveys a number of possible identity criteria for a biological organism (functional integration, autonomy, genetics). After showing their flaws, Brzović shifts her analysis to different versions of the *immunology criterion*. She discusses the *self-nonself theory* (Burnet), several versions of the *systematic*

theories of immunity—the self as an *autopoietic* entity (Maturana and Varela, Jerne), the *danger theory* (Matzinger) and the *continuity theory* (Praudu)—but concludes that all of them share a problem of presupposing the identity of the organism, and thus cannot serve as the criterion of identity.

The second chapter of the book, titled “Self-Knowledge” starts with “The Value of Self-Knowledge” by Nenad Mišćević (University of Maribor). Mišćević starts by drawing a distinction between two kinds of self-knowledge—knowledge of one’s inner phenomenal states (such as knowing that my back is in pain) and knowledge of one’s causal and dispositional properties (knowing that I am easily frustrated or impatient). Mišćević then turns to the question of their intrinsic and instrumental value. He argues against the claim that our knowledge of our phenomenal states has no instrumental value (Cassam). In addition, he insists that it also has an enormous intrinsic value for our conception of the self. Also, following Lehrer, he defends the instrumental value of knowledge of one’s causal and dispositional properties as a prerequisite for a wise life.

In “The Self-ascription of Conscious Experiences” Luca Malatesti (University of Rijeka) wonders how do we make the step from experiencing X (Malatesti uses color perception) to knowing (consciously) that we are experiencing X. He analyses two models of self-awareness (Armstrong’s quasi-perceptual model and Moore’s transparency of experience) but finds them both unsatisfactory. He concludes by offering a conception of the self which he believes to be a prerequisite for the possibility of conscious self-ascription of experiences. Here he follows Millar and claims that the concept of the self “involves the capacity to think about ourselves as entities that have sense organs and internal states that are determined by interactions with certain sorts of stimulation of these sense organs” (135).

The third chapter, “Self in the History of Philosophy” starts with “The Logical Positivists on the Self” by Boran Berčić (University of Rijeka). After analysing the logical positivists’ (Shlick, Ayer, Carnap, Weinberg, Reichenbach) critique of the Cartesian *Cogito*, Berčić shifts our attention to various ways in which the positivists understood the self. He draws a distinction between *conceptual*, *epistemological* and *ontological* reductionism about the self and concludes that, although the positivists were reductionist in all three senses, their reductionism should be understood primarily in its epistemological sense, meaning that we come to know about the self only *a posteriori*, that is, when we know what its elements are.

Ljudevit Hanžek (University of Split) in “Brentano on Self Consciousness” examines a theory of self-consciousness by Franz Brentano, as put forward by Brentano in his *Psychology from an Empirical Stand-point* (1874). In order to avoid an infinite regress of mental states, Brentano argued that in addition to being aware of an object, mental states possess a certain kind of self-awareness. While analysing arguments *pro et contra* Brentano’s views, Hanžek considers a number of similar proposals—Uriah Kriegel’s theory which rests on the distinction between focal and peripheral awareness; transitive and intransitive awareness (Kriegel, Gennaro, Rosenthal) and Amie Thomasson’s adverbial theory—but finds them all unsatisfactory.

Goran Kardaš (University of Zagreb) in the “No-Self View in Buddhist Philosophy” analyses the Buddhist claim that there exists no such thing as the self. Buddha believed that the self is an illusion rooted in bad cog-

nitive mechanisms and linguistic practices. Kardaš surveys a number of arguments made by Buddha and his followers which were directed against earlier Indian metaphysicians who held a substantialist position on the self. In doing so, Kardaš reveals the empiricism, reductionism and eliminativism which is present in the Buddhist view of the self.

In “The Self in Ancient Philosophy” Ana Gavran Miloš (University of Rijeka) argues against the claim that ancient Greeks didn’t possess the idea of subjectivity (Gill). She draws on the texts by Plato, Aristotle and Epicurus and concludes that, although the Ancients didn’t possess the modern, Cartesian conception of the self—understood “in terms of epistemic certainty and primacy of the pure subjective self-consciousness” (212)—their ontological and (especially) ethical deliberations presuppose a subjective, first-personal point of view and a subjectivity/objectivity distinction.

The first essay in the fourth chapter, titled “Self as Agent” is “Ideal Self in Non-Ideal Circumstances” by Matej Sušnik (University of Rijeka). As an internalist about the (normative) reasons of one’s actions, Sušnik analyses three suggested answers on the question about the relation between the real and the ideal self. He rejects the first two (the *straight-forward model* and the *advice model*) and accepts the third, according to which an agent has a reason to do x only if there exists a “sound deliberative route” from the agent’s actual motives to his doing x (Williams).

Filip Čeč (University of Rijeka) in “The Disappearing Agent” analyses the *disappearing agent objection* (Pereboom) which is directed against the *event causal* version of the libertarian position about free will. According to the objection, the event-causal ontological framework (based on events and states) doesn’t secure the agent’s role in the decision-making process, especially in the so-called *torn decisions* which figure prominently in several event-causal accounts (Kane, Balaguer, Franklin). Torn decisions involve indeterminism in the decision-making process which seems to undermine the agents role in it. Čeč analyses five possible ways in which an event causal libertarian might respond. He concludes by choosing the last option which claims that the event-causal libertarian can secure the agents role in spite of there being some indeterminacy in the decision-making process.

Marko Jurjako (University of Rijeka) in “Agency and Reductionism about the Self” starts with an analysis of the psychological-continuity criterion of personal identity developed by Parfit (Parfit 1984, 1995). He then presents some of the problems which seem to follow from the reductionism entailed by Parfit’s account, especially those related to our moral and prudential concerns. Both of these seem to presuppose a “deep unity” underlying our personal identity and that unity is what Parfit’s account seems to eliminate. Jurjako finds the agency based accounts of the self (Korsgaard, Bratman) capable of meeting these problems. He believes them to be compatible with the reductionist view and argues that their focus on our ability to act and deliberate as the source of personal identity provides the unity needed to vindicate our practical concerns.

The fifth chapter, titled “The Non-existent Self” starts with “On never Been Born” by Marin Biondić (University of Rijeka). Biondić wonders whether we can make meaningful judgements about people who were never born. For example, can we feel sorry about those who were never born?

Biondić follows Parfit and concludes that we can not. Here he relies on the view (also by Parfit) that we can evaluate (in terms of good or bad) only lives of actual people.

Iris Vidmar (University of Rijeka) in “Fictional Characters” takes the “literary aesthetics” approach to the nature of fictional characters which focuses on “the way fictional characters come to life within established literary practices”. Within it, Vidmar discusses the ways fictional characters are brought into existence and what makes up their identity. As the main protagonist of her analysis, Vidmar takes Flaubert’s *Emma Bovary*. She concludes that the identity of fictional characters is multilayered and relational in nature and discusses some of the objections to her view.

The sixth and final chapter, titled “Metaphysics and Philosophy of Language” starts with “Haecceity Today and with Duns Scotus” by Márta Újvári (Corvinus University of Budapest). She discusses the historical and contemporary understanding of haecceity. Traditionally, haecceity was understood as an entity, while contemporary authors see it as a “relational property of being identical with itself” (332). Relying on Chisholm and Scotus, Újvári analyses several contemporary views (Rosenkratz, Diekemper, Gracia) which offer an account for the notion of haecceity but finds them all unsatisfying.

Arto Mutanen (Finnish National Defence University) in “Who Am I?” gives an analysis of the same question. He finds it to be a cluster-question with two possible answers. One is about identification, the other about identity. He then expands on the distinction between identification and identity, which he analyses by relying on Gleason, Hintikka, Quine and Kripke. He finds identification to be a methodological notion concerned with determining *who* somebody is or what their *location in society* is (Gleason). On the other hand, identity is an ontological notion concerned with determining *what* kind of entity one is. Mutanen finds Descartes’ dualism to be the prime example of an answer to an identity question.

The last essay in the collection is “Meta-Representational *Me*” by Takashi Yagisawa (California State University). He starts by differentiating between the notion of *me* (which applies absolutely) from the notion of the *self* (which he believes to be relative). He then proceeds in developing an account of the notion of *me*. In doing so, he analyses the standard indexical theory, but finds it incomplete and offers his own theory which relies on the “way-to-thing shift” strategy. The theory claims that we represent the world in a certain first-person way (Yagisawa calls it *me-way*). The *me-way* of representing enables me to pick myself as the recipient of that representation. Only then, Yagisawa believes, I come to postulate myself as an entity, as *me*.

Each of the seventeen essays found in the collection *Perspectives on the Self* makes a valuable contribution to contemporary explorations and discussions about the notion of the Self. Particularly significant is the fact that the collection brought together a team of international authors, alongside with philosophers working in Croatia. Coming from different areas of philosophical interest, the authors surveyed and analysed a variety of contemporary and historical arguments, theories, traditions and problems related to the notion of the self. In doing so, they covered a wide range of topics related

to the self, such as identity, agency and mind. This volume will primarily be of interest to professional philosophers, psychologists, graduate students in philosophy, and other scientists interested in the philosophical themes related to the self, but with the help of a detailed and admirably clear introduction by the editor, it can perhaps even be accessible to the general, intellectually curious, public. We can conclude that the *Perspectives on the Self* is a successful exercise in contemporary analytic philosophy which brings valuable insights into areas of epistemology, metaphysics, ethics and history of philosophy, and is a major contribution to the philosophical literature published in Croatia.

MARKO DELIĆ
University of Split, Split, Croatia

Croatian Journal of Philosophy is published three times a year. It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in *The Philosopher's Index*, *PhilPapers*, *Scopus*, *ERIH PLUS* and in *Arts & Humanities Citation Index (Web of Science)*.

Payment may be made by bank transfer

SWIFT PBZGHR2X

IBAN HR4723400091100096268

Croatian Journal of Philosophy is published with the support of the Ministry of Science and Education of the Republic of Croatia.

Instructions for Contributors

All submissions should be sent to the e-mail: cjp@ifzg.hr. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

ISSN 1333-1108



9 771333 110001