Hypotheses fingo.
Epistemological considerations on philosophical thought experiments

Abstract

The present dissertation focuses on the notion of philosophical thought experiment. In particular, the work aims at assessing a number of fundamental logico-epistemological features which seem to characterize both their internal structure and their possible use. The first chapter intends to show the philosophical relevance of the central epistemological question raised by scientific thought experiments. The second chapter aims at drawing attention on the central role played by our intuitions in the execution of philosophical thought experiments. The third chapter, finally, delves into logical aspects of the argumentative structure of most philosophical thought experiments and puts forward a proposal concerning the kind of reasoning they instantiate.

Abstract

Oggetto primario della presente tesi è la nozione di esperimento mentale in filosofia. Il lavoro si propone in particolare di mettere in luce una serie di aspetti logico-epistemologici fondamentali, che sembrano caratterizzarne tanto la struttura interna, quanto il possibile utilizzo. Il primo capitolo intende far emergere la rilevanza filosofica della questione epistemologica fondamentale sollevata dagli esperimenti mentali scientifici. Il secondo capitolo si sofferma sul ruolo centrale svolto dall'intuizione nell'esecuzione di esperimenti mentali filosofici. Il terzo capitolo, infine, è dedicato a un approfondimento della struttura logica fondamentale di gran parte degli esperimenti mentali filosofici e avanza una proposta relativa al tipo di ragionamento che in essi ha luogo.

Acknowledgements

Over the last four years, several people have valuably contributed, in one way or another, to the completion of the present dissertation. I would especially like to thank Guido Cusinato and Ferdinando Marcolungo of the University of Verona, Vincenzo Fano of the University of Urbino, Achille Varzi and John Collins of Columbia University. During the same period of time, I have also immensely benefited from illuminating conversations with, as well as thought-provoking criticisms by, Pierluigi Graziani, Claudio Calosi, Marco Nathan, Daniele Santoro, Patrizia Pedrini and Linda Selmin, to all of whom I am extremely grateful. I would finally like to express my gratitude to the University of Verona for founding my four-semester stay as a visiting scholar at Columbia University, and to Patricia Kitcher, of Columbia University, for making it possible.

Contents

Introduction

"*There are more things in heaven and earth, Horatio,*
than are dreamt of in your philosophy"

Hamlet Act 1, scene 5.

If he could look down into our restless world from the heaven of poets, William Shakespeare would probably be surprised to find out how far down the path of history his verses have echoed. "My own suspicion", once wrote evolutionary biologist J.B.S. Haldane, "is that the universe is not only queerer than we suppose, but queerer than we *can* suppose"[1]. Half a century later, Richard Dawkins' preface to the first edition of his ground-breaking work *The Selfish Gene*, opens with the following words: "This book should be read as though it were science fiction. It is designed to appeal to *imagination*. But it is not science fiction: it is science. Cliché or not, 'stranger than fiction' expresses exactly how I feel about the truth"[2].

The belief that reality could have been different from what it actually is, it seems, has been around for a while in the Western world. In what is now the fifth book of his *Metaphysics*, Aristotle wrote: "We say that that which cannot be otherwise is necessarily as it is"[3]. This definition implicitly credits our species with a remarkable ability to imagine that things 'could have been otherwise'. As soon as we acknowledge the finite nature of our cognitive abilities, the following two hard facts are immediately put in front of us as inquirers: while human beings, on the one hand, often imagine things which *do not* exist, their intellectual powers, on the other hand, have just as often been far from sufficient to imagine the existence of things which *do* in fact exist. In our relentless effort to understand reality, that is, imagination has often proved to be a very deceptive ally. And yet very few people, I believe, would be willing to deny that *imagination* plays a fundamental role in human *understanding*.

[1] Haldane (1927: 286).
[2] Dawkins (2006 [1976]: xxi). My emphasis.
[3] Aristotle, Metaphysics 1015b. Translated by W. D. Ross.

The present work originated from a personal interest in what in the last chapter I have proposed to call *the boundaries of heuristic fertility of our imagination*. While indeed the very existence of such boundaries, on the one hand, seems to most of us just an obvious fact, any attempt at charting their exact topography, on the other, strikes us just as obviously as a nearly impossible task. Starting from such premises, my attention was naturally drawn to the almost ubiquitous use of *thought experiments* in contemporary analytic philosophy.

As a matter of fact, despite their popularity among philosophers, I could not help but finding this philosophical practice deeply disappointing. Annoyed by my insistently professed inability to understand the "wonderful" piece of modern art that he was trying to draw my attention to, an art historian friend once spoke up his frustration by impatiently telling me that all I really needed to understand a painting was a chair. Now, be that as it may for the case of modern art, the idea that that very same chair, or a more comfortable version of it for that matter, say an armchair, could also be sufficient to understand the world we live in, seemed to me as utterly wrong minded, if not slightly pathological. And yet, over the last few decades, thought experiments have undeniably become an apparently fundamental item within the bag of tools of most analytic philosophers, and I felt the need to inquire more carefully into their nature.

As a matter of fact, starting at least from the early seventies, an abundant use of thought-experimental reasoning has characterized a considerably large number of philosophical debates in areas as far apart as ethics, philosophy of language, philosophy of mind and epistemology. As a consequence, starting from the early eighties, a consistent amount of attention has been directed by several authors both to the inner workings and to the general purposes of these strange philosophical creatures. Their reflections on the topic have soon generated a complex and very lively debate which touches upon various philosophically relevant issues. My considerations revolve around that debate and try to focus on the aspects of it that I find most relevant from an epistemological point of view.

Although the work focuses primarily on *philosophical* thought experiments, it traces some of the fundamental epistemological problems they raise back to *scientific* thought experiments, and begins therefore by introducing this latter family of mental procedures. This choice was also intended to stress and to help appreciating better the close resemblance that the two kinds of thought experiments bear to one another. As I hope will emerge in the next three chapters, I mainly think of philosophical thought experiments as of a sort of *tools* available to the philosopher to put his intellectual house in order, as someone has written.

In my effort to understand their nature, and the limits of their fruitful use, I like to think of myself as of an inquisitive naturalist, who is trying to come to grips with a very puzzling feature of human cognitive life. Our ability to perform philosophical thought experiments, I believe, is

largely dependent on our non-philosophical ability to transcend the actual in thought by reflecting on hypothetical states of affairs, which seems to appear quite early in our cognitive lives. As a consequence, I think that the more or less useful or successful performing of philosophical thought experiments will depend crucially on the more or less careful reliance on this ability, to which the bulk of the third chapter will be devoted. In trying to present the main lines of the debate mentioned above, many positions which are certainly worth of investigation were unavoidably left out. I am well aware of the fact that this choice, if not argued for, runs the risk of appearing exceedingly idiosyncratic. As a promissory note in my defence, I can only say that the reasons of my choice will become clearer along the way.

In a somewhat collateral way, this work intends to contribute to a larger methodological reflection on philosophical activity. In this I share company with Richard Foley, who, borrowing a line from one of my favourite poets, Robert Frost, has suggested to envision philosophy as "a momentary stay against confusion"[4]. The fundamental idea, he goes on to say, is that "one is doing the best he can, given one's present tools […] to provide a coherent intellectual picture"[5].

As I hope will become clear in what follows, I am also very sympathetic to the position advocated by one of the pioneers of the debate I mentioned earlier, namely Roy Sorensen, who has envisioned thought experiments as instances of what he has called a *cleansing model* of armchair inquiry, according to which the ultimate end of this practice would be that of making us more rational by detecting various forms of inconsistencies contained in our beliefs.

An important side effect of this activity, I contend, is that of enhancing our understanding of what each one of us chooses to call *reality*. Although some might find it intellectually disturbing, this last formulation is meant to suggest that I do not mean to address head-on the centuries old debate between *realists* and *antirealists*. I believe that, insofar as thought experiments contribute to a better understanding of our conceptual framework, they make us aware of the more or less tacitly held folk theories relying on which we navigate the world, and this, in turn, indirectly enhances our understanding of it. The emphasis Thomas Kuhn has placed on the tight link between nature and our conceptual apparatus, I think, is on the right track when one is trying to assess the cognitive powers of thought experiments.

Here, in short, is the plan of the work. The first chapter aims at assessing the philosophical relevance of thought experiments in the natural sciences. It tries, in other words, to address the question as to why philosophers should be interested in scientific thought experiments. In order to provide a rough idea of the kind of reasoning involved in a thought experiment, a few

[4] Foley (1998: 247).
[5] Foley (1998: 247).

paradigmatic instances of such curious devices are presented. These fictional narratives, it is argued, are bound to raise a puzzling epistemological question, namely: How can a scientific thought experiment manage to be informative about the world we live in without adding new empirical data by means of actual experimenting? Repeated attempts at addressing this question, it is observed, have generated over the last decades what I suggest to look at as various different *epistemologies* of thought experiments. Consequently, a brief survey is provided of what I take to be the most influential positions that have been advocated on this topic. The main purpose of the overview is that of 'measuring the price' of maintaining the central claims of each different position. In the light of the philosophical debate generated by the different stances, an attempt is made at assessing the tenability of a widely spread empiricist view which tends to commit science to a strictly *a posteriori* methodology. The extensive use of thought experiments made by several modern scientists, it is claimed, suggests the plausibility of a less drastic divide between scientific and philosophical investigations, insofar as it seems to extend the jurisdiction of philosophy beyond the scope of pure conceptual analysis, while at the same time proving the effectiveness of armchair methods within empirical science itself. A further intent of the chapter is to show how this search for a viable epistemology of thought experiments has contributed to make the existence of a purportedly sharp divide between a context of *discovery* and a context of *justification* appear less obvious.

In the second chapter, I move on to consider philosophical thought experiments, which I take to be one of the most striking methodological features of the new mainstream in analytic philosophy, which followed the gradual waning of the logical empiricist program and started a new season of theorizing in many areas of philosophy. Following the lead of the first chapter, I begin by introducing a few examples of what I think may be regarded as paradigmatic instances of philosophical thought experiments. The main intent is that of drawing attention to the fact that these argumentative strategies rely heavily on *intuitions*, thereby proving to consider them as essentially reliable epistemic sources. I argue that skepticism about thought experiments is generally and ultimately amenable to skepticism about intuitions, and I therefore try to inquire into these psychological states by considering some of the proposals that have been advanced in the literature concerning their nature and their possible uses. In what follows, I analyse a typical use of intuitions by focusing on what many have taken to be the canonical example of intuition-based philosophical methodology, namely *conceptual analysis*. As traditionally practiced, conceptually analysis presupposes a *classical theory of concepts*, according which concepts would possess a specifiable definitional structure that expresses the necessary and sufficient conditions a particular object has to satisfy in order to fall under a given concept. Nonetheless, as

I try to argue, a more careful consideration of its internal structure suggests that conceptual analysis should rather presuppose a *theory theory of concepts*, according to which conceptual categorization is a process that strongly resembles scientific theorizing. Conceptual analysis, I contend, should therefore be regarded and carried out as a continuous process whose two poles would be concepts or theories, on the one hand, and intuitions, on the other. In the final part of the chapter I point to the fact that recent work in cognitive psychology seems to offer strong empirical support to previous concerns about the general reliability of our intuitions. The relevance of these empirical findings to philosophical analysis, I argue, while having been questioned on the basis of important epistemological considerations, can hardly be denied. The challenge for philosophical methodology, I suggest, is that of taking into account the results produced by cognitive psychology while at the same time preserving its normative role.

The third and final chapter of my work approaches philosophical thought experiments from a different, although I believe complementary, angle. Temporarily departing from the *psychological* considerations of the previous chapter, I move on to consider some broadly *logical* aspects of their inner workings. The chapter should be regarded, in general, as a critical attempt at singling out the most salient features of what, following others, I take to be the typical inferential structure of most philosophical thought experiments. Human beings, as I mentioned above, can be credited with a fascinating ability to transcend the actual in thought by mentally entertaining hypothetical states of affairs. Insofar as philosophical thought experiments make large and systematic use of this ability, the kind of reasoning they instantiate cannot but be regarded as inherently *modal*. An assessment of the potential epistemic virtues of a philosophical thought experiment, in particular, as I try to argue at some length, should start by acknowledging the fact that claims concerning the *way* in which a certain proposition *holds* or a certain state of affairs *obtains*, ought to be thought of as always *relative* to a set of truths that we, more or less consciously, decide to keep, or at least treat as, fixed. It follows that the purported outcome of any single thought experiment, as I try to show by means of an example, will carry a very different epistemic weight according to the way in which we decide to interpret the modal notion featuring in it. In what follows, as an attempt to delve deeper into their typical argumentative structure, I consider two regimentation attempts that have been recently put forward in the literature an try to assess their merits. The regimentation proposal that I find more promising is subsequently applied to one of the thought experiments introduced in the second chapter and showed, by means of additional examples, to be a very useful analytical tool. A further fundamental feature of thought-experimental reasoning is subsequently taken into consideration, namely their appeal to *counterfactual* scenarios, and attention is drawn to the fact that

counterfactual conditionals raise a justification problem the solution of which is neither purely *a priori* nor purely *a posteriori*. In the last part of the chapter I try to assess the epistemological pros and cons of envisioning philosophical thought experiments as forms of *defeasible reasoning*, i.e. as special instances of non deductive inferential patterns. In particular, I consider a new classification of inference rules that has recently been proposed within the literature on non-monotonic reasoning and I suggest that it might shed new light on philosophical thought experiments.

In the conclusions I try to take stock of the considerations developed in the previous chapters and to make place for them into as coherent a picture as I am capable of.

1. Thought experiments in the natural sciences

1.1 *A shift of interest*

Historically, the first use of the expression 'thought experiment' is associated with the name of the Danish physicist and chemist Hans Christian Oersted (1777-1851), who has been credited for having introduced the term *Gedankenexperiment* into the philosophical and scientific discourse of the Nineteenth century[6]. It is nonetheless fairly uncontroversial that the first extensive enquiry concerning the nature and purpose of thought experiments in the natural sciences is due to the Austrian physicist and philosopher Ernst Mach (1838-1916)[7]. Mach, as a matter of fact, was amongst the first authors to emphasize the theoretical relevance of thought experiments. His evolutionary approach to the development of physics, in particular, led him to deny the existence of any kind of qualitative gap either between *everyday* thinking and *scientific* thinking or between *human* and *animal* thought. Correspondingly, as one of his commentators has recently pointed out, according to Mach "there's nothing essentially human about experiments"[8]. In particular, he envisioned though experiments as *necessary preconditions* of physical experiments[9], thereby granting them a foundational role deeply rooted in the very structure of physical enquiry.

Along the 20th century thought experiments have gradually passed from a condition of relative disrepute, under logical positivism, to a growing interest both in their inner functioning and in their epistemological implications. The general distrust showed by the logical empiricists toward thought experiments, in particular, can be traced back to the instrumentalist views of the French physicist, historian and philosopher of science Pierre Duhem (1861-1916) concerning the relation between *theory* and *experiment*. The use of the "fictitious experiment" (*expérience fictive*) in physics, according to Duhem, had to be firmly rejected "because of the false ideas it deposits in the minds of students"[10]. Appealing to fictional scenarios, lamented Duhem, does not enhance our understanding of nature, insofar as it simply amounts to offering "an experiment *to*

[6] Witt-Hansen (1976).
[7] Mach (1976 [1905]).
[8] Sorensen (1992: 190).
[9] See Mach (1976 [1905]: 136).
[10] See Duhem (1914: 201).

be done for an experiment *done*"[11]. The practice, moreover, would be circular, insofar as it purports to justify a physical principle by appealing to facts whose existence has not been observed but merely hypothesized, and this hypothesis, in his turn, "has no other foundation than the belief in the principle supported by the alleged experiment"[12].

One of the main epistemological tenets of the program of rational reconstruction of science championed by the logical positivists was the sharp distinction between the so called context of *discovery* and the context of *justification*[13]. The *psychological* processes relying on which scientific hypotheses or theories are arrived at, according to this distinction, ought to be kept separate from the *experimental* or *logical* procedures by means of which those hypotheses and theories are tested or justified. Correspondingly, thought experiments, from a logical positivist perspective, may well play the role of powerful heuristic tools, thereby contributing to the process of discovery, but they cannot contribute in any way to the process of *justification*, they do not have, that is, the power to warrant any conclusion arrived at by their means and do not have, therefore, any demonstrative value. A clear formulation of this stance, for instance, is still to be found during the mid-sixties in the highly influential epistemological views of the German philosopher of science Carl Gustav Hempel (1905-1997). According to Hempel,

> "Scientific objectivity is safeguarded by the principle that while hypotheses and theories may be freely invented and *proposed* in science, they can be *accepted* into the body of scientific knowledge only if they pass critical scrutiny, which includes in particular the checking of suitable test implications by careful observation and experiment"[14].

The fundamental creative aspect of scientific activity leading up to new discoveries, in other words, certainly "calls for imaginative, insightful guessing"[15], but when it comes to the justification of new hypotheses the only legitimate evidential source is provided by real experiments.

During the second part of the last century, a slow process of critical reconsideration of the logical empiricist epistemology and of the corresponding approach to scientific activity was followed by a new wave of interest in *thought experiments*. The focus of attention gradually shifted from the prescriptive and definitional concerns who characterized the endeavors of the previous tradition, to a growing interest in more pragmatic aspects of science. New energies were invested in

[11] See Duhem (1914: 202). My emphasis.
[12] *Ibid* (202).
[13] See Reichenbach (1938: 3-7).
[14] Hempel (1966: 16).
[15] *Ibid* (17).

exploring the actual practices and patterns of reasoning scientists made use of and this shed new light on the significant role played in their work by thought experiments.

1.2 *Paradigm instances*

Since any definition is largely a matter of stipulation, there's almost as many definitions of 'thought experiment' available on the market as many authors have written on this topic. As a consequence of this fact, some have regarded the matter as a merely terminological issue and have very reasonably suggested that an exact definition might be something we could easily dispense with, focusing our attention instead on a limited number of 'paradigm instances', which usually prove themselves sufficient to recognize a thought experiment when we see one[16]. I sympathize with this approach, which I hold to be the most promising, and I find therefore convenient to begin by providing, in what follows, a few examples of thought experiments displaying a certain number of features that have generally been regarded as typical, in order to set the stage for further inquiry.

1.2.1 *Galileo's falling bodies*

One of the most famous examples, regarded and treated by the literature as paradigmatic, comes, perhaps not surprisingly, from physics. In his 1638 *Discourses and Mathematical Demonstrations Relating to Two new Sciences*, Galileo set out to refute the Aristotelian theory of motion by showing that all bodies fall at the same speed. He did this by means of a thought experiment. It must be recalled that, according to Aristotelian physics, heavier bodies fall faster than lighter ones. Galileo's thought experiment was specifically designed to show that this very assumption leads to a paradox, and must therefore be rejected.

His argument requires us to imagine that a heavier body be attached to a lighter one and then to ask ourselves what would happen if the two bodies where to be released together. The point is to show that, within this setting, Aristotelian physics allows for two different and mutually exclusive answers. As a matter of fact, if Aristotle's assumption were correct, then the lighter body, 'naturally' falling slower, should be expected to slow down the heavier one, and consequently the compound as a whole. On the other hand though, that very same assumption seems to justify the exact opposite expectation, according to which the compound of the two

[16] Brown (1991:1). A similar strategy is adopted by Kuhn (1977: 241).

bodies linked together, being heavier, should fall faster than the previously heavier body taken alone. The two opposite answers lead to the absurd conclusion that the compound body must fall both faster and slower than the heavier of its components. According to some, as we shall see, Galileo's thought experiment had the further merit of offering a positive account of the behaviour of falling bodies, insofar as the only way to avoid the paradox seemed to him that of assuming that bodies, regardless of their weight, fall at the same speed.

1.2.2 *Newton's rotating bucket*

Sir Isaac Newton, the father of classical mechanics, is to be held responsible for a further thought experiment, which is contained in a Scholium to the definitions given in the first book of his *Philosophiae Naturalis Principia Mathematica*, and which has long been regarded by philosophers as paradigmatic. One of the general aims of the Scholium was that of supporting Newton's *substantivalist* view of space, according to which, roughly, space would be a real entity, distinct from body, and to reject a *relationist* view, widely spread in 17th century and held notably by Descartes, according to which space would be nothing over and above a relation amongst bodies. It must indeed be recalled that Newton's epoch-making treatise was based on a sharp distinction between what he held to be the *true*, *absolute* and *mathematical* notions of space, time and motion, on the one hand, and their *apparent*, *relative* and *common* counterparts, on the other. The thought experiment we are concerned with, in particular, was specifically designed in order to show that an adequate analysis of *true* motion must involve reference to *absolute* space, thereby indirectly proving its existence. Absolute space must exist, so reasoned Newton, because it is the only way to account for the undeniable existence of absolute motion.

The thought experiment invites us to imagine a bucket half-filled with water and suspended from a long cord which has been twisted up on itself several times. Given these initial conditions, it does not seem necessary to enter a laboratory and actually perform the procedure in order to acknowledge that, were the bucket to be suddenly released, we would observe both water and bucket going through the following two, ostensibly different, stages.

When the bucket is released, it begins spinning quickly on itself. At this time, water and bucket are in *relative motion* and the water surface displays a *flat* shape. As the bucket continues to rotate, the water begins to rotate with it and to climb up its sides, thereby entering the second (ideal) stage. At this time water and bucket are at *relative rest* and the water surface displays a *concave* shape. It will now be sufficient to reflect on the setting just sketched in order to conclude that the concave shape of the water surface, due to the receding of the water from the

axis of circular motion, cannot possibly be explained by appealing to the relative motion of the water with respect to the bucket. One can easily infer this by recalling that in the first stage, in which water and bucket are indeed in relative motion, the water surface appears flat. It follows, maintains Newton, that the only way to account for the effect is to postulate the existence of *absolute space*, with respect to which the corresponding *absolute* circular *motion* of both bucket and water is now occurring.

1.2.3 *Maxwell's demon*

A third example is due to the Scottish physicist James Clerk Maxwell and is aimed at enhancing our understanding of the second law of thermodynamics by showing that it has only a statistical certainty and that it can therefore, in principle, be violated. A consequence of this law is normally take to be that "it is impossible to design a machine whose sole effect is to transfer heat from a colder heat reservoir to a hotter reservoir"[17]. As a matter of fact, according to classical thermodynamics, two bodies of different temperature, when brought into contact, will reach a thermodynamic equilibrium in which they both have the same temperature. *Entropy*, in particular, is a measure of this process and it approaches a maximum value once the equilibrium is reached. Intuitively, heat can never flow spontaneously from a colder to a warmer object. This can also be stated by saying that in an isolated system, entropy never decreases. Now, the kinetic theory of heat endorsed by Maxwell models gases as if they were composed of molecules. The large number of molecules requires that the mathematical description of their behaviour be statistical. Accordingly, temperature, in this model, is the average kinetic energy of these molecules. A troublesome consequence of the statistical treatment of these systems is that it allows for the possibility of an entropy decrease. In other words, once we buy into the statistical framework, we are committed to the claim that it is at least possible, although very improbable, for heat to flow from a colder body to a warmer one, thereby violating a fundamental law of classical thermodynamics.

Maxwell's thought experiment is aimed precisely at making this consequence appear less obviously absurd. We are asked to imagine a couple of adjoining chambers containing gases at different temperatures[18]. The chambers are divided by a small door, controlled by a little intelligent being. It has to be kept in mind that the temperature of a gas, according to the kinetic theory of heat, is given by the average kinetic energy of its molecules. Accordingly, in the

[17] This formulation was slightly adapted from the one given in Norton (1991: 32).
[18] In what follows I draw on Brown (1991: 37).

present case, each of our two gases contains molecules which are faster or slower than the average molecule. This fact allows the little 'demon' envisioned by Maxwell to quickly open and close the small door in order to let fast molecules from the cold gas into the hot one, and slow molecules form the hot gas into the cold one. As a result of repeatedly performing this procedure, the average speed of the molecules contained in the cold chamber will decrease, therefore lowering its temperature, whereas the average speed of the molecules contained in the hot chamber will increase, therefore raising its temperature. This, in turn, amounts to saying that the cold chamber will become colder, and the hot one hotter. There will occur, in other words, a flow of thermal energy from a colder body to a hotter one, contrary to what stated by the second law of classical thermodynamics. It follows that this law enjoys a merely statistical certainty.

1.2.4 *Einstein's elevator*

A still further instance, perhaps one of the most famous thought experiments of the past century, comes from Albert Einstein. It is designed in order to justify the so called *principle of equivalence*. This principle establishes the physical equivalence between a uniformly accelerated frame of reference outside of a gravitational field and a frame of reference at rest within a homogeneous gravitational field. As a matter of fact, gravitational force as described by Newton appears to be incompatible with the new framework for the laws of mechanics put forward by Einstein's special theory of relativity. Accordingly, the first step towards extending this last theory to gravitational phenomena was for Einstein to show that special relativity still applies locally within gravitational fields. He did this by means of a long celebrated thought experiment[19].

The experiment features an observer inside of an elevator in a region of space remote from gravitational sources. Both the elevator and the observer are being pulled upwards by a mysterious force which, being constant, accelerates them uniformly. Now, if our observer were to release a body from her hand she would see it behave exactly in the same way as it would do if the elevator were at rest within a homogeneous gravitational field, i.e. she would see the body swiftly moving towards the floor of the elevator. The main point of the thought experiments is to show that, in the described situation, no real experiment could possibly enable the observer to decide whether she is in fact in a gravitational field or not. This amounts to say that the behaviours of two bodies, one with respect to a uniformly accelerating frame of reference in a gravitation free region, and the other with respect to a frame of reference at rest in a

[19] I draw here on Norton (1991: 136-138).

homogeneous gravitational field are *observationally* indistinguishable. Indeed, the same observer's inability to distinguish between a region of space subject to a gravitational field from one that is not, would be produced by an analogous situation in which the unlucky observer were to find herself in an elevator in free fall within a homogeneous gravitational field. It is now the case that, according to a general verificationist assumption endorsed by Einstein and concerning the proper relationship between theory and observation, a good physical theory should not distinguish between two separate state of affairs which are not also observationally distinct. It follows that the observer's inability to distinguish between the two scenarios described by the thought experiment provides strong intuitive evidence for the plausibility of the principle of equivalence.

1.3 *Instances of what?*

The examples given above should be at least sufficient to convey a rough idea of the kind of reasoning involved in a thought experiment. Nevertheless, it may still not be immediately clear precisely why or in what sense these imaginative stories, these "forays of the imagination", as someone has written[20], have come to be referred at by the expression 'thought experiments'. While indeed, on the one hand, the word 'experiment' is not usually associated with the mere armchair contemplation of the structure of an empirical question, on the other hand it is not exactly straightforward how 'thinking' could be envisioned as a strictly 'experimental' activity. As a matter of fact it is certainly possible to argue that, in the light of our every day use of the terms 'thought' and 'experiment', the locution 'thought experiment' itself seems fatally bound to sound oddly oxymoronic. For similar reasons many have indeed found the phrase more or less intentionally misleading, lamenting that "a thought experiment is no more an experiment than a plastic flamingo is a flamingo"[21], and have preferred to adopt the more neutral and general term 'argument'. In their opinion, as we shall see, any thought experiment, despite of its rhetorical drapery, can ultimately be reduced, in principle, to a corresponding and explicit argument. Others, by contrast, have felt that a similar reduction, regardless of its feasibility, would be seriously wrongheaded insofar as it would utterly fail to capture the real nature of these procedures. The only way to make sense of the peculiar form of persuasion produced by a thought experiment, they have consequently claimed, is to focus on its inherently psychological elements.

[20] Wilkes (1988: 2).
[21] Rescher (2005: 8). On 'veridical' and 'falsidical' adjectives as applied to thought experiments see Sorensen (1992: 217).

Performing a thought experiment, intuitively, does not require the use of laboratories or of more or less sophisticated instrumentation, insofar as its outcomes are not grounded on any kind of measurement but are *qualitative* in nature. Moreover, thought experimenting does not involve any physical manipulation of variables, unless of course we metaphorically stretch the meaning of the term 'manipulation' and take it to refer to a purely mental performance[22]. Scientists usually resort to thought experiments when ordinary experiments either *cannot* or do not *need* to be performed. This last eventuality, in particular, appears to be more directly relevant to the present discussion. It is, in fact, for this reason that many have decided to characterize these procedures as *thought*-experimental insofar as, contrary to actual experiments (performed or only envisioned), they are carried out, as a popular metaphor suggests, entirely in the "laboratory of the mind"[23]. Their conclusions, so it is claimed, usually strike us as compelling, and their evidential value, at first glance, does not seem to depend on their material execution at all. This is due to the fact that thought experiments seem to instantiate cases in which the mere reflecting upon a hypothetical scenario has the power to make us grasp some intuitive truths concerning the investigated phenomena. New and unsuspected features or properties of an entity or of a process seem to become immediately perspicuous to us after the 'story' has been told, and this is precisely what seems to render superfluous the material execution of the experiment.

The *experimental* character of these procedures, on the other hand, is usually taken to lie in the fact that, similarly to ordinary experiments, their outcomes do not seem to bear exclusively on our conceptual framework. Both the intentions of the experimenter and the outcome of a thought experiment, it has been argued, explicitly purport to enhance our knowledge and understanding of the actual world, not our linguistic competence or our logical skills. It has been further observed that "all experiments work by raising the experimenter's status as an epistemic authority"[24]. Similarly, as a consequence of the exposition to these fanciful narratives, the subject seems to be epistemically altered by the experience and to *see* the world differently.

1.4 *The fundamental question*

Despite our previous considerations, the reason why philosophers should be interested in this practice is still in need of careful considerations of a more abstract sort. We might begin by noticing that, while going through our examples, we might have had the impression, long shared

[22] By this I do not mean to suggest that this is not a viable option, it indeed characterizes the cognitive approach to thought experiments put forward by Gooding (1992, 1992b) and Nersessian (1992).
[23] The metaphor is due to James Robert Brown. See Brown (1991).
[24] Sorensen (1992: 5).

by many, that the very argumentative structure of a thought experiment has something rather puzzling in it. As a matter of fact, when first confronted with one of these more or less fanciful narratives, we usually find ourselves trying to assess what is 'really' going on in it. The paradoxical nature of these devices becomes suddenly apparent to us once someone brings to our attention an important general feature shared by every thought experiment: A thought experiment, as we already mentioned, purports to tell us something new and substantial about the actual world without at the same time introducing the slightest bit of fresh empirical information beyond the one that was already available at the beginning of the argument. By performing a thought experiment, it seems, a scientist sets out to learn something new about the world he lives in, and not merely about the time-bound and culture-bound conceptual apparatus relying on which he has been trained to describe that world. And yet, again, his reasoning does not appeal to any new empirical data beyond the ones he already had at the beginning of his inquiry. By the end of the 'story', then, we have undeniably gained new information, this information seems to be empirical in nature, and yet we are embarrassingly clueless as to where it might come from!

It has been convincingly argued, I believe, that the puzzlement effect just described could be easily placed within the traditional debate between rationalists and empiricists[25]. As it is well known, one of the fundamental questions around which this epistemological debate has traditionally revolved concerns the extent to which our knowledge of the external world depends on our sense experience. According to the standard rationalist construal, insofar as no inference drawing process could be entirely justified on purely empirical grounds, at least part of our knowledge of the external world, and of the justification thereof must be *a priori*. Thought experiments, when taken at face value, seem to offer strong support for this idea. In other words, they seem to instantiate the perfect cases in which some contents of our knowledge intuitively seem to outstrip the information provided by the senses, which means, according to the rationalist, that this additional information cannot but be provided by 'pure reason' itself, however we might wish to define it. Galileo, for instance, seems to be *a priori* entitled to believe truly, and therefore to have *a priori* knowledge of the fact, that 'all bodies fall at the same speed'. On the other hand, one of the central tenets of the opposite, empiricist tradition, requires that knowledge can only be gained, if at all, by sense experience, and since thought experiments provide no exception to this rule, a satisfactory explanation of their achievements cannot appeal to any mysterious epistemic process but must occur along empiristic lines. Where this is not possible, the existence of that very knowledge which the single thought experiment purports to provide is called into question.

[25] *Ibid* (15).

Once placed within its proper epistemological background then, the question that has set in motion the whole philosophical debate about thought experiments seems to take the following form: How can a successful thought experiment manage to be informative about the world we live in without adding new empirical data by means of actual experimenting?[26] In Galileo's *Dialogue Concerning the Two Chief World Systems*, for instance, we find an astonished and sceptical Simplicio, the spokesman of Aristotelian physics, asking Salviati, Galileo's stand-in: 'So you have not made a hundred tests, or even one? And yet you so freely declare it to be certain?' To which Salviati confidently replies: 'Without experiment, I am sure that the effect will happen as I tell you, because it *must* happen that way'[27].

1.5 *Different epistemologies*

The question raised above could be rephrased in a perhaps more prosaic way as follows: How exactly do thought experiments achieve what they do (when they do)? Over the years, several attempts to address this question have been made. Those attempts have generated what we may consider as various different 'epistemologies' of thought experiments. In what follows, I'll try to sketch briefly, and without any pretension of completeness, some of the main stances that have been taken in this regard. My survey, in particular, will focus on five different positions…

1.5.1 *Brown's Platonism*

Following a tradition that can be traced back to the philosophical views of the French historian of science Alexandre Koyré (1892-1964), according to whom "good physics is made *a priori*"[28], James Robert Brown has tried to offer an account of thought experiments along *a priori* and Platonistic lines, the plausibility of which has been long debated[29]. As we shall see, some thought experiments, in his view, make a rationalist interpretation of science believable by showing that there is a part of our knowledge of the external world that cannot be accounted for along empiricist lines.

In order to set the stage for his main thesis, Brown has developed a fairly elaborated taxonomy of thought experiments, the details of which do not need concern us here. It is sufficient for our

[26] The credit for having singled out this fundamental epistemological question goes to Kuhn (1964: 241).
[27] Galilei (1967: 145), quoted in Brown (1991: 3). My emphasis.
[28] Koyré ([1960] 1968: 68), quoted in Brown (2004b:1130).
[29] See Brown (1986, 1991a, 1991b, 1993, 2004a, 2004b).

purposes to start by noticing that thought experiments, in his view, can be divided according to the role they play within our knowledge. In general, they can be regarded as falling under two broad general kinds, which Brown labels respectively *destructive* and *constructive*. While a *destructive* thought experiment consists of a "picturesque *reductio ad absurdum*"[30], which purports to refute a rival theory by showing that a particular thesis endorsed by that theory leads to an absurd conclusion, a *constructive* thought experiment positively aims at establishing a new phenomenon and suggesting the best explanation for it.

A small third class of thought experiments, which Brown dubs *platonic*, have the merit of playing both a negative and a positive role at the same time. Their *pars destruens* is specifically designed to refute an established theory along the lines of a destructive thought experiment, whereas their *pars construens* plays the creative role of bringing a new theory into being. Brown provides the following characterization of this third class of thought experiments:

> "A *platonic thought experiment* is a single thought experiment which destroys an old or existing theory and *simultaneously* generates a new one; it is *a priori* in that it is not based on new empirical evidence nor is it merely logically derived from old data; and it is an advance in that the resulting theory is better than the predecessor theory"[31]

A platonic thought experiment then, according to the above characterization, displays two fundamental features, namely (1) it does not introduce new empirical data, and (2) it does not establish new logical truths. In Brown's opinion, these two features constitute enough evidence to support his central claim: platonic thought experiments are epistemically very remarkable insofar as they "transcend empirical sensory experience"[32], thus providing us with *a priori* knowledge of nature.

Brown regards the famous thought experiment on falling bodies performed by Galileo and briefly sketched above as providing the standard example of a platonic thought experiment. In this case it is indeed certainly possible to show, by means of a deductive argument, that a logical contradiction can be derived from one of the fundamental tenets of the Aristotelian theory of motion. As we have seen, the claim that heavier bodies fall faster than lighter ones allows for two different and mutually exclusive consequences and must therefore be rejected as inconsistent. On the other hand though, according to Brown, no straightforward inferential process seems suitable, in principle, to lead the transition from the old theory to the new one, for

[30] Brown (1991b: 34).
[31] *Ibid* (1991b: 77). The second emphasis is mine. As a matter of fact, as will become apparent in what follows, the adverb 'simultaneously' is crucial to Brown's account, since it indicates that the passage from the old theory to the new one does not involve any further inferential step, neither deductive nor inductive.
[32] Brown (2004b: 1130).

the simple reason that the principle according to which 'all bodies, regardless of their weight, fall at the same speed' is not a logical truth. In order to acknowledge this, it is enough to realize that the speed of a falling body, from a logical point of view, might depend just as well on its color, or on its chemical composition. It follows, according to Brown, that in order to explain this transition we need to appeal to some other kind of cognitive process.

It is now the case that Brown's Platonistic view of mathematics has led him to believe that, besides ordinary physical perception of material objects, human beings enjoy a kind of non sensory perception, which grants them epistemic access to some abstract entities. It also happens that a realist account of the laws of nature, recently put forward by David Armstrong, Fred Dretske, and Michael Tooley[33], and readily endorsed by Brown, construes these laws precisely as abstract entities. The laws of nature, according to this interpretation, "are relations among universals, that is, among abstract entities which exist independently of physical objects, independently of us, and outside of space and time"[34]. This last metaphysical claim, then, provides the final ingredient of Brown's Platonistic epistemology. "Just as the mathematical mind can grasp (some) abstract sets", writes Brown, "so the scientific mind can grasp (some of) the abstract entities which are the laws of nature"[35]. This same sort of non sensory perception, in his opinion, is exactly what enabled Galileo to 'see' the relevant law he was trying to read off the book of nature, and to perform the leap which led him from the old Aristotelian theory to his own new one.

One last important thing to be noticed is that, according to Brown's construal of the notion of *a priori*, our perception of abstract entities is in itself fallible. As a consequence, Brown feels compelled to observe that, when referring to the epistemic achievements of a thought experiments, "the term 'knowledge' may be too strong as it implies *truth*; 'rational belief' might be better since, on my view, what is *a priori* could be false"[36]. If this wasn't indeed the case, wrong thought experiments, i.e. thought experiments which, as it often happens, lead us to a false belief, would be left totally unexplained. As a matter of fact, according to Brown, just as the physical world may at times contribute to the production of rational, but false beliefs in the existence of various theoretical entities, such as, for instance, phlogiston, caloric, or aether, the abstract world may, in the same mysterious way, cause the belief in the wrong conclusion of a thought experiment[37].

To conclude, Brown's epistemology of thought experiments seems to have two fundamental requirements, namely (1) the existence of a non-sensory perception of abstract entities, which

[33] See Armstrong (1983), Dretske (1977), and Tooley (1977).
[34] Brown (1991b: 82).
[35] *Ibid* (ix).
[36] *Ibid* (79, note 4).
[37] *Ibid* (92-93).

allows us to 'see' the relevant laws of nature, and (2) the plausibility of a realist account of the laws of nature themselves.

1.5.2 *Norton's empiricism*

By Brown's own admission "if and antidote to [his] gung-ho Platonism should be needed, then it can be found in either Norton's empiricism or Sorensen's naturalism"[38]. While Sorensen's views will be taken care of in one of the next sections, I will now try to provide a brief outline of a stance first put forward by John Norton[39] and which can be regarded as an empiricist reaction to Brown's own proposals.

The enthusiasm emphasized in the above quotation, as we have seen, has led Brown to see thought experiments as epistemically very remarkable. It is important to observe that, according to his general line of reasoning, the existence of (a particular kind of) *a priori* knowledge of nature has to be hypothesized in order to explain a supposedly 'peculiar [epistemic] phenomenon', i. e. the thought-experimental discovering of new facts[40]. It follows that, in case we were able to come up with a better explanation for that very same phenomenon in some other way, we could light-heartedly get rid of *a priori* knowledge of any kind.

Now, Norton believes that, since "all knowledge of our world derives from experience"[41], this last challenge can and must be met. His central claim is indeed that thought experiments, though certainly constituting a very useful and at times practically indispensable heuristic tool, are nonetheless *epistemically* quite unremarkable, insofar as they rely on our standard epistemic resources, namely "ordinary experiences and the inferences we draw from them"[42]. In particular, according to his analysis, "Thought experiments are usually introduced when the straight argument would be difficult to develop"[43]. In other words, they facilitate the accomplishment of a task which, *in principle*, could be completed even without their help, by means of an argument. Norton's line of reasoning can be roughly summarized by the following conditional:

> *if* it is possible to reconstruct every thought experiment as an argument, *then* (this means that) thought experiments are epistemically unremarkable.

---

[38] *Ibid* (x).
[39] See Norton (1991, 1993, 1996, 2004a, 2004b).
[40] Brown (1991: 98).
[41] Norton (1996: 335).
[42] *Ibid* (334).
[43] Norton (1991: 131).

Norton calls the antecedent of this conditional, which he believes to be true, *elimination thesis*[44]. According to this thesis, in principle, "any thought experiment can be *replaced* by an argument without the character of a thought experiment"[45]. This would follow from the fact that since, as already established, thought experiments do not introduce any new empirical data, all they can do is *reorganize* or *generalize* what we already know. While in the former case they would function as a *deductive* argument, in the latter they would function as an *inductive* one[46]. Despite all appearances then, according to Norton's deflationary stance, thought experiments may be rightfully seen as merely "picturesque argumentation"[47]. Accordingly, his view is neatly summarized by the following *reconstruction thesis*:

> "All thought experiments can be reconstructed as arguments based on tacit or explicit assumptions. Belief in the outcome-conclusion of the thought experiment is justified only insofar as the reconstructed argument can justify the conclusion"[48].

Since the conclusion of any thought experiment is indeed reached by ordinary inference, the fundamental epistemological point, according to Norton, is that "the degree of belief conferred by the thought experiment on it's outcome *coincides* with the degree to which the reconstructed argument supports its conclusion"[49]. As in any argument, maintains Norton, that degree of belief depends on our degree of belief in the argument's premises and in the deductive validity (or inductive strength) of the argument itself. As a consequence, to put it in his own words, "a good thought experiment is a good argument, a bad thought experiment is a bad argument"[50].

Norton's main target being Brown's platonic account, it is helpful to recall that Brown's view is based on an analogy between ordinary physical perception and some sort of non sensory perception[51] in virtue of which we are able to grasp some abstract entities, amongst which are the laws of nature. Now, in order to fully appreciate the spirit of Norton's proposal, it is important to stress that he does not rule out in principle the possibility that thought experiments provide us epistemic access to such abstract entities; rather, his claim is once again a conditional one,

---

[44] Norton (1996: 336).
[45] *Ibid* (336). My emphasis.
[46] *Ibid* (335).
[47] Norton (2004a: 1142). The thesis according to which thought experiments are arguments is also endorsed by Rescher (1991), Irvine (1991) and Forge (1991).
[48] Norton (1996: 339).
[49] *Ibid* (340), my emphasis.
[50] *Ibid* (335).
[51] This platonic perception, according to Norton, would be very different from ordinary perception "insofar as it obeys no known regularities that would allow us to control misperception". *Ibid* (335).

namely: *if* thought experiments provide us epistemic access to abstract entities, *then* "they may do so only insofar as these universals can be accessed via argumentation"[52].

Norton's analysis of Galileo's thought experiment on falling bodies provides an example of the way in which the reconstruction of any thought experiment, according to his view, should be carried out. It is quite uncontroversial that the first part of the thought experiment, in which Galileo rejects the Aristotelian theory of motion, is a straightforward *reductio* argument. Accordingly, Norton reconstructs its characteristic inferential structure in eight steps, leading from the assumption required for the *reductio* proof (1), namely that 'the speed of fall of bodies is proportionate to their weights', to the denial of that same assumption (8) [53]. It is not important here to follow Norton's reconstruction step by step. It is enough for our purposes to observe that, according to Norton, step (8) is where Brown's 'platonic leap' into the new theory occurs, the moment in which the platonic law according to which 'all bodies fall at the same speed' is directly 'perceived'. Norton, by contrast, holds that our degree of belief in that law, and hence in Galileo's new theory, depends rather on our belief in the tacit assumption (8a) according to which 'the speed of fall of bodies depends *only* on their weights'[54]. It is in virtue of this assumption, according to his analysis, that the final outcome can be reached via a simple inferential process, reconstructed by Norton as follows[55]:

8a. Assumption: The speed of fall of bodies depends *only* on their weights.

8b. Assumption: The speed of fall of bodies is some *arbitrary* monotonic increasing function of their weights.

8c. From 3, 5[56]. If the function is anywhere strictly increasing, then we can find a composite body whose speed of fall is intermediate between the speed of fall of its lighter components.

8d. The consequent of 8c contradicts 8b.

9. From 8d. The function is constant. All stones fall alike.

[52] Norton (1991: 129, note 1). This also explains why, according to Norton, empiricism is not strictly speaking fundamental to his view. See, for instance, Norton (1996: 336): "the view that thought experiments are arguments is not equivalent to empiricism. It is entailed by empiricism, but the converse implication does not obtain […] In principle, one may hold the argument view without any commitments concerning the origin of the premises used in the arguments and their connection with experience".

[53] Norton (1996: 341-342).

[54] *Ibid* (342). My emphasis.

[55] *Ibid* (343).

[56] The assumptions 3 and 5 of the *reductio* argument had established, respectively, that "if a slower falling stone is connected to a faster falling stone, the slower will retard the faster and the faster speed the slower", and that "the composite of the two weights has greater weight than the larger".

In particular, according to Norton, assumption 8b constitutes "the most general viable theory"[57] of the behaviour of falling bodies, a special case of which would be Galileo's own theory. As a matter of fact, Norton observes, any function other than Galileo's would be rejected by a similar *reductio* argument. The above considerations, according to Norton, are enough to show that the *reductio* argument 1-8 can be generalized to yield Galileo's theory without being compelled to rely on any mysterious 'platonic leap', or to appeal to any kind of *a priori* knowledge.

1.5.3 *Kuhn's constructivism*

A third, highly influential view is due to the historian and philosopher of science Thomas Kuhn (1922-1996). In a famous paper first published in 1964 Kuhn addresses the three following questions:

1. What do we learn from a thought experiment?
2. How does a thought experiment increase our knowledge of nature?
3. Which conditions must a thought experiment satisfy in order for us to learn something from it?[58]

In what follows I will try to provide a brief outline of his answers to these questions.
At the very heart of Kuhn's ideas concerning thought experiments lies the general assumption according to which *nature* and the *conceptual apparatus* relying on which we try to understand it are jointly implicated. Based on this assumption, Kuhn rejects as misdirected the received view according to which the kind of understanding produced by thought experiments would not really be an understanding of nature, but rather of the scientist's conceptual apparatus. The only function of a thought experiment, according to this view, would be that of correcting previous conceptual mistakes by enabling the scientist to recognize contradictions inherent in his way of thinking. This would also explain why, given the merely logical nature of its task, a thought experiment does not need to appeal to any new empirical data[59]. This account of thought experiments, according to Kuhn, is not tenable, in so far as it is arguably at odds with the actual development of physical science, which could hardly be envisioned as a process whose only

[57] Norton (1996: 342).
[58] Kuhn (1964: 241).
[59] *Ibid* (242).

function is that of gradually dispelling logical confusions[60]. Although Kuhn doesn't explicitly ascribe this view to any specific author, it is important for our purposes to recognize that the account just sketched reflects the stance taken towards philosophy by the Logical Positivists. Indeed, the linguistic formulation of empiricism endorsed by these philosophers, as it is well known, firmly denied the capacity of philosophical activity to exceed the limits of pure conceptual analysis[61].

Thought experimental scenarios, according to Kuhn, have the power to generate an experience of *paradox* by confronting us with a situation in which two previously well established criteria for the use of a certain concept happen to conflict.

In order to exemplify this situation Kuhn compares a real experiment performed by the Swiss developmental psychologist Jean Piaget (1896-1980) with a thought experiment due to Galileo[62]. Both experiments, the details of which are not relevant here, focus on the notion of speed and share the common feature of producing two very similar cases of conceptual revision. Indeed, when confronted with specific questions concerning the motions of the objects observed (or envisioned), both Piaget's children and the Aristotelian physicists are often forced to attribute the relational properties 'faster' and 'slower' at the same time to the same object, thereby revealing the presence, in their reasoning, of two different but mutually inconsistent criteria for applying the concept of *speed*. This is due, in particular, to their failure to distinguish between *instantaneous* and *average* speed. Eventually, the acknowledging of this situation will lead some of the subjects exposed to the experiment to a careful reassessment of the relevant notion.

While on a standard logical empiricist account this outcome would be taken to constitute evidence for the fact that the notion possessed by the subjects prior to the experiment was self-contradictory, the matter at issue here, according to Kuhn, is not the consistency of the concept itself, but rather the consistency of its *use*. Prior to the puzzling effect induced by the experimental setting, maintains Kuhn, the Aristotelian notion of speed was not self-contradictory at all[63], at least not in the same way as the notion of a square-circle would be. While in fact the latter notion could not be instantiated in any possible world, the former could. We could indeed certainly conceive of a world in which all motions occurred at uniform speed and in such world, according to Kuhn, the Aristotelian notion of speed would be perfectly consistent, for the trivial

---

[60] *Ibid* (253). This account, claims Kuhn, trivializes the function of thought experiments in so far as it "is too reminiscent of the familiar position which regards the Ptolemaic theory, the phlogiston theory, or the caloric theory as mere errors, confusions, or dogmatisms which a more liberal or intelligent science would have avoided from the start".

[61] See Rosenberg (2005: 23-24).

[62] The theoretical usefulness of drawing a parallel between a *thought* experiment and a *real* experiment, according to Kuhn, is based on the fact that "the effects of thought experimentation […] are much closer to those of actual experimentation than has usually been supposed". *Ibid* (242).

[63] "It is significantly misleading to describe as "self-contradictory" or confused the situation of the scientist prior to the performance of the relevant thought experiment". *Ibid* (242).

fact that in such world *instantaneous* and *average* speeds would always coincide. The shortcomings of the Aristotelian concept then, in Kuhn's own words, "lay not in its logical consistency but in its failure to fit the full fine structure of the world to which it was expected to apply"[64].

This last quotation introduces us to the next fundamental feature of Kuhn's view. For the same reason, adds Kuhn, we should rather say of the Aristotelian physicist living in the boring possible world just sketched that "consciously or unconsciously, he had embodied in his concept of speed his expectation that only uniform motion occurs in his world. We would, that is, conclude that his concept functioned in part as a law of nature"[65]. As a matter of fact, according to Kuhn, in so far as concepts are integral parts of the theories to which they belong, they do not work just as abstract definitions do in hypothetical-deductive systems, but also and fundamentally as natural laws[66]. Concepts, according to his view, display an essential legislative content or function in so far as they reflect specific *ontological expectations* of their users. It follows that the use a scientist makes of a concept ought to be regarded as an "index of his commitment to a larger body of law and theory"[67]. This fact, in turn, hints toward an answer to the first question raised by Kuhn, namely the one concerning the object of the cognitive achievement made possible by a thought experiment. As a matter of fact, by acknowledging the mismatch between his concept and the structure of the world to which it applies, the scientist, according to Kuhn, does in fact learn about the concept *and* about the world at the same time[68].

This construal of the potential achievements of a thought experiment offers further support to the thesis previously endorsed by Kuhn, according to which the historical role played by thought experiments would bear a strong resemblance to the one played by real experiments. In both cases, maintains Kuhn, a failure of nature to conform to a previously held set of expectations is disclosed to the scientist, and in both cases the specific ways in which this failure occurs can provide him with instructions about how to revise his concepts.

Thought experiments, in particular, as opposed to real experiments, play this role by giving the scientist access to information which is already at hand and yet simultaneously "somehow inaccessible to him"[69]. The notion of 'unassimilated observations' (or 'unassimilated anomalies' or 'incongruous experience') plays indeed a central role in Kuhn's picture and provides an answer to the second question raised above, concerning the functioning of a thought experiment.

---

[64] *Ibid* (258).
[65] *Ibid* (255). In the same situation, Kuhn asks rhetorically: "Ought we demand of our concepts, as we do not and could not of our laws and theories, that they be applicable to any and every situation that might conceivably arise in any possible world?".
[66] *Ibid* (257-258).
[67] *Ibid* (260).
[68] *Ibid* (258).
[69] *Ibid* (261).

It is precisely in this form, indeed, that nature fails to fit the scientist's mental equipment and is therefore to be held responsible for the conceptual confusion induced by the thought experimental scenario[70]. Thought experiments, in other words, put the scientist in front of a situation which is potentially capable of contributing to the production of what Kuhn elsewhere calls a *paradigm shift* or a *scientific revolution*[71].

The seemingly steady advance of scientific research, according to Kuhn, is due to the fact that scientists tend to restrict their attention to problems defined by the conceptual apparatus they are trained to rely on[72]. This mode of problem selection, in turn, is bound to push observations which do not fit their theory-induced expectations to the fringe of their scientific attention. The paradigm shift mentioned above is then induced by the fact that, while some of those anomalies can be readily taken care of by small, local adjustments of the conceptual apparatus available to the scientist, some others cannot, and spark off a process of deep conceptual revision, eventually leading to the adoption of an entirely new conceptual framework, to which the previous anomalous observations are gradually assimilated. This process, according to Kuhn, far from being restricted to the actual experimental practice, is at the heart of thought-experimental situations themselves[73]. It is in this sense that thought experiments, in his view, are "essential analytic tools" which can "enable the scientist to use as an integral part of his knowledge what that knowledge had previously made inaccessible to him"[74].

What we have said so far should at this point allow us to anticipate Kuhn's answer to his third and last question, concerning the conditions that a thought experiment must satisfy in order to be effective, i.e. to teach us something. As a matter of fact, precisely because it draws on a sort of receptacle of previously 'unassimilated observations', a thought experiment, in order to allow us to learn something from it, must put us in condition to employ our concepts in the same ways they have been employed before. This requires in particular that "nothing about the imagined situation may be entirely unfamiliar or strange"[75]. It finally follows from the above that, in order to learn from thought experiments, the imagined situation need not only be one that nature itself could present, but it must also be one that 'however unclearly seen', has confronted the scientist before[76].

[70] *Ibid* (261).
[71] See Kuhn (1962).
[72] Kuhn (1964: 261).
[73] *Ibid* (263).
[74] *Ibid* (263).
[75] *Ibid* (252).
[76] *Ibid* (265).

1.5.4 *Sorensen's naturalism*

One of the most extensive and systematic attempts at offering a fully articulated general theory of thought experiments is due to Roy Sorensen[77]. The intent of providing a comprehensive and detailed survey of his views lays therefore far beyond the reach of this section. I will rather focus, in what follows, on two main aspects of Sorensen's account, which I find to be particularly relevant to the present discussion, namely (1) his evolutionary epistemology of thought experiments, and (2) his analysis of their logical structure.

Ernst Mach, according to Sorensen, has to be credited for having fully grasped the far-reaching epistemological implications of Darwin's biology, as well as for having successfully applied them to the study of thought experiments. In so doing, Sorensen holds, Mach indicated a viable and liberating way to solve the age-old tension between rationalist and empiricist accounts of knowledge. Mach's leading assumption was that natural selection favours in minds an ability to mimic natural patterns. This would allow humans, as well as other animals, to store large quantities of tacit information which their seemingly *a priori* intuitions would unconsciously draw upon when formulating synthetic judgements about the external world.

Along the lines of the same evolutionistic account, Sorensen champions a form of what he calls *metaphilosophical gradualism*, according to which the difference between scientific and philosophical endeavours should be regarded as a difference in *degree* and not in *kind*. Accordingly, he claims that both, scientific and philosophical thought experiments, regardless of their level of sophistication, evolved from a vast array of *practical* experiments which populated the every day life of our remote hunter-gatherer ancestors. Thought experiments, in particular, would not constitute new kinds of entities, but ought to be understood and treated as limiting cases of real experiments, gradually evolved by an attenuation of the execution element and a corresponding elaboration of the design element proper of every experiment[78]. While indeed, according to Sorensen, real experiments are aimed at raising or answering questions concerning specific relationships between independent and dependent variables by means of a process of actual manipulation, though experiments, on the other hand, purport to reach the same results by mere rational reflection on their experimental design.

In addition to this, Sorensen explicitly parallels his evolutionistic thesis with a further developmental thesis, according to which children's ability to perform thought experiments, as opposed to their ability to perform real ones, would appear later in life. Accordingly, he claims that "by studying the order in which children learn the skills, we gain evidence about the order in

[77] Sorensen (1992).
[78] *Ibid* (212).

which their ancestors acquired the skills leading up to thought experiment"[79], thereby implicitly endorsing the well known biological principle according to which *ontogenesis* would recapitulate *phylogenesis*.

Elaborating on Mach's views, Sorensen's account also aims at including a further feature of thought experiments within the same evolutionary framework. As a matter of fact, the seemingly *a priori* intuitions elicited by thought experimental scenarios appear to be modal in nature. They purport to give us access to modal truths about the world, insofar as they provide us with clear, ready to use instructions as to what *could* or *could not* happen given certain known circumstances. Sorensen is confident that 'the right kind' of evolutionary theory would prove itself capable of providing a naturalistic account of the modal knowledge at work in thought experiments. An adequate evolutionary epistemology, in particular, should rely solely on the evolutionary forces accepted by contemporary biology. Within this updated theory, according to Sorensen's brand of reliabilism, the belief-forming mechanism responsible for the existence of our innate pre-theoretical knowledge of *possibilities* is to be found in the generate-and-eliminate process of natural selection[80]. This knowledge would be 'innate' insofar as, strictly speaking, it is not *learned*, but inherited from the experiences of our ancestors, and would thus constitute what Sorensen dubs "a poor man synthetic a priori"[81].

Evidence for this construal, according to Sorensen, would be provided by the fact that *hypothetical* reasoning, which makes extensive use of modal notions, reveals itself particularly useful for practical purposes[82]. The existence of tight links between *practical* and *theoretical* skills suggested by the cognitive gradualism mentioned above is then appealed to by Sorensen in order to account for the transition from these practical purposes to more theoretical ones. Adopting this kind of evolutionary framework, according to Sorensen, would allow us to envision the highly developed theoretical skills possessed by humans as welcomed 'side-effects' of those same biological mechanisms responsible for the selection of their practical counterparts.

It is also important to keep in mind that, according to Sorensen, claiming that selection for a trait has occurred does not immediately guarantee the epistemic reliability of that trait. Natural selection, as Sorensen points out, does not logically entail reliable belief formation[83]. As a matter of fact, insofar as nature only selects for traits that prove sufficient to enhance reproductive success, "there is little hope of a perfect fit between an organism's representation of the world

[79] *Ibid* (213).
[80] Sorensen (1992b: 24, 33).
[81] *Ibid* (24). The rationale for the deflationary attitude conveyed by this characterization is provided by Sorensen himself by pointing out that: "The relevant 'a priori' is a relative one: it refers to what can be known prior to the trial, not prior to *all* experience whatsoever". It is also fair to add that Sorensen leaves room for the doubt that "under stiff standards of belief attribution, none of our innate knowledge may turn out propositional". *Ibid* (25).
[82] *Ibid* (28-31).
[83] *Ibid* (35).

and the real world"[84]. Our instinctive or innate modal knowledge, in other words, might be reasonably expected to be fallible. Nonetheless, it has proved effective along the evolution of our species, or else it would not have been selected for at all. The claim that evolution leads (or displays a tendency to lead) to reliable belief formation, according to Sorensen, is not immune to counterexamples, but enjoys the status of a very reasonable inductive generalization. This means, in particular, that although the outcomes of thought experiments will be subject to 'wide fluctuations in reliability'[85], this reliability is nonetheless underwritten by the principles of natural selection.

As we mentioned above, Sorensen's general theory of thought experiments includes a detailed analysis of their logical structure. This analysis serves the systematic purpose of providing a general framework under which particular instances can be subsumed. Thought experiments, according to this account, conform themselves to the standards of what Sorensen calls a *cleansing model* of armchair inquiry, insofar as their main function is to make us more rational by revealing and eliminating inconsistencies contained in our beliefs. As a consequence of this general view, he holds that every thought experiment is ultimately reducible to, and in a sense generates, a *paradox*, i.e. to "a small set of individually plausible yet jointly inconsistent propositions"[86].

Focusing on their negative function, Sorensen characterizes thought experiments as *alethic refuters*, arguments, that is, specifically designed in order to refute statements "by disproving one of their modal consequences"[87]. Accordingly, Sorensen suggests to envision thought experiments as "expeditions to possible worlds"[88]. If a statement, for instance, implies that *p* fails to hold in any possible world, then this statement, intuitively, can be refuted by finding a possible world at which *p* is true. This is also the sense in which Sorensen, according to the modal notion implied by the original target-statement, divides the class of all thought experiments into the two sub-classes of '*necessity refuters*' and '*possibility refuters*'.

Once reduced to the form of a paradox, every thought experiment, according to Sorensen, contains exactly five members, or propositions. In the case of *necessity refuters*, which he takes to be the most frequent, this reduction takes the form of the following schema:

---

[84] *Ibid* (25).

[85] Mach's 'instinctive knowledge', after all, was held by Mach himself to be reliable only with respect to highly familiar phenomena, whereas many thought experiments present us with scenarios which are very far from our every day experience.

[86] Sorensen (1992: 5).

[87] By using the expression 'alethic modalities', as opposed for instance to 'deontic modalities' or 'epistemic modalities', Sorensen intends to refer to the notions of *necessity* and *possibility*. *Ibid* (135).

[88] *Ibid* (135).

1. S
2. $S \supset \Box I$
3. $(I \wedge C) \Box\!\rightarrow W$
4. $\neg \Diamond W$
5. $\Diamond C$

The above schema is interpreted as follows. Step (1) symbolizes the proposition Sorensen calls 'modal source statement'[89]. This expression presumably refers to the fact that the statement is one from which *modal* consequences are likely to be drawn. Typical source statements, according to Sorensen, include items such as, for instance, semantic theses or law statements. The proposition formalized in step (2), in turn, draws from the source statement the modal implication that the thought experiment is aimed at rejecting ("I" stands therefore for "implication"). Step (3) is slightly more complex. It formalizes a subjunctive conditional, namely a counterfactual[90], claiming that the modal implication (I) drawn from the source statement in conjunction with the thought-experimental scenario, (C, where 'C' stands precisely for "counterfactual scenario") has an odd consequence ("W" stands for "weird"). Step (4) explains the odd consequent of the above counterfactual as an impossibility. Step (5), finally, claims that the situation envisioned in the thought experimental scenario (C) is indeed possible.
The following, in turn, is the schema of what Sorensen calls a *possibility refuter*:

1. S
2. $S \supset \Diamond I$
3. $(I \wedge C) \Box\!\rightarrow W$
4. $\neg \Diamond W$
5. $\Diamond I \supset \Diamond (I \wedge C)$

While the first four steps of this schema are similar to the corresponding steps of the previous one, its last step claims that the possibility of the modal implication (I) implies the possibility of its conjunction with the counterfactual scenario appealed to by the thought experiment.
Sorensen is willing to grant that, in order to make every thought experiment fit the above set theoretic characterization, one must regiment its exposition in an often very artificial way. Nonetheless, one of the advantages of the present schemas, according to Sorensen, is that they

[89] *Ibid* (135).
[90] The epistemological problems raised by this particular kind of conditional sentence will be treated in section 3.3 below.

double as a *theory of fallacy*[91]. This is due to the fact that, by making explicit the different components of a thought experiment, they facilitate the task of detecting potential flaws. As a matter of fact, since, as already pointed out, the propositions 1-5 are jointly inconsistent, at least one of them has to be rejected in order to solve the paradox. This allows us to classify thought experiments according to which member of the set has to be rejected, while at the same time providing us with a reliable criterion we can use in order to separate good thought experiments from bad ones. Good or effective though experiments will indeed be those in the case of which the proposition to be rejected in order to solve the corresponding paradox is the first member of the set.

To summarize, then, the advantage of Sorensen's reductionist move is twofold. While on a *systematic* level the schema he proposes constitutes a general mould into which single thought experiments can be cast, on an *epistemic* level it provides us with a straightforward and profitable criterion which can be readily applied in order to assess whether a single thought experiment has been successful.

1.5.5 *Häggqvist's holism*

The position I would like to close my survey with has been advocated until very recently by the Swedish philosopher Sören Häggqvist[92]. Häggqvist's own reflections on the topic can be said to have developed further the approach to thought experiments first put forward by Roy Sorensen, whose views I considered in the last section. Nonetheless, I believe that it would be utterly unfair to label Häggqvist's ideas as generally "Sorensenian", as it were. As a matter of fact, far from simply commenting on the results of his ingenious predecessor, Häggqvist, as we shall see, has made a highly valuable contribution to the understanding of our present topic by insightfully elaborating further on Sorensen's original formal framework. For this reason, I think that his achievements fully deserve a treatment of their own.

Häggqvist's general epistemological aim is that of contributing to the development of a *normative* theory of philosophical thought experiments. On a normative level, he maintains indeed, thought experiments purport to yield *justified belief revision*, i.e. to play a significant cognitive role in rational argumentation and theory choice[93]. The tenability of this claim, Häggqvist feels, naturally calls for assessment. While declaring himself skeptical about the possibility of providing a decision method which would allow us to deductively establish the

[91] *Ibid* (132).
[92] See Häggqvist (1996, 2009a, and 2009b).
[93] See Häggqvist (1996: 12; 2009a: 55).

conclusiveness of any given thought experiment, a satisfactory theoretical framework, in his view, should at least be able to devise a reliable criterion which might be applied in order to discriminate between *successful* thought experiments and *unsuccessful* ones[94]. Moreover, as other inquirers before him, Häggqvist is not interested, nor event confident in the possibility of providing a definition of 'thought experiment' capable of encompassing all the different items to which the term has been applied[95]. Accordingly, he starts by circumscribing the phenomenon he finds most worth of investigation. He characterizes a *thought experiment* in the following words:

> "Something functioning, or intended to function, as an experiment, in the following sense. It aspires to *test* some hypothesis or theory. It is performed in thought – and is hence "real" – but need not thereby shun such prosthetic devices as pencil and paper, encyclopaedias, or computers"[96].

It should be further observed that the *hypothetical* nature of the scenarios appealed to by thought experiments in order to test theories, in his view, should not be taken to entail *non-actuality*, but rather "only that the situation contemplated in the thought experiment is entertained as a possibility in thought"[97].

John Norton, as we saw is section 1.4.2 above, has famously advocated a view now known as *elimination thesis*, according to which, roughly, every thought experiment can, in principle, be replaced by an argument, and thus dispensed with. In order to defend the opposite view, according to which thought experiments, just as ordinary experiments, *cannot* be identified with arguments, Häggqvist appeals to a very natural distinction between linguistic or *truth-valued entities*, such as theories or hypotheses, on the one hand, and non linguistic or *non truth-valued entities*, such as processes, events, or procedures, on the other. Contrary to arguments, he maintains, thought experiments, regardless of their purported epistemic virtues or vices, are neither composed of truth-valued entities, nor may meaningfully be said to be valid in any formal sense[98]. Qua psychological processes, maintains indeed Häggqvist, "experiments […] cannot, properly speaking, have a logical structure"[99].

---

[94] Häggqvist (1996: 12-13).
[95] "It seems quite obvious", he has recently claimed, "that the class of things to which the term 'thought experiment' has been applied does not constitute any natural kind or category". Häggqvist (2009a: 58).
[96] Häggqvist (1996: 15). Häggqvist has recently couched a shortened version of the above characterization in the following words: "hypothetical cases intended to function as experiments, in the following sense: they aspire to *test* hypotheses or theories". Häggqvist (2009a: 57).
[97] *Ibid* (59).
[98] *Ibid* (61).
[99] Häggqvist (1996: 87).

While fending off by this line of reasoning the radically eliminativist bent of Norton's deflationary proposal, though, Häggqvist wishes to retain what he holds to be its most valuable underlying insight. He does this by maintaining that both thought experiments and ordinary experiments, while not exactly *reducible to* arguments, would necessarily need to be *connected with* arguments in order to achieve their ends. This connection, in his view, would be *causal* in the sense that, just as the *physical events* taking place in a laboratory while performing an ordinary experiment *cause* observers to hold certain observational statements true, the *psychological events* taking place within the thought experimenter's head while performing a thought experiment *cause* thought experimenters[100] to hold certain non-observational statements true[101]. Both observational and non-observational statements, according to Häggqvist, would be subsequently employed in arguments whose targets are the theories or hypotheses to be tested. Thought experiments, in other words, would be mental procedures ultimately aimed at generating acceptance, or providing grounds for believing in certain arguments[102]. Correspondingly, he acknowledges the existence of a peculiar family of *thought experiment-based arguments*[103].

As I mentioned earlier, one of Häggqvist's most valuable achievements is that of having developed further a formal regimentation of thought experiments which, compared to Roy Sorensen's original proposal, should be given credit of representing and improvement both in generality and in usefulness. While its increased *generality* comes from the fact that Häggqvist's regimentation, contrary to Sorensen's one, is intended to apply to ordinary experiments as well as to thought experiments, its increased *usefulness* consists in its being closer to the actual dialectical structure of most thought experiments. Indeed, by Häggqvist's own admission, the model he proposes "doesn't assume that thought experiments in fact *manage* to achieve epistemic justification, but it allows us to see how they *aspire* to do so"[104]. While searching for a "feasible idiom" in which important features of thought experiments may be expressed, he adopts Quine's *maxim of shallow analysis*, which reads: "*expose no more logical structure then seems useful* for the […] inquiry at hand"[105]. The schema for arguments connected to ordinary experiments then, according to Häggqvist, would be the following[106]:

[100] Interestingly, Häggqvist invests with the status of 'thought experimenter' both the performer and his audience.
[101] Häggqvist (1996: 87, 2009a: 62).
[102] Häggqvist (2009a: 62, 64).
[103] Häggqvist (1996: 17). My emphases.
[104] Häggqvist (2009a: 56).
[105] Quine (1960: 160), quoted in Häggqvist (1996: 88).
[106] Häggqvist (1996: 99, 2009a: 63-64).

(A) $T \supset (I \supset O)$

I

$\neg O$

$\therefore \neg T$

The first premise of this argument schema claims that a given target theory or hypothesis entails a conditional, namely $(I \supset O)$, predicting that a certain outcome (O) will follow certain initial conditions (I). The second and third premise, in turn, claim that, while the initial conditions hold, the expected outcome does not occur. From this three premises a conclusion is drawn which establishes the falsity of the initial target theory or hypothesis.

The argument schema he proposes for thought experiments would be just a "modalized version" of schema (A)[107], and could be regimented in the following way:

(α) $\Diamond C$

$T \supset (C \mathrel{\square\!\!\rightarrow} W)$

$C \mathrel{\square\!\!\rightarrow} \neg W$

$\therefore \neg T$

"A thought experiment", writes Häggqvist, "is typically designed to invite the conclusion that the target thesis is false"[108]. Accordingly, the above schema is interpreted thus. 'C' is the counterfactual scenario appealed to by the thought experiment. 'T', according to the terminology we established above, is the truth-valued entity, for instance a theory, which the thought experiment aims at rejecting. This target theory, according to the second premise of our argument schema, entails that, were the circumstances described by the counterfactual scenario ('C') to occur, then a certain "weird" state of affairs ('W') would also occur. The third premise claims that, in the circumstances described by the counterfactual scenario, the "weird" state of affairs would not, in fact, occur. Hence, as the conclusion claims, the target theory is false. With respect to the causal connection mentioned above between thought experiments and arguments, in particular, it should now be observed that, according to Häggqvist, thought experiments should be seen as generating belief in the premises of the form $\Diamond C$ and $C \mathrel{\square\!\!\rightarrow} \neg W$. On the contrary, he maintains, it does not seem plausible to maintain that the thought experiment generates belief in the nested conditional $T \supset (C \mathrel{\square\!\!\rightarrow} W)$, which is best seen as prior to the thought experiment itself.

[107] Häggqvist (1996: 99-102, 2009a: 63-64).
[108] Häggqvist (2009a: 63).

Häggqvist follows Sorensen in holding that thought experimental scenarios are explicitly introduced in order to generate an *inconsistent set* of statements, or a *paradox*. With respect to schema proposed by Häggqvist, in particular, the inconsistent set would be the following:

$$\{\ T, C \mathrel{\Box\!\!\rightarrow} \neg W, T \supset (C \mathrel{\Box\!\!\rightarrow} W), \Diamond C\ \}.$$

Now, observes Häggqvist, insofar as the inconsistency of this set can be eliminated only by dropping at least one of its elements, it follows that the argumentative moves available to both the thought experimenter and her critic in order to resolve the anomaly will coincide in number with the members of the set. This consideration, in particular, accounts for the fact that the same thought experiment often leads thinkers to different conclusions. For this reason, every given thought experiment, according to Häggqvist, ought to be seen as connected with *four* different arguments, corresponding to the four different minimal ways of resolving the inconsistency. This means that besides schema (α) above, which is the one corresponding to the thought experimenter's intentions, we will now have three further schemas, namely (β), (γ), and (δ), which correspond to the argumentative strategies available to her critic in order to defend the target theory. The three new schemas would be the following:

(β) T
$\Diamond C$
$T \supset (C \mathrel{\Box\!\!\rightarrow} W)$
$\therefore \neg(C \mathrel{\Box\!\!\rightarrow} \neg W)$

Häggqvist dubs this schema the "*biting the bullet*" strategy for defending the theory attacked by the thought experiment. As a matter of fact, (β) aims at resisting the conclusiveness of the thought experiment by rejecting his core modal intuition, i.e. by denying that, were the situation envisioned by the counterfactual scenario (C) to occur, the "weird" consequence (W) would not follow.

(γ) T
$\Diamond C$
$C \mathrel{\Box\!\!\rightarrow} \neg W$
$\therefore \neg\ (T \supset (C \mathrel{\Box\!\!\rightarrow} W))$

This, in turn, is the argument schema Häggqvist calls "*irrelevance*" strategy. (γ) aims at defending the target theory (T) by denying that it is committed to the "weird" consequence's being true in the counterfactual scenario (C), i.e. by claiming that the target theory does not entail the counterfactual conditional (C □→ W). This is generally due to the fact that the counterfactual scenario, according to the defender, would be too far-fetched to be "relevant" for the theory under attack.

(δ) T
T ⊃ (C □→ W)
C □→ ¬W
∴ ¬◊ C

This last schema has been called, , for obvious reasons, "*impossibility*" strategy. (δ)'s defense of the target thesis, that is, points to the fact the scenario appealed to by the though experiment is indeed impossible. From this impossibility, the defender claims, both the counterfactual advocated by the thought experimenter (C □→ W) and it opposite (C □→ ¬W) can be drawn.
At the beginning of this section I mentioned the fact that a satisfactory theoretical framework, according to Häggqvist, should at least be able to devise a reliable criterion which might be applied in order to discriminate between *successful* thought experiments and *unsuccessful* ones. In the light of the above four schemas, it is now possible to introduce the criterion he has proposed. Insofar as a thought experiment purports to achieve justified belief revision, a *successful* thought experiment, according to Häggqvist, will be one in which "the premises of a regimentation with the form (α) are all justified"[109]. A failed or *unsuccessful* thought experiment, on the other hand, will be "one whose regimentation as an instance of (α) is such that its premises are not all justified"[110]. Now, the question as to whether and to which extent the premises of a thought experiment *can* be justified will not be addressed here. The reason is that this the problem cannot be tackled without previously introducing further considerations concerning the unavoidably modal nature of all thought experimental claims, which will be dealt with in the third chapter.

[109] *Ibid* (2009a: 68).
[110] *Ibid* (2009a: 68).

2. Thought experiments in philosophy

2.1 *Methodological concerns*

If it were allowed to squeeze a whole season of reflection on philosophy into a few paragraphs, our bonsai historiographical opus would probably be something like the following.

Along the first three decades of last century, a self-styled linguistic formulation of empiricism known as *logical positivism*[111] conclusively upgraded philosophy from his medieval position as handmaid of theology to a new and promising role as handmaid of a purportedly much more respectable owner, science. The logical empiricists, in particular, substituted specific preconditions of 'cognitive significance' to the well known Kantian preconditions of 'knowledge', i.e. they imposed standards that any proposition had to meet in order to be rightfully considered cognitively *meaningful*[112]. Faithful to their own avowed empiricism, and following a tradition that can be traced back to David Hume, they divided all meaningful propositions into two general classes: propositions concerning 'relations of ideas', which they called *analytic*, and propositions concerning 'matters of fact', which they called *empirical*. Both classes were subject to specific methods of validation, the details of which do not need concern us here.

To be relevant for our purposes is rather the fact that, contrary to empirical propositions, analytic propositions, as for instance those of logic or mathematics, were taken to be *necessary*, i.e. not revisable in the light of further experience. They enjoyed, that is, the epistemic status of certainty. This status, however, came at a remarkably high price. Analytic propositions were indeed considered certain, i.e. not empirically defeasible, precisely because, trivially, they were not taken to be claims about the empirical world at all. Their proper function, as it has famously been put, was rather to "record our determination to use symbols in a certain fashion"[113]. Their necessity, in other words, was taken by the logical empiricists to be a merely linguistic one.

In a sense, the program pursued by the logical empiricists represented at the same time both a gain and a loss for philosophy. The gain, on the one hand, consisted in the fact that philosophical *methodology*, as Leibniz had dreamed three centuries earlier, had finally reached a level of rigor

[111] In what follows, the labels 'logical positivism' and 'logical empiricism' (and the corresponding forms 'logical positivists' and 'logical empiricists') are used interchangeably.

[112] As we shall see, this is precisely the sense in which, as it has been observed, the same tradition "made epistemology into a philosophical theory of scientific language". See Rosenberg, (2005: 24).

[113] Ayer (1952: 31).

comparable to that of mathematics. The loss, on the other hand, seemed to lye in the fact that, contrary to what the ingenious German philosopher had hoped, the *subject matter* of philosophy ceased once and for all to be the world itself, to become instead the various *descriptions* of that world produced by the natural sciences. A striking consequence of this self-proclaimed "logical outcome"[114] of British empiricism was indeed that claims formerly taken to be about the *world* were now found by its enthusiastic supporters to be in fact, although not explicitly, statements about *language*[115].

This seemed the be the case, in particular, of all meaningful *philosophical* propositions. Since philosophical propositions did aspire to some sort of necessity, and since, as we just saw, the only existing kind of necessity was taken by the logical empiricists to be a linguistic one, i.e. one concerning 'relations of ideas' rather than 'matters of fact', philosophical practice did not seem to have any other choice but to give up its age-old ambition of contributing directly to the knowledge of the empirical world. If a philosopher wants to contribute to human knowledge, Ayer had confidently declared, he "must […] confine himself to works of clarification and analysis"[116], and the proper subject matter of such practice were, needless to say, the "factual" propositions produced by the natural sciences[117]. To make the same point from a different, perhaps opposite angle, we might as well say that, according to the same tradition, in so far as any claim, in order to be *meaningful*, must either be empirically verifiable or analytic, i.e. true in virtue of its meaning, and insofar as most philosophical claims were obviously not empirically verifiable, philosophical analysis was clearly not in the business of producing any knowledge of the empirical world, but rather in that of providing us with the logical-linguistic *conditions* which had to be met if such knowledge was ever to be achieved[118]. Philosophy in other words, borrowing terminology from Rudolf Carnap, was to become a *logic of science*. It had to be replaced, that is, by a "logical syntax" of the language of the latter, i.e. by a formal theory of its linguistic forms[119]. Failure to fully appreciate this point, according to the same tradition, had

[114] *Ibid* (31).
[115] Carnap (1971 [1934]), in particular, elaborated on the idea of linguistic questions "on the sly".
[116] Ayer (1952: 51).
[117] The function of philosophical propositions, wrote Ayer, is "to clarify the propositions of science by exhibiting their logical relationships, and by defining the symbols which occur in them". *Ibid* (32).
[118] "The propositions of philosophy", wrote Ayer elsewhere, "are not *factual*, but *linguistic* in character – that is, they do not describe the behaviour of physical, or even mental, objects; they express definitions, or the formal consequences of definitions". *Ibid* (57), my emphasis. Interestingly, this was also taken to be the reason why, as Ayer points out, philosophy and science "cannot conceivably contradict one another". *Ibid* (57). On a similar basis, Ludwig Wittgenstein famously and firmly resisted the possibility of counting philosophy amongst the natural sciences. See Wittgenstein (2009 [1953]), sections 89-113.
[119] See Carnap (1971 [1934], § 1, § 72, § 73). Central to the idea of a 'logic of science', according to Philip Kitcher, were the following two contentions: (1) "that methodological principles can be formulated is ways that emulate Frege's preferred mathematical idiom; and (2) "that such methodological principles apply independently of subject matter". See Kitcher (1992: 57).

brought over the centuries to notoriously interminable and unfruitful philosophical disputes, sparked off by the literally meaningless propositions of most western metaphysics.

A thoroughly informed critical reconstruction of the multifarious and complex reasons which finally brought to the dismissal, or at least to the substantial revision, by most philosophers, of the original logical empiricist program, certainly does not fall within the much narrower scope of the present work. It will be sufficient, from the point of view of our discussion, to point out that, while retaining the logical-semantic apparatus relied upon and further developed by the logical empiricists, and valuing it as one of the major contributions ever made to both clarity and rigour of philosophical discourse, most leading figures of the subsequent philosophical tradition, starting from the fifties, have at least partially given up the idea that philosophical practice ought to be envisioned as a *uniquely* linguistic endeavour[120]. One of the most relentless critics of logical empiricism for instance, Willard Van Orman Quine, as it is well known, regarded philosophical work as continuous with the work of science, and granted to ontological questions a status "on a par with questions of natural science"[121]. According to the same author, one of the consequences of abandoning what he regarded as two ill-founded "dogmas" of modern empiricism, namely reductionism and the existence of a sharp analytic/synthetic distinction, would be "a blurring of the supposed boundary between speculative metaphysics and natural science"[122].

Even if one disagreed with Quine's prediction, it would still be difficult to deny that one of the long term consequences of the gradual waning of the logical empiricist program has been, as a matter of historical fact, the opening of a lively new season of metatheoretical reflection. Over the last three decades, as it appears, an increasing number of philosophers have felt the pressing need to address questions concerning the nature and limits of philosophical inquiry, as well as to clarify its possible relations to other disciplines, such as empirical psychology for instance[123]. As a consequence of this general tendency, philosophy has entered a phase of deep revision of what it was previously, and often uncritically, taken to be its standard methodology. In recent years, a growing amount of philosophical literature has been devoted to the analysis of current philosophical practice[124], with an eye toward its several epistemological implications.

---

[120] "Translatability of a question into semantical terms", wrote Quine, "is no indication that the question is linguistic". Quine (1980 [1954]: 54). For a recent and thoroughly argued pronouncement in a similar sense see also Williamson (2007).

[121] Quine (1980 [1954]: 45).

[122] *Ibid* (20).

[123] This is the case, for example, of the several naturalistic approaches to epistemological issues put forward over the last decades, on which see Kitcher (1992).

[124] The editors of one of the most representative surveys of this literature, for instance, explicitly mention in the preface to their anthology that their central goal was that of initiating "a self-examination of philosophical method that we believe is long overdue". See DePaul and Ramsey (1998: x). The movement which has been recently

Starting from the seventies, original and substantial work has been made in several areas of analytic philosophy, such as philosophy of mind, philosophy of language and epistemology for instance. A striking methodological feature of this large body of literature is the confident and abundant use of a practice which, when taken at face value, seems to fly in the face of the stubborn anti-psychologism notoriously professed by the previous philosophical tradition[125], namely *thought-experimental reasoning*. Although this practice, as we shall see in the next section, is not new to modern philosophy, most contemporary analytic philosophers seem to assign to hypothetical or counterfactual scenarios, and to the intuitions generated therefrom, a decidedly unprecedented cognitive weight within their theoretical inquiries. "Thought experimentation", some has indeed gone as far as claiming, "has come to supplant meaning analysis as the distinctive method of contemporary analytic philosophy"[126]. Regardless of the tenability of the latter claim, thought experiments have undeniably played a fundamental role in most philosophical theorizing over the last few decades, and this centrality has sparked off a host of extremely interesting epistemological explorations concerning what have been called the powers and limits of imaginary cases[127].

In the following section, I will present a few examples of philosophical though experiments which, I believe, may be regarded as paradigmatic. While indeed, on the one hand, they have generated lively and not yet concluded debates within the analytic community, they display, on the other hand, a number of typical methodological features which I will try to analyse both in the present and in following chapter.

2.2 *Paradigm instances*

2.2.1 *Mind swaps*

In the second book of his *Essay*, concerning 'ideas', we find one of the fathers of modern empiricism, John Locke, grappling with the general notion of *identity*, and with the problem of *personal identity* in particular. Since this last subject, as it has been put, "has probably exploited the method [of thought experiments] more than has any other problem area in philosophy"[128], it

---

referred to under the label of 'experimental philosophy', it could be further observed, has issued directly from the process of methodological self-examination we are considering. See Knobe and Nichols (2008).

[125] On psychologism, see Kusch (1995, 2011).

[126] Horowitz and Massey (1991: 1).

[127] Szabó Gendler (2000).

[128] Wilkes (1988: 6).

will certainly serve our present purposes to begin by restating briefly Locke's views on the topic and to dwell on the nature of the arguments he deemed apt to support them.

According to Locke, both our ideas of *identity* and *diversity* originate when: "[…] considering anything as existing at any determined time and place, we compare it with itself existing at another time"[129]. In such occasions, explains Locke, we are compelled to acknowledge the existence of the above relations by the fact of "never finding, nor *conceiving* it possible, that two things of the same kind should exist at the same place at the same time"[130]. It is now the case that *substance*, *man*, and *person*, according to Locke's tripartite ontology[131] and to his ideational views on meaning, "are three names standing for three different ideas"[132]. It follows that each one of these ideas will enjoy a different identity criterion (*principium individuationis*), insofar as "such as is the idea belonging to that name, such must be the identity"[133].

An animal, for instance, is a living organized body, and Locke does not seem to have any doubts concerning the fact that, if not direct observation, then at least what he calls an "*ingenious* observation"[134] cannot but finally lead us to establish that "the idea in our minds of which the sound *man* in our mouths is the sign, is nothing else but of an animal"[135]. Although Locke is not explicit on this point, one might happen to wonder whether there is any difference between a mere observation and what is here referred at as an "*ingenious* observation". A difference, that is, capable of justifying his preferring the latter expression to the former. I think the following passage suggests a plausible answer:

> "I think I may be confident, that whoever should see a creature of his own shape and make, though it had no more reason all its life than a cat or a parrot, would call him still a *man*; or whoever should hear a cat or a parrot discourse, reason, and philosophize, would call or think it nothing but a *cat* or a *parrot*; and say, the one was a dull irrational man and the other a very intelligent rational parrot"[136].

A dumb man then, according to Locke, stays a man; a graduated parrot, stays a parrot. The identity criterion of both man an animal seems therefore to be, in Locke's opinion, the

[129] Locke (1978: 182).
[130] *Ibid* (183). My emphasis.
[131] See Noonan (1989: 33-34).
[132] Locke (1978: 187).
[133] *Ibid* (186).
[134] *Ibid* (187).
[135] *Ibid* (187). My emphasis.
[136] *Ibid* (187).

possession of a body[137]. It is vitally important for our purposes to stress the fact that Locke, besides recording his own private intuition, declares himself confident that other people will very likely share it. The present case displays indeed a characteristic feature of his writing, insofar as he almost always introduces his fictional scenarios by means of phrases such as "it is evident", "every one sees that", "everyone finds that", "everyone one who reflects will perceive that", "every intelligent being must grant that", or "everyone would say that". I take the abundant use of such expressions to play more than a mere rhetorical function within the economy of his argument and to suggest that Locke explicitly meant to draw attention on the compelling nature of certain intuitions[138].

This being said of the notion of *man*, we now need to follow the same procedure in order to establish what *personal* identity consists of. We need, that is, to establish what kind of idea the name *person* stands for. Here is Locke's own proposal:

> "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places; which it does only by that *consciousness* which is inseparable from thinking, and as it seems to me essential to it; it being impossible for anyone to perceive without perceiving that he does perceive. When we see, hear, smell, taste, feel, meditate, or will anything, we know that we do so"[139].

It is precisely in virtue of *consciousness* then, in the sense of *introspection*[140], that "everyone is to himself that which he calls *self*"[141]. The spatial and temporal bounds of the self, according to Locke, coincide with those of introspection. A unity and continuity of consciousness, in other words, often referred at as the "Lockean condition"[142], would be essential to personal identity.

Our consciousness though, by Locke's own admission[143], is often interrupted, for instance during sleep, and this makes it legitimate to raise a further question, namely: is the self sameness of *substance*, too? Is it, to put it in his own words, "the same identical substance which always thinks in the same person"[144]? This last question, in particular, is further analyzed by Locke into the two following subquestions[145]: (1) If 'the substance which thinks' is changed, can we have

---

[137] "I presume it is not the idea of a thinking or rational being alone that makes the idea of a man in most people's sense, but of a body, so and so shaped, joined to it", *Ibid* (187).
[138] Alvin Goldman has recently endorsed a similar reading of Locke's writing. See Goldman (2007: 1).
[139] Locke (1978: 188).
[140] I follow here Perry, insofar as this last word seems to me to capture well in current terminology the sense in which Locke uses terms such as 'reflection' or 'inner sense'. See Perry (2008, 13).
[141] Locke (1978: 188).
[142] See Wilkes (1988: 103).
[143] Locke (1978: 188-189).
[144] *Ibid* (189).
[145] *Ibid* (190-191).

the same person? (2) If 'the substance which thinks' stays the same, can we have different persons?

Question (1), Locke tells us, can be restated by asking whether it is possible "that that may be represented to the mind to have been, which really never was". Dreams, after all, provide us with clear cases of representations which lack what he calls "reality of matter of fact". Their very existence, therefore, seems to render the situation envisioned in question (1) a genuine possibility. A possibility, that is, which cannot be a priori excluded "till we have clearer views of the nature of thinking substances"[146]. Question (2) in his turn can be restated, in a similar fashion, by asking "whether the same immaterial being […] may be wholly stripped of all the consciousness of its past experience" and "beginning a new account from a new period, have a consciousness that cannot reach beyond this new state"[147].

In general though, Locke tells us, similar questions, concerning *possibility*, are "difficult to conclude from the nature of things"[148]. They cannot be settled, one would like to add once more, by mere observation. A further 'ingenious observation', that is, seems to be required, and this is indeed where Locke introduces the famous thought experiment of the prince and the cobbler. Here are his own words:

> "Should the soul of a prince, carrying with it the consciousness of the prince's past life, enter and inform the body of a cobbler, as soon as deserted by his own soul, *every one sees* he would be the same *person* with the prince, accountable only for the prince's actions: but who would say it was the same *man*?"[149].

Despite the fact that ordinary language tends to systematically conflate the two notions then, the idea of *man* and the idea of *person*, according to Locke, ought to be kept separate. Indeed, insofar as it is consciousness alone that "unites existence and actions […] into the same person", maintains Locke, "it matters not whether this present self be made up of the same or other substances"[150]. Consciousness, to conclude, is what confers identity to the person, regardless of its being joined to one or several substances. Substance, that is, can vary without change of personal identity, provided that "the consciousness of past actions can be transferred from one thinking substance to another"[151].

---

[146] *Ibid* (191).
[147] *Ibid* (191).
[148] *Ibid* (191). This passage seems to me to imply that modal facts or modal notions, according to Locke's empiricism, do not belong to "the nature of things".
[149] *Ibid* (193). My emphases.
[150] *Ibid* (193-194).
[151] *Ibid* (190).

2.2.2 *The categorial hypothesis*

At the very outset of his wonderfully written introductory textbook on *modality*, Joseph Melia tries to show the apparent unavoidability of positing the existence of a large class of *de re* modal truths about the world over and above the purely categorial ones, in order to account for the meaning of our modal thought and talk. He does this by means of a simple yet, in my opinion, very well designed thought experiment[152].

The reader is asked to imagine being in possession of a complete *theory of the world*. This theory, explains Melia, enjoys the three following basic features. Its language contains a *name* for every object, that is for every single thing existing in the universe, and a *predicate* for every property which is actually instantiated in it. Everything our theory says about the universe, we are further asked to assume, is *true*. Our theory, in other words, accurately lists *all* there is in the universe. It further provides us with *true* descriptions of what every object is like, down to its most recondite details, and of the relations in which any object stands with respect to any other or others.

At this point Melia asks the crucial question: "Does *every* truth appear within the theory? Would the theory account for *every* single matter of fact?"[153]. Giving a positive answer to this question, in his view, would amount to endorsing what he calls *the categorial hypothesis*. And yet, he maintains, "a number of philosophers believe that such a theory would *not* be a theory of *everything*"[154]. Many of us, that is, once put in front of a similar scenario, would tend to resist, upon reflection, the categorial hypothesis and would seem rather willing to grant that there *must* indeed be a whole class of truths about which our purportedly *complete* theory of the world is utterly silent, a class of truths, that is, that our theory simply lacks the linguistic resources to describe.

An easy way to see this, is to consider the following two *true* sentences[155]:

(1) Claudio is an ontologist

(2) Claudio is a human being

[152] See Melia (2003: 1-4).
[153] *Ibid* (1).
[154] *Ibid* (2).
[155] I am adapting here Melia's own examples. See Melia (2003: 2).

Since (1) and (2) both ascribe a certain categorial property to Claudio, namely the property of being an ontologist and the property of being a human being, it follows, by hypothesis, that they both have to appear in our theory. And yet (1) and (2) seem to concern, so to speak, different *ways* in which Claudio possess or instantiates the properties that are being ascribed to him, some of which are *contingent*, whereas others are *necessary*. As a matter of fact, we might plausibly reason, our friend Claudio *could* have certainly been, or thought of himself as being, a theoretical physicist instead of an ontologist, without thereby ceasing to be what he is, namely Claudio. On the contrary, or so at least our ordinary intuitions seem to go, he *could not* have failed to be a human being, without thereby ceasing to be our friend Claudio.

This difference between *essential* and *accidental* properties, according to Melia, would not be *linguistic* in nature. It would not merely concern, that is, the concepts we happen or decide to use in order to describe our world, but would rather tell us something substantial about the way in which the world *is*, something, moreover, the truth of which most of us would be willing to recognize as necessary. Melia's point is that our purportedly complete theory of the world cannot, for the reasons we have seen, possibly capture this difference. It follows, according to the same author, that in order to capture this difference we must acknowledge the existence of class of *de re* modal truths over and above the purely categorial ones, and this gives us reason to doubt the correctness of the categorial hypothesis.

2.2.3 *Mary*

On a standard and very broad construal of the term, *physicalism* is the metaphysical view according to which everything that exists is either physical or supervenes on the physical. This claim, in particular, entails that, were physicalism to be true, then an ideally complete physical account of what our world is like would necessarily encompass all knowable truths about it. The same idea, many have felt, can be profitably couched it in terms of possible worlds by saying that "physicalism is true at a possible world *w* iff any world which is a *physical* duplicate of *w* is a duplicate of *w simpliciter*"[156]. Physicalists, in other words, hold that, if we ever came to learn *everything physical* there is to know about our world and whatever is contained in it, human beings included, there would not be *anything* left to know[157].

In a relatively short seminal article published in 1982 Frank Jackson famously launched a frontal attack to physicalism by means of a much celebrated and yet apparently very simple argument,

[156] Stoljar (2009, section 3). My emphasis.

[157] "Physicalism", to put it in Jackson's words, "is not the noncontroversial thesis that the actual world is largely physical, but the challenging thesis that it is *entirely* physical". Jackson (1986: 291). My emphasis.

which was nevertheless bound to spark off one of the most heated debates in contemporary philosophy of mind. Jackson's *knowledge argument*, as he himself called it, was originally meant to develop further the intuition, already shared by several philosophers, according to which no amount of purely physical information would ever be able to account for the introspectively accessible phenomenal sides of our mental lives, also known as *qualia*. Towards the beginning of his article, the "qualia freak" Jackson put down his credo in the form of an explicit challenge:

> "Tell me everything physical there is to tell about what is going on in a living brain, the kind of states, their functional role, their relation to what goes on at other times and in other brains, and so on and so forth, and be I as clever as can be in fitting it all together, *you won't have told me* about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky"[158].

A person, in other words, might come to learn, and therefore know, all the physical and functional facts concerning a specific brain event, down to its most recondite details, and still not *know*, as Thomas Nagel would have it[159], *what it is like* to subjectively undergo that brain event. "Nothing you could tell of a physical sort", could therefore be the catch line of this philosophical stance, "captures the smell of a rose"[160]. If one were willing to grant him this "intuitively obvious" premise, Jackson confidently maintained, one would be forced to admit, by sheer logical necessity, that physicalism is false. Unfortunately, by his own admission, not many people in the philosophical community seemed ready to regard the cornerstone of his argument as *obvious*. This dissent, however, as it is often the case, proved very fruitful, in that Jackson, in order to make his point even more compelling, gave birth to one of the most popular and debated fictional entities of the last three decades: Mary.

Jackson's justly famous thought experiment introduces us to the life and deeds of a rarely talented prisoner scientist, Mary, who, as we are asked to imagine, is forced to investigate the nature of the outside world from a black-and-white room via a black-and-white television monitor. Within her room, according to the scenario, she does not have, nor ever had, access to anything colored. Nonetheless, despite her heavily impaired condition, and in virtue of her extraordinary intellectual capacities, Mary somehow manages to fit all the information that she has painstakingly acquired over the long years of her studies into a fully complete and consistent account of our world in general, and of the neurophysiology of human vision in particular. In

[158] Jackson (1982: 127). My emphasis.
[159] See Nagel (1974: 437).
[160] Jackson (1982: 127).

short, she knows everything there is to know about both brain events and the specific functional roles played by those events. As a consequence, based on her impressive physical, chemical and biological knowledge of both a single human being and her environment, she is perfectly able to predict, to an astonishing degree of precision, each and every event which will occur, from a microphysical up to a behavioral level, once that person is put in front of, say, a midsummer blue sky. To put it in Jackson's own words, Mary, from her room, will be perfectly able to anticipate "just which wave-length combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'"[161]. Now, and this is the crucial point, given the above scenario, if physicalism is true, then there isn't *anything* that Mary does not know.

At this stage of our story the setting is ready for introducing the fundamental question. What will happen, the ingenious thought experimenter asks, once poor Mary is released from her black-and-white room and finally enters our "colored" world? Is she going to *learn* anything new and thereby acquire new *knowledge*? It appears really difficult not to answer in the affirmative to this last question. It just seems overwhelmingly evident, that is, that she will learn, on her release, about an aspect of our world, and of our visual experience of it in particular, which she didn't previously have any access to. Which term indeed, other than *knowledge*, could we be possibly use in order to describe the relation that occurs between Mary and the new experience she is undergoing? And yet, didn't we say that her physical knowledge was, *ex hypothesis*, complete? The following line of reasoning seems therefore inescapable: Mary's knowledge before release was incomplete, therefore physicalism is false. The following is a slightly modified version of the way in which Jackson proposes to reconstruct the structure of his argument[162]:

> Premise 1. Mary (before her release) knows *everything physical* there is to know about other people.
>
> Premise 2. Mary (before her release) does not know *everything* there is to know about other people (because she *learns* something about them on her release)
>
> Conclusion. There are truths about other people (and herself) which escape the physicalist story. Hence physicalism is false.

[161] *Ibid* (130).
[162] *Ibid* (293).

Generalizing from the case of vision to other mental states which arguably display phenomenal features or qualia, we end up with the same result: "The material or physical story about us is not the *complete* story about us"[163].

2.2.4 *Twin-Earth*

Some thirty years ago, by means of a very famous thought experiment, Hilary Putnam proposed a view on meaning which has ever since generally come to be referred at as *semantic externalism*. That ingenious piece of "science-fiction", as Putnam himself called it[164], has had a very deep impact on both philosophy of language and, as we shall see, philosophy of mind over the last few decades. As a matter of fact some have, perhaps slightly too enthusiastically, claimed that "in 1975 Hilary Putnam changed the face of philosophy forever"[165]. Be that as it may, it will certainly be of interest for the present inquiry to introduce briefly the gist of his endeavors.

Once upon a time, so the story goes, there reigned a widely shared view about *concepts* and about our knowledge of them[166]. According to this view, concepts were entities capable of being completely contained in the thinker's mind, which was in its turn construed as a sort of Cartesian private theatre. Since the Middle Ages, continues Putnam, the notion of meaning had been parsed into the two notions of *extension* and *intension*. The *extension* of a term was generally taken to be the set of things to which the term applied, or, alternatively, of which the term was true. By *intension* of a term, on the other hand, was usually meant the concept associated with it. In particular, insofar as concepts were construed as mental entities, meanings (in the sense of *intensions*) were also construed as such. As a consequence, writes Putnam, no one belonging to this tradition "doubted that understanding a word (knowing its intension) was just a matter of being in a certain psychological state"[167]. It was further believed that the concept in the mind *totally* determined the extension of the term associated with it. A concept, in other words, was expected to "*always* provide a necessary and sufficient condition for [a given item to fall] into the extension of the term"[168], and this implied that "sameness of intension entails sameness of extension"[169].

---

[163] Braddon-Mitchell and Jackson (1996: 129).
[164] Although he elsewhere uses the German term '*Gedankenexperiment*'.
[165] Pessin and Goldberg (1996: xi).
[166] According to Putnam, this view is "millennia-old" and can be traced back to the writings of Plato and Aristotle. See Pessin and Goldberg (1996: xv).
[167] Putnam (1996 [1975]: 6).
[168] *Ibid* (6).
[169] *Ibid* (6).

On a sunny day, a dauntless knight from Harvard described a family of scenarios which were intended to show, amongst other things, that the received view rested ultimately on a false theory. This theory, in particular, relied on the two assumptions just recalled, maintaining that (1) "knowing the meaning of a term is just a matter of being in a certain psychological state", and that (2) "the meaning of a term (in the sense of "intension") determines its extension"[170]. An unavoidable consequence of the two joint assumptions was that *mental state determines extension*. As a matter of (logical) fact, if we grant the truth of the two, apparently quite plausible premises, according to which (1) the mental state of the speaker determines intension, and (2) intension determines extension (in the sense of providing necessary and sufficient conditions for membership in the extension), it follows, by modus ponens, that mental state determines extension.

Finding this consequence utterly unpalatable, Putnam set out to show that the two premises of the above syllogism were in fact mutually inconsistent, that they could not both be true at the same time. To set the stage for (what he regarded as) the fatal blow, he reasoned as follows: if it is possible to show that two subjects can be in the same mental state "even though the extension of the term A in the idiolect of the one is different from the extension of A in the idiolect of the other"[171], then the seemingly unassailable inference breaks down, and we are forced to give up, i.e. to acknowledge as false, at least one of the two premises. In order to do this, he resorted to the newly rediscovered Leibnizian notion of a *possible world* and gave birth to one of the most heuristically fruitful fictional entities philosophers have been grappling with in the last few decades: Twin Earth. The planet we are asked to imagine is "*exactly* like Earth", but for one little, apparently harmless detail, consisting in the fact that on Twin Earth:

> "the liquid called "water" is not H_2O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as XYZ. I shall suppose that XYZ is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain XYZ and not water, that it rains XYZ on Twin Earth and not water, etc.
>
> If a spaceship from Earth ever visits Twin Earth, then the supposition at first will be that "water" has the same meaning on Earth and on Twin Earth. This supposition will be corrected when it is discovered that "water" on Twin Earth is XYZ, and the Earthian

[170] *Ibid* (6).
[171] *Ibid* (9).

spaceship will report somewhat as follows: "On Twin Earth, the word "water" *means* XYZ"[172].

The above scenario, according to many, has the power to elicit in us strong intuitions concerning the nature of *meaning*. In particular, as I said, it purports to show that it is possible for two speakers to be in the same mental or brain state, "neuron for neuron"[173], and yet for the terms that they use to differ in extension. And since extension, according to Putnam, while not identifiable with, certainly contributes to the *meaning* of a term, it follows that knowing the meaning of a term cannot simply amount to being in a certain psychological or brain state. Meanings, according to Putnam, do not indeed exist "in quite the way we tend to think they do"[174], in the sense that they do not seem to be, and therefore ought not to be construed as, private mental properties. Twin Earth examples, according to Putnam, enable us to realize that knowledge of meanings, contrary to the deeply rooted methodological solipsism of the previous tradition[175], displays both *social* and *environmental* components. Such knowledge would not be possible for a thinker in isolation, but presupposes interactions with other language users, on the one hand, and with the world, on the other. In a nutshell, Putnam's idea is the following: In order for us to *mean* x when we use the term "x" it is neither sufficient nor necessary to share certain mental images or brain states with other speakers. What is necessary is that the (particular) entity referred to by that term *actually be* an (instance of) x. And whether this is the case or not, depends ultimately on empirical research. "Cut the pie any way you like", he famously concludes, ""meanings" just ain't in the head"[176].

2.2.5 *Justified true beliefs*

It is probably true that epistemology, as it has been written, has been mainly driven by "what may seem to be the purely argumentative power of counterexamples"[177]. Nonetheless, amongst these counterexamples, some have happened to enjoy a remarkably higher favor than others, and thereby immediately to acquire the status of paradigm instances. This is indeed undeniably the case of the two examples by means of which Edmund Gettier, in a famous three pages article

[172] *Ibid* (9-10). My emphasis.
[173] Pessin and Goldberg (1996: xvii).
[174] Putnam (1996 [1975]: 3).
[175] Putnam defines *methodological solipsism* as the assumption according to which "no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom the state is ascribed". *Ibid* (7).
[176] *Ibid* (23).
[177] Gutting (2009: 51).

published in 1963, rejected a widely shared analysis of knowledge, the refutation of which, if we are to believe what we are told, "was accepted almost overnight by the community of analytic epistemologists"[178]. Whether the unusual success of Gettier's cases directly contributed to, or was rather partly explained by, contemporary philosophers' growing interest in thought-experimental methods is perhaps an interesting matter for historians to settle. From the point of view of the present discussion anyways, Gettier's counterexamples certainly represent a privileged case to end our short list of philosophical thought experiments with. In fact, some have recently gone as far as to maintain that "to determine whether Gettier's thought experiments succeed is in effect to determine whether there can be successful thought experiments in philosophy"[179].

Borrowing terminology from James Robert Brown[180], we might say that Gettier's counterexamples constitute a clear-cut case of *destructive* thought experiment, in that they are not properly aimed at establishing a new definition[181] of knowledge, but rather at rejecting a received and generally accepted one. According to the received view, which might be regarded as the standard analysis of knowledge up to Gettier's paper, being a justified true belief is a *necessary* and *sufficient* condition for being knowledge. This widely shared tripartite analysis of knowledge, that is, holds that, in order for an epistemic agent *S* to know (i.e. to be in the relation of 'knowing' to) a proposition *P*, it is both necessary and sufficient that the three following conditions be satisfied: (1) *P* is true, (2) *S* believes that *P*, and (3) *S* is justified in believing that *P*.

Despite the undeniably high plausibility of the above claim, Edmund Gettier thought otherwise. Being a justified true belief, according to him, although necessary, was *not* a sufficient condition for being knowledge. It seemed to him just as plausible, that is, that a person may be fully justified in believing something without thereby automatically being in the relation of 'knowing' to the object which she believed. The intuitive plausibility of a similar situation rested, according to Gettier, on the previous acceptance of two basic assumptions, which he seemed to take for granted, concerning the notion of "justification". The only sense of "justification" in which a person's *justified* belief is a necessary condition of his knowledge, according to the first assumption, is the sense in which someone can be justified in believing something false. Moreover, for any proposition *P*, according to the second assumption, "if *S* is justified in believing *P* and *P* entails *Q* and *S* deduces *Q* from *P* and accepts *Q* as a result of this deduction,

[178] Williamson (2007: 180).
[179] *Ibid* (180).
[180] Brown (1991b: 34). See section 1.4.1 above.
[181] In what follows, the words 'definition' and 'analysis' are used interchangeably.

then *S* is justified in believing *Q*"[182]. If one were willing to accept these two assumptions as obvious, maintained Gettier, then cases would be certainly possible in which a justified true belief intuitively would *not* seem to constitute knowledge. Applying the term 'knowledge' to such cases, to put it in Russell words, "would not accord with the way in which the word is commonly used"[183].

In order to show this, Gettier used two puzzling examples, which spawned one of the most hair-splitting yet prolific epistemological debates in recent analytic philosophy. In the first example we are asked to consider the following possible set of circumstances. Two men named Smith and Jones have applied for the same job, and Smith has evidence for the following proposition:

(a) Jones is the man who will get the job, and Jones has ten coins in his pocket[184].

Smith is further aware of the fact that (a) entails the following proposition:

(b) The man who will get the job has ten coins in his pocket.

As a consequence, accepting (b) on the grounds of (a), Smith is justified in believing that (b) is true. Suppose now that, unknown to Smith, he (Smith) is the one who has been selected for the job, and who further *happens* to have ten coins in his pocket. Were a similar situation to occur, the following three conditions, according to the two assumptions recalled above, would obviously be satisfied: (1) (b) is *true*, (2) Smith *believes* that (b) is true, (3) Smith is *justified* in believing that (b) is true. Unfortunately for the standard analysis, Gettier points out, despite the fulfilment of the three conditions, it is just as obvious that Smith cannot be said to *know* that (b) is true.

The second example, presented here in a somewhat simplified version, requires us to consider a similar scenario. Our imaginary epistemological guinea pig, the highly respectable Mr. Smith, has again evidence for the following proposition:

(a) Jones owns a Ford[185].

[182] Gettier (1963: 122).
[183] Russell (1912: 205). Although, in Russell's case, to fly in the face of established linguistic use would be the application of the term to a mere true belief.
[184] We are free to imagine, Gettier suggests, that Smith has come to know from a reliable source that Jones will be selected for the job, and that Smith has actually counted the coins in Jones' pocket.
[185] The evidence this time may be taken to consist in the fact that Smith has repeatedly seen Jones driving a Ford over the years.

At this point, a further acquaintance of Smith, Brown, enters the thought experimental scenario. In the absence of any evidence whatsoever concerning the current geographical location of Brown, Smith entertains in his mind the following proposition:

(b) Either Jones owns a Ford, or Brown is in Barcelona.

Smith is aware of the fact that (b) is entailed by (a) and, having inferred the former from the latter, is justified in believing that (b) is true. Suppose now again that as a matter of fact, and unbeknownst to the poor Smith, Jones does not *own* a Ford but drives a rented car and that Brown, "by sheer coincidence", really *happens* to find himself in Barcelona. Even in this case then, the conditions required by the standard analysis would be certainly met, since: (1) (b) is *true*, (2) Smith *believes* that (b) is true, (3) Smith is *justified* in believing that (b) is true. Nonetheless, as in the previous example, despite the apparent plausibility of the standard analysis, we would not be willing to admit that Smith really *knows* that (b) is true.
In both cases, it seems, Mr. Smith's beliefs are in order, in that they are both arguably justified and true. And yet, as Gettier's scenarios are intended to show, our intuitions are, at least *prima facie*, stubbornly reluctant to grant them the status of knowledge. Being a justified true belief therefore, concludes Gettier, contrary to what the standard tripartite analysis holds, is not a sufficient condition for being knowledge. Q. E. D.

2.3 *Intuitions we live by*

Towards the beginning of Saul Kripke's epoch-making lectures on *Naming and necessity* we find the following Neo-Cartesian pronouncement:

> "Some philosophers think that something's having *intuitive* content is very inconclusive evidence in favour of it. I think it is very heavy evidence in favor of anything, myself. I really don't know, in a way, what more conclusive evidence one can have about anything, ultimately speaking"[186].

Starting with Plato, intuitions have been typically appealed to by philosophers in developing their arguments. It seems indeed difficult to deny that, even if one were reluctant to grant them any evidential value or conclusive weight within philosophical debates, intuitions, as it has been

[186] Kripke (1980: 42). My emphasis.

put, would certainly still constitute "the inevitable starting point of any intellectual inquiry"[187]. In virtue of this fact, some have moreover felt that our disagreements concerning both nature and epistemic authority of intuitions fundamentally reflect a struggle for the preservation of philosophy as an autonomous field of inquiry[188]. Be that as it may, these highly controversial epistemic sources have figured prominently in most contemporary analytic philosophers' argumentative strategies and are therefore certainly worth a closer look at.

Philosophical thought experiments, as the above examples show, rely heavily on intuitions and call therefore naturally for a careful discussion of both their nature and the limits of their proper use. In fact, insofar as skepticism about thought experiments can arguably be taken to rest ultimately on skepticism about certain purportedly wrongheaded uses of intuitions, attempted general assessments of the proper epistemic status of thought-experimentally generated beliefs will hinge largely on a brief preliminary look at the current debate concerning the use of intuitions in philosophical methodology.

A striking feature of the latter debate is that, while almost every philosopher seems willing to grant intuitions some role or other in the process of generating, supporting, or updating our beliefs, a remarkable amount of disagreement still reigns over their specific nature. Granted indeed that intuitions are *psychological* states, one may want to ask, what *kind* of psychological states are they? On a first approximation, a general yet very plausible answer to this question might go as follows: Intuitions are *doxastic* states not formed on the basis of any explicit (i.e. conscious) reasoning process; they are opinions we find ourselves willing to assent to, while unable to support our assent by means of any explicit argument[189].

Moreover, while accepting this formulation as fundamentally correct, one might also want to raise questions concerning the proper *phenomenology* of such doxastic states. In order to push the inquiry further, that is, one might profitably ask questions such as: What are the salient *features* (if any) which allow us to distinguish intuitions from other doxastic states? Or else, what are the possible *contents* of an intuition, i.e., roughly, what kinds of things can we have intuitions *about*? A casual use of the term may indeed bring one to conflate intuitions with other similar doxastic states, or worst suggest the idea that the range of possible contents of an intuition may encompass just about anything. Having come to grips with similar questions, different authors have proposed several sensibly divergent accounts of the notion of intuition they hold to be most relevant to philosophical theorizing. The account provided by George Bealer, being one of the most thorough and insightful attempts at describing intuitions proper phenomenology, is the one we will focus on in what follows.

---

[187] Gutting (1998: 8).
[188] *Ibid* (7).
[189] This formulation draws on the one provided by Gopnik and Schwitzgebel (1998: 77).

Bealer construes intuitions as *intellectual seemings*. As such, intuitions would be opposed to sensory, introspective, or imaginative seemings, which he refers to as e*xperiential*. "For you to have an intuition that A", Bealer writes, "is just for it to *seem* to you that A"[190]. Intellectual and experiential seemings, according to his view, differ markedly in phenomenology, insofar as their contents cannot overlap. Drawing on Descartes' famous example, Bealer maintains that, while it is possible for a subject to "intuit" the existence of a chiliagon, the very existence of the same polygon cannot be "imagined"[191]. These intellectual seemings, in his view, can further be divided into two subclasses, namely *physical* intuitions, which would typically not present themselves as necessary, and *rational* intuitions, which would. Accordingly, he offers the following, tentative, analysis of rational intuitions: "necessarily, if x intuits that P, it seems to x that P, and that necessarily P"[192].

Bealer, in particular, regards intuition as differing substantially from *belief*. In fact, he maintains, contrary to intuition, this last propositional attitude is usually taken to be highly "plastic". The strength, that is, with which a particular belief is held, is liable to be sensibly increased or diminished by various circumstances[193]. Even in this case, therefore, each state could occur without the other. One could, for instance, "believe" but not "intuit" that Rome is the capital of Italy or that a certain mathematical theorem holds, just as one could "intuit" but not "believe" the naïve comprehension axiom of set theory[194]. A further, illuminating example appeals to a well known law of propositional calculus: "when you first consider one of De Morgan's laws, often it neither seems true nor seems false; after a moment's reflection, however, something happens: it now just *seems* true"[195]. Intuition, in other words, manifestly survives grounded contrary belief, and this is exactly the respect under which, according to Bealer, its phenomenology can be regarded as paralleling that of a common sensorial kind of experiential seemings, namely perceptual illusions. The analogy proposed in this case, as one might expect, is that with the famous Müller-Lyer illusion, where, in Bealer's words, "it still *seems* to me that one of the two arrows is longer than the other […] despite the fact that I do not believe that one of the two arrows is longer (because I have measured them)"[196].

---

[190] Bealer (2002: 73).
[191] *Ibid* (73).
[192] Bealer (1998: 207).
[193] *Ibid* (208). The immediate polemical target here is the view endorsed by Richard Foley, according to which, while not full-fledged beliefs, intuitions ought nonetheless to be regarded as "*belief-like* states", on a par with hunches, forebodings, premonitions and suspicions. Foley's account, as a matter of fact, insists on the role played by the "depth" of an opinion, thereby meaning precisely the strength with which the opinion is held, and distinguishes, accordingly, between "deeply" and "shallowly" held opinions, all of which he takes to be empirically revisable. See Foley (1998: 244-45).
[194] Bealer (1998: 208; 2002: 73). The same example would also apply, in Bealer's view, to other propositional attitudes such as judgements, guesses, and hunches. See Bealer (1998: 210; 2002: 74).
[195] Bealer (1998: 207). My emphasis.
[196] *Ibid* (208).

Bealer further holds intuition to differ markedly from a spontaneous (i.e. unprompted) *inclination to believe*. This phenomenological difference, in his view, is a consequence of fact that intuitions, contrary to dispositional states such as inclinations, would be intrinsically *episodic*. "As I am writing this", Bealer points out, "I have spontaneous inclinations to believe countless things about, say, numbers. But I am having *no* intuition about numbers"[197].

A similar example, in his view, would also exclude the possibility of reducing intuitions to a *raising to consciousness* of some nonconscious background beliefs[198]. While appealing to unconscious background beliefs might indeed help explaining *conscious belief*, the same explanatory pattern, in Bealer's view, would not do the job in the case intuitions as so far construed. The reason is that, on many occasions, the conscious beliefs associated to the intuition would differ from it in truth value. In order to show this, Bealer invites us to reflect upon a situation in which a subject is asked whether the naïve comprehension axiom *and* the rules of classical logic both hold. Confronted by such question the subject would have the conscious belief that that they do not both hold, while having the intuition that they do. Moreover, the same reductionist proposal, according to Bealer, would fall short of accounting for intuitions concerning novel questions, in the case of which there simply are no background beliefs one could appeal to[199].

To sum up then, based on their peculiar phenomenology, rational intuitions, according to Bealer, ought to be carefully distinguished from: experiential seemings, beliefs, inclinations to believe, and raising to consciousness of nonconscious background beliefs. His proposal is therefore to construe intuition as a "*sui generis* propositional attitude"[200], more specifically a sui generis, irreducible *propositional* attitude that occurs episodically"[201].

2.4 *Conceptual analysis*

The use of intuitions about possible cases is generally held to be deeply entrenched in a traditionally practiced philosophical activity, namely *conceptual analysis*. In fact, it is precisely because of its systematic reliance on this highly controversial epistemic source, that conceptual

[197] *Ibid* (209). The polemical target, in this case, is the counterfactual account of seemings put forward by Ernest Sosa, according to which intuition would be best construed as a kind dispositional state, namely as an *inclination* to believe. Sosa's account focuses mainly on linguistic competence. An intuited proposition, in his view, is simply a proposition that *would* be believed if understood. See Sosa (1996: 260). See, in particular, Bealer (1998: 233) for a critique of Sosa's counterfactual account.

[198] The polemical target here is the view endorsed by Hilary Kornblith, which we shall consider later on.

[199] Bealer (1998: 209-10).

[200] Bealer (2002: 73).

[201] See Bealer (1998: 207).

analysis has come to be regarded as a canonical example of *intuition*-based methodology. Before taking a closer look at its typical functioning, it is worth pointing out that conceptual analysis presupposes a specific view on the internal structure of concepts, often referred to as *classical theory of concepts*. According to this view, concepts in general would possess a specifiable definitional structure, i.e. they would be composed of simpler concepts which express the necessary and sufficient conditions a particular object has to satisfy in order to fall under a given concept[202].

The generally acknowledged aim of conceptual analysis is that of determining the *essential properties* of some abstract notion *F*, such as 'knowledge', 'mind', or 'meaning' for instance. Essential properties, on a first, rough approximation, can be understood as those properties in virtue of which an *F* is what it is. In order to identify such properties, philosophers, from Socrates on, have typically started by looking for *definitions*. Giving a definition of *F*, according to the classical theory of concepts recalled above, amounts to providing a set of necessary and sufficient conditions a particular object *a* has to satisfy in order to be rightfully considered an instance of *F*. Once the definition is given, philosophers will typically start looking for *counterexamples* to it.

These counterexamples will be of two general kinds, corresponding to the two ways in which a definition may go wrong. One could, that is, either provide an example of a particular object which *does* satisfy the conditions provided by the definition, but which one would be *unwilling* to count as an instance of *F*, or one could provide an example of a particular object which *does not* satisfy those conditions, but which one would nonetheless be *willing* to count as an instance of *F*. While in the former case the conditions posited by the definition would be shown to be not *sufficient*, in the latter case they would be shown to be not *necessary* for an item *a* to be rightfully considered an instance of *F*. Once a definition is thus undermined by means of specific counterexamples, a new definition will be usually looked for, which is able to account for those counterexamples. The provisional goal of this open-ended dialectical process is that of reaching what we might call a *stable* definition, i.e a definition which appears to lack counterexamples.

It is not hard to see that the practice sketched above, in the process of finding counterexamples to given definitions, relies heavily on *intuitions* concerning the applicability of specific terms (those terms, namely, whose referents are the objects appealed to in the counterexamples) to possible scenarios. A fundamental part of conceptual analysis, that is, to put it in familiar jargon, turns on 'what one *would* (or *should*) say' if something were the case[203].

---

[202] See Margolis and Laurence (2011, section 2.1).

[203] David Papineau has proposed to schematize this practice as follows. In order to undermine a target analysis of the form $\Box \forall x\ (Ax \rightarrow Bx)$, we look for an *A* which, as we put it above, we would not be willing to count as a *B*, and which would therefore fit a schema of the form $\Diamond \exists x\ (Ax\ \&\ \neg Bx)$. See Papineau (2009, section 2.4). With regard to

A definition then, according to what we said so far, is usually accepted or rejected depending on whether it captures or fails to capture our conceptual or linguistic insights concerning the matter at issue. It is indeed a widely shared opinion amongst philosophers that definitions ought to be constantly revised in the light of intuitive judgements. According to many, though, this ongoing revision cannot just be a matter of testing our concepts, and the theories they stem from, against our intuitions, but should rather run in both directions. Conceptual analysis, in their view, should therefore be regarded, and carried out as a continuous process whose two poles would be concepts or theories, on the one hand, and intuitions, on the other.

Adapting to the case of concepts Nelson Goodman's famous account of justification in terms of *reflective equilibrium* amongst beliefs, I propose to describe this process as follows:

> A *definition* should be revised if it does not account for an intuition we are unwilling to neglect; an *intuition* should be neglected if it is not captured by a definition we are unwilling to revise[204].

It is important for our purposes to observe that, contrary to traditional conceptual analysis, this further development seems to presuppose a different view of the internal structure of concepts, often referred to as *theory theory of concepts*. According to this view, "concepts stand in a relation to one another in the same way as the terms of a scientific theory and […] categorization is a process that strongly resembles scientific theorizing"[205]. This way of thinking about concepts, as I would like to suggest, is often implicit in much theorizing on the epistemic role of intuitions[206] and plays a fundamental role in philosophical thought experiments.

---

the same strategy William Ramsey has stressed the role played in conceptual analysis by what he calls *intuitive categorization judgements*. "The process of appraising definitions", he writes, "requires comparing and contrasting the definitional set of properties with *intuitively* judged instances and noninstances of the target concept". See Ramsey (1998: 164).

[204] Goodman's original formulation, as it is well known, is the following: "A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend". See Goodman (1983: 64).

[205] Margolis and Laurence (2011, section 2.2).

[206] George Bealer, for instance, has recently defended a position according to which the aim of what he dubs the *standard justificatory procedure* in philosophical enquiry would be precisely that of reaching reflective equilibrium amongst our intuitions. Such procedure, in his view, would fundamentally consist in "canvassing intuitions, subjecting those intuitions to dialectical critique, constructing *theories* that systematize the surviving intuitions, testing those theories against further intuitions, and so on until equilibrium is approached". See Bealer (1998: 205), my emphasis. The results of such procedure, in his view, would be *collectively* reached and *provisional*. See Bealer (1998: 206).

2.5 *The tribunal of experience*

Gary Gutting has recently maintained that: "a modus tollens argument based on an *utterly obvious* counterexample is *decisive*"[207]. This bold claim immediately raises a problem concerning the epistemic status of our intuitions, which could be formulated roughly as follows: how can we decide whether a certain counterexample's being *obvious* to anyone can be taken to provide *decisive* evidence for anything?

General skepticism about the use of intuitions in philosophical theorizing is not uncommon amongst philosophers. Hypothetical scenarios such as the ones we have been considering above naturally (and reasonably, I believe) invite skepticism. While some authors, on the one hand, seem willing to grant to thought experiments the power to confer *justification* on philosophical beliefs, other authors, on the other, regard the use of such argumentative strategies as a rather persuasive invitation to banish intuitions from our philosophical inquiries once and for all. Given the often wildly far-fetched nature of the fictional scenarios appealed to by thought experiments, some have indeed reasoned, why should we regard the intuitions elicited by them as epistemically warranted judgements?

Over the last century, thinkers of a more or less scientistic bent, have often made a point of signalling the potential *deceptiveness* of our intuitions. While reflecting on the ultimate epistemic foundations of our knowledge of the external world, for instance, Bertrand Russell voiced his concerns in the following words:

> "It is true that intuition has a *convincingness* which is lacking to intellect: while it is present, it is almost impossible to doubt its truth. But if it should appear, on examination, to be at least as fallible as intellect, its greater subjective certainty becomes a demerit, making it only the more irresistibly deceptive"[208].

As a matter of fact, examples of the potential deceptiveness of *empirical* intuitions abound. A simple one is apparently to be found in one of the arguments traditionally used by Aristotelians to prove that the earth was at rest[209]. Their argument can be cast into the form of a counterfactual conditional as follows:

---

[207] Gutting (2009: 55). The second emphasis is mine.
[208] Russell (2005 [1912]: 35). My emphasis.
[209] I draw here on Sorensen (1992: 197-98).

> *If* the earth rotated from west to east, *then* unattached objects (e.g. clouds) should move from east to west, (or, alternatively, …*then* a rock dropped from a tower would land west of the tower).

According to the same Aristotelians, this would be clearly shown by a simple thought experiment. It is evident, so the story goes, that a rock dropped from the mast of a moving ship would land at the rear of the ship. Thinking that it wouldn't would indeed be absurd. As it happened, subsequent experiments proved that the 'evident' fact did not occur at all, and this seems to teach a disturbing lesson about empirical intuitions which is somehow reminiscent of Russell's concerns, namely that their alleged *obviousness*, at times, can prevent people from putting them to the test!

Why should *rational* intuitions, i.e. intuitions about abstract matters of the sort described by Bealer, fare any better? With respect to the study of *meaning*, for instance, some have very reasonably held that "just as our ordinary intuitions about physics provide only a rudimentary starting point for the creation of physical laws […] the fact that our ordinary understanding of the mind presupposes an externalist, supra-individual semantics says very little about whether an externalist semantics is the best choice for a theory of meaning"[210].

Warnings of this sort, moreover, have usually pointed to the fact that most intuitions are likely to display several rather unwelcome epistemic shortcomings, which have usually been held responsible for rendering their deliverances intolerably prone to error. Intuitions, in particular, have typically been regarded as neither *reflectively stable*, nor *universally shared*. While indeed, on the one hand, intuitive beliefs generally do not outlive more careful rational scrutiny, empirical research, on the other, might end up showing that epistemic agents belonging to different cultural groups (or to the same cultural community, for that matter) do not in fact share the same intuitions[211]. For this reason, many have deemed intuitions to be excessively *context-sensitive*, too much dependent, that is, on a number of psychological or cultural factors. Others have seen intuitions as unconsciously and hence unavoidably *theory-driven*. A change in the privileged theoretical framework, according to this line of reasoning, would be very likely to produce markedly different intuitions on the very same subject matter. Similar drawbacks have contributed to depict intuitions as dangerously unreliable sources of knowledge.

These concerns have found strong support, in recent years, in the advancement of cognitive science. As a matter of fact, over the last three or four decades, the epistemic value of intuition-based philosophical methodologies has been radically questioned on strictly empirical grounds.

[210] Gopnik and Schwitzgebel (1998: 85).

[211] Stich (1993) and Weinberg, Stich, and Nichols (2001), for instance, have pursued a similar line of reasoning with respect to epistemic intuitions of subjects belonging to two different cultural groups.

A powerful case has indeed recently been made, according to which a growing corpus of empirical findings concerning the nature of human concepts and categorization judgements would cast serious doubts on the cognitive value of a methodology which, as we have seen, has traditionally played a crucial role in philosophical theorizing, namely conceptual analysis[212]. A troublesome diagnosis has accordingly been formulated, according to which "Western analytic philosophy is, in many respects, undergoing a crisis where there is considerable urgency and anxiety regarding the status of intuitive analysis"[213].

A long tradition in Western philosophy has thought of *categories* as logical entities, membership to which has been usually taken to be, as it has been put, "a *digital*, all-or-none phenomenon"[214]. Several authors, starting from the seventies, have argued that some natural categories would be best thought of as possessing rather an *analog* structure[215]. Elaborating on this view, Eleanor Rosch has proposed to think of categories as internally structured into *prototype* and *nonprototype* members, reflecting an order that ranges from better to poorer examples of a given category. She exemplifies thus the psychological data which her research tries to account for: "As speakers of our language and members of our culture, we know that a chair is a more reasonable exemplar of the category *furniture* than a radio"[216]. Empirical research, rather than *a priori* speculation is obviously needed in order to specify the principles of learning and information processing which determine the formation of prototypes. Nonetheless, Ludwig Wittgenstein's considerations concerning the structure of concepts are acknowledged by Rosh as the immediate philosophical antecedent of such empirical work[217].

Wittgenstein's general idea was that, in order for a word to be consistently used in a language, its referents need not possess common elements, but rather display what he called a *family resemblance*. This kind of relation can be profitably understood as a set of items of the form: ab, bc, cd, de. The feature of this set which is relevant to our discussion, according to Eleanor Rosh and Carolyn Mervis, would be that: "each item has at least one, and probably several, elements in common with one or more other items, but no, or few, elements are common to all items"[218]. The same authors take their work to provide an empirical foundation for Wittgenstein's original idea, insofar as their research would reveal that "subjects can reliably rate the extent to which a member of a category fits their idea or image of the *meaning* of the category name […] and such ratings predict performance in a number of tasks"[219].

---

[212] See, in particular, the articles contained in DePaul and Ramsey (1998).
[213] *Ibid* (x).
[214] Rosch and Mervis (1998 [1975]: 17).
[215] See Rosch and Mervis (1998 [1975]), Rosch (1978), and Smith and Medin (1981).
[216] Rosch and Mervis (1998 [1975]: 17).
[217] See Wittgenstein (2009: § 67).
[218] Rosch and Mervis (1998 [1975]: 18).
[219] *Ibid* (17). My emphasis.

Some have felt that the story sketched above would have, if true, very heavy consequences for conceptual analysis as traditionally practiced. The same empirical results, according to their line of reasoning, would explain in particular why definitional analyses of concepts are typically very controversial. In fact, according to the *prototype theory*, and contrary to the classical view, the way in which we represent concepts would simply lack a clear-cut definitional structure, but would rather display a prototypical and hence probabilistic one[220]. If this were indeed the case, then any attempt at specifying the *essential properties* of an abstract notion *F* in terms of a set of necessary and sufficient conditions that an object *a* has to satisfy in order to be a member of *F* would be radically undermined, insofar as those sets of necessary and sufficient conditions will *never* be able to capture the full range of our intuitions, and hence never become reflectively stable, i.e. immune to further counterexamples.

According to William Ramsey[221], for instance, the abundant use of intuitive judgements would reveal that conceptual analysis as practiced by philosophers is committed to specific, but of course not empirically founded, assumptions concerning the nature of our cognitive system. These assumptions would stem, in his view, from a widely, yet wrongly shared folk-theoretical story about the way in which human beings represent concepts, namely the *classical view of concepts* recalled above. A fundamental tenet of this view, according to Ramsey, would indeed be an empirically unjustified reliance on intuition, construed as a faculty capable of providing access to some sort of *tacit* knowledge concerning the *essence* of abstract notions and their intrinsic structure.

2.6 *Questioning the verdict*

In the light of all this, other philosophers have responded by pointing out that a careful assessment of the general theoretical relevance of empirical findings of any kind for philosophical methodology cannot avoid taking into account the *normative* role traditionally assigned by philosophers to their epistemological endeavours. Overlooking the normative and evaluative aspects of philosophical inquiry, in their view, would run the obvious risk of falling pray to a gross and well known psychologistic fallacy.

Basic normative considerations, as a matter of fact, seem to suggest that what are usually referred to as *typicality* or *prototype effects* of our categorization judgements, what we might call the degree of *F-ness* of an *F*, for instance, are not immediately relevant to the "real nature" of an

[220] See Margolis and Laurence (2011, section 2.2).
[221] Ramsey (1998 [1992]: 164).

F. Conceptual analysis, as it has been frequently argued, would not be after the former, but rather after the latter. From a normative point of view, in other words, the position of someone claiming that the *classical view of concepts* could be taken as "a possible underlying motivation"[222] for practicing conceptual analysis seems just wrong. The claim, it might be plausibly maintained, from a perhaps less generous point of view, just trades on the conflation between *psychological* and *logical* aspects of our cognitive lives. How could the *actual* way in which humans represent concepts, one could in fact reasonably argue, possibly bear on conceptual analysis? While indeed the former is a purely *descriptive* business, the latter, once again, thinks rather of itself as a strictly *normative* one[223].

An early reaction along these lines has been thoroughly articulated in a seminal article by Georges Rey[224], whose central aim was not that of evaluating the empirical results produced by Eleanor Rosch and her associates[225], but rather to provide a general assessment of the *relevance* of those results for the philosophical study of concepts.

Traditionally, observes Rey, concepts have been held to perform several different functions within human cognition beyond pure *categorization*, none of which would have been properly addressed by the afore mentioned empirical findings. Concepts, for instance, would perform what he calls a *stability* function with respect to our ordinary explanations of behaviour, in that they provide "the links between different cognitive states that are 'about the same thing'"[226]. They would further perform, in his view, fundamental *metaphysical* and *epistemological* functions in that they provide the basis for claims about *universals*, on the one hand, and about *a priori* knowledge, on the other.

The incapability of accounting for these functions, according to Rey, would be due to the fact that the theory of concepts developed by cognitive psychologists "hopelessly confuses metaphysical issues of conceptual *identity* with (roughly speaking) epistemological issues of conceptual *access*"[227]. Cognitive scientist, in Rey's view, just as the logical positivists before them, would have illegitimately conflated *metaphysics* with *epistemology*[228]. Their theory of concepts, as a consequence, while certainly revealing important psychological features of people's access to their system of beliefs, would do little or nothing to explain either a concept's

[222] Ramsey (1998 [1992]: 170).
[223] It must be observed that, by Rosch and Mervis' own admission, *family resemblance* would be a *descriptive* principle. Rosch and Mervis (1998 [1975]: 41).
[224] See Rey (1983).
[225] Much of this work is reviewed in Edward Smith and Douglas Medin's 1981 book *Categories and Concepts*, which constitutes the direct polemical target of Rey's considerations. See Rey (1983: 238).
[226] Rey (1983: 242).
[227] Rey (1983: 238).
[228] See Rey (1983: 245).

identity conditions, or the conditions of its *competent use*[229]. As a matter of fact, he suggests, a satisfying account of the these conditions "may not be a piece of *psychology* at all"[230].

An implicit assumption of the work in cognitive science, Rey observes further, holds that "*having a concept* consists in *knowing* the defining conditions"[231]. Contrary to this assumption, he maintains, "recent" work in philosophy of language, due mainly to Hilary Putnam and Saul Kripke, has abundantly showed that *competent use* of a concept by a member of a linguistic community and *knowledge* of its correct definition ought to be thought of and treated as different cognitive performances. Appealing to our ordinary linguistic intuitions concerning proper names and natural kind terms, in particular, these authors would have showed that "*whether or not there actually are defining conditions*, competent users of a concept may still not know them"[232]. The arguments of these philosophers would indeed be precisely aimed at challenging the view "that the proper definition of a term need play some epistemological role"[233]. This, in particular, is the reason why, in his view, producing empirical evidence of the fact that *competent* users of concepts might not be able to provide defining conditions for them, does *not* amount to rejecting, or undermining the classical view of concepts. Such evidence, he maintains, might equally well be explained by acknowledging that "while at least many such natural kind terms *do in fact* have definitions, competent users may simply be ignorant of what they might be"[234].

Sharing similar concerns, George Bealer has more recently championed a form of *moderate rationalism*[235], perhaps placing a stronger emphasis on modal considerations. Central questions of philosophy, according to Bealer, can, *in principle*, be answered without relying on the results of the empirical sciences[236]. This is due to the fact that the "central questions" of philosophy, in his view, would be questions concerning "the nature of *x*", and the answers to these questions, in order to be legitimately considered such, must enjoy the status of *necessity*. While inquiring into things such as the nature of *x*, maintains indeed Bealer, "philosophers do not want to know what those things just happen to be, but rather what those things *must* be, what they *are* in a strong sense"[237]. The epistemic status of conceptual analysis, in his view, would be on a par, for

---

[229] "It is not at all clear", writes Rey, "how people's responses to categorization queries bear upon the question of the identity of the concepts, or even on the conditions under which they are competent to use them". Rey (1983: 241). And elsewhere, perhaps more explicitly, "The fact that subjects might be unclear about whether something falls under a particular concept *will not show anything* one way or another about whether there are metaphysically defining conditions for that concept unless it can be shown that the unclarity is metaphysical and not merely epistemological". Rey (1983: 248). My emphasis.
[230] *Ibid* (238). My emphasis.
[231] *Ibid* (238).
[232] *Ibid* (252).
[233] *Ibid* (253).
[234] *Ibid* (254). My emphasis.
[235] Starting from the late eighties, Bealer has defended this position in a series of papers. See Bealer (1987, 1992, 1994, 1996, 1999).
[236] See Bealer (1998: 202, 209, 213).
[237] Bealer (1998: 204).

instance, with the mathematical investigation of computability in the 1930's[238]. Accordingly, he summarizes his view in what he calls the *autonomy of philosophy* thesis, which he expresses in the following terms:

> "Among the central questions of philosophy that can be answered by one standard theoretical means or another, most can in principle be answered by philosophical investigation and argument without relying substantively on the sciences"[239].

Insofar as the above thesis is a *modal* claim, maintains Bealer, it constitutes a "cognitive ideal"[240]. It follows that, in order to reject it, one must show that the sort of knowledge it appeals to is not only *not available* to humans, but strictly speaking *impossible*. His position seems therefore to be in principle unassailable by empirical means. In particular, he maintains, despite the fact that "many philosophers enjoy the pastime of "intuition bashing""[241] on the basis of the empirical findings of cognitive psychologists, rational intuitions, construed as *seemings*, while certainly corrigible, remain the proper epistemic source of our *standard justificatory procedure* in all *a priori* disciplines[242].

2.7 *Naturalizing intuitions*

As it stands, the debate that we have been outlining so far seems to suggest that the real challenge that philosophical methodology has to face is that of taking into serious account the growing corpus of empirical data coming from cognitive science, while at the same time preserving its normative role. Some have indeed lamented that even long after the experimental findings recalled above have been disclosed "there has been surprisingly little effort […] to articulate a plausible method for philosophical inquiry that does not run foul of current empirical research"[243]. Methodological traditionalists and cognitive scientists, in other words, while investigating the same object, namely *human cognition*, run the serious risk of just talking past each other. This outcome, I believe, can and should be avoided, and a natural way to do this would certainly be that of inquiring into the possibility of there being interesting links between *normative* and *descriptive* aspects of the study of human cognition.

---

[238] *Ibid* (201).
[239] *Ibid* (201).
[240] *Ibid* (203).
[241] *Ibid* (202).
[242] See Bealer (1993: 164-67).
[243] DePaul and Ramsey (1998: ix).

In this regard, the central aim of Philip Kitcher's 1992 seminal paper on *epistemological naturalism* was precisely that of exploring "the feasibility of preserving the normative enterprise within a naturalist framework"[244]. Starting from the work of Frege, the two central tenets of philosophical antinaturalism have notoriously been *anti-psychologism*, on the one hand, and the belief in the existence of a class of *a priori* truths, on the other. Kitcher's brand of naturalism, which he calls *traditional naturalism*, occupies an "uncomfortable middle ground"[245] in that it aims at rejecting both, while at the same time preserving the normative project of traditional epistemology.

Playing a *normative* function, in his view, would amount to achieving *corrigible* formulations of the goals of our cognitive enterprise, and *corrigible* accounts of the strategies for achieving those goals[246]. This requires, in particular, making room for *psychology* within epistemological theorizing. Insofar as human beings are the products of a long evolutionary process, Kitcher maintains indeed, their psychological capacities are highly relevant to the study of human knowledge. It is for this reason that prescriptions for thought, in his view, should be grounded "in facts about how systems like us could attain our epistemic goals in a world like ours"[247].

It is important for our present purposes to observe that, in Kitcher's view, the reintroduction of psychology into epistemology would be reflected by the passage from the first to the second of two different strategies which have been adopted in order to tackle the problem raised by Gettier's much celebrated thought experiment. Central to the first strategy, according to his view, would be the question as to which *logical conditions* a belief has to satisfy in order to avoid Gettier-like counterexamples, while central to the second would rather be the question as to which *belief-generating processes* a belief has to originate from in order to avoid those counterexamples. It cannot go unnoticed that a similar change of strategy presupposes a fundamental transition from a *foundationalist* approach to epistemology to a *reliabilist* one. This transition, according to some, is a fair price to be paid in order to 'make room for psychology'.

Moving from a similar reliabilist approach to justification, for instance, Hilary Kornblith has been trying to embed an account of the role of intuition in philosophical inquiry within a purely naturalistic framework[248]. Insofar as epistemology, in his view, is to be seen as a continuous with the empirical sciences, the practice of appealing to intuitions in "constructing, shaping, and refining […] philosophical views"[249] can and should be accounted for in purely naturalistic terms. Epistemology, in particular, according to Kornblith, would not be some kind of ordinary

[244] Kitcher (1992: 59).
[245] *Ibid* (77).
[246] *Ibid* (58).
[247] *Ibid* (58-61).
[248] Kornblith (1998). His view was further developed in Kornblith (2002).
[249] Kornblith (1998: 129).

language analysis, but the investigation of a certain mind-independent natural phenomenon called 'knowledge', single cases of which, displaying a high degree of theoretical unity, allow us to treat it as a *natural kind*. It is precisely for this reason, he maintains, that by appealing to our intuitions about knowledge we are not inquiring into the nature of same shared yet mind-dependent *concept*, but we rather "make salient"[250] certain instances of the phenomenon we are investigating.

In order to clarify his view, Kornblith appeals to the following analogy. "What we [epistemologists] are doing", he writes, "is much like the *rock collector* who gathers samples of some interesting kind of stone for the purpose of figuring out what it is that the samples have in common"[251]. What we are trying to determine, that is, are the specific traits which confer theoretical unity to the natural kind. Philosophical inquiry, in other words, would be on a par with scientific investigation of natural kinds.

Once the main task of epistemology is framed in similar terms, Kornblith maintains, the proper role of intuition can readily be discerned. "The examples that prompt our intuitions", he writes, "are merely *obvious* cases of the phenomenon under study"[252]. Their obviousness resides solely in the wide agreement that is to be found among the beliefs generated by those intuitions, and such agreement, in his view, has nothing to do with a priority, but is as a posteriori as the judgements of any 'rock collector' can be. Such intuitions therefore, according to Kornblith, would not only be *corrigible* but largely *theory-dependent*. This implies, in particular, that the intuitive judgements formulated at early stages of the investigation would carry less epistemic weight. Their reliability, in his view, will be most likely to increase as the privileged background theory progresses, thereby deepening our understanding of the phenomenon under investigation. This entitles him to subscribe to the following maxim: "The greater one's theoretical understanding, the less weight one may assign to untutored judgement"[253]. This is why, in particular, as the scope of our theories expands, "old intuitions give way to well-integrated theoretical judgements, and, in addition, to new intuitions about matters not yet fully captured in explicit theory"[254].

While sharing his reliabilist approach to epistemology, some philosophers have found Kornblith's account not entirely satisfying. Alvin Goldman, for instance, has recently proposed a naturalistic account of the philosophical use of intuitions which, while rejecting as wrongheaded

[250] *Ibid* (133).
[251] *Ibid* (134).
[252] *Ibid* (134). My emphasis.
[253] *Ibid* (135).
[254] *Ibid* (135).

Kornblith's appeal to natural kinds, favours an approach that I consider interesting in its own right and find therefore worth mentioning[255].

Goldman's general approach to the matter is apparently very simple: we cannot even start assessing the epistemic reliability of philosophical intuitions, he maintains, until we decide what the proper targets of philosophical analysis are. This is due to the fact that intuitions will most likely turn out to be a reliable or unreliable epistemic source depending on how we construe the proper target of philosophical inquiry. That is to say that, trivially, we cannot decide whether a given intuition constitutes a piece of evidence, if we haven't previously decided what that intuition is supposed to be evidence *for*. The targets of philosophical analysis, Goldman observes, have been alternatively construed by different philosophers as:

1. Platonic forms
2. Natural kinds
3. Concepts$_1$ (in the Fregean sense)
4. Concepts$_2$ (in the psychological, personal sense)
5. Concepts$_3$ (shared concepts$_2$)[256].

According to Goldman, the first stage of philosophical inquiry consists in the study of commonsense or folk concepts[257]. As a consequence, the way in which philosophical analysis is standardly practiced would reveal that it is concerned uniquely with concepts$_2$, i.e. with private mental representations, or concepts in the psychological, personal sense. Once the target of philosophical inquiry is thus properly construed, he maintains, intuitions become again an evidential source worth appealing to[258], and an epistemic agent can rightfully be said to *intuit* that a case *a* is or isn't an instance of concept *F* (or that a concept *F* applies or fails to apply to case *a*).

In order to clarify the nature of the evidential relation which links, in his view, a single intuition to the concept it is evidence for, Goldman distinguishes between what he calls *constitutive* and *non-constitutive* groundings of an evidential relation. An example of constitutive grounding would be provided by the idealist doctrine usually referred to as *phenomenalism.* According to this doctrine, as it is well known, the evidential status of appearances is grounded in the

[255] Goldman (2007).
[256] See Goldman (2007: 6).
[257] "We must first identify the features of folk epistemology", he writes, "in order to figure out how it might be transcended by scientific epistemology, while ensuring that the latter project is continuous with the former". *Ibid* (22).
[258] A similar view has been defended by Gopnik and Schwitzgebel (1998).

constitution of physical objects, insofar as "what it *is* to be a physical object of a certain sort is that suitably situated subjects will experience perceptual appearances of an appropriate kind"[259]. While rejecting phenomenalism, Goldman holds that the evidential status of the intuitions by means of which an epistemic agent applies a concept *F* to a case *a* is grounded in the concept itself. "It's part of the nature of concepts (in the personal psychological sense)", he writes, "that possessing a concept tends to give rise to beliefs and intuitions that accord with the contents of the concept"[260].

This does not imply, in his view, that application intuitions of this sort should be regarded as *infallible*. An epistemic agent might for instance be misinformed or insufficiently informed about a case *a*, or act on the basis of a false theory concerning concept *F*[261], and similar circumstances might well affect the correctness of his corresponding intuitions. This, in particular, would be the reason why philosophers usually prefer carefully stipulated examples and try to avoid appealing to intuitions that are clearly influenced by a theory of the target concept.

The two main advantages of Goldman's account, explicitly recognized by the author, are certainly worth signalling. A standard line of skepticism toward the philosophical use of intuitions, as Goldman points out, draws on the fact that different people might have conflicting intuitions about specific cases. His account blocks this line of criticism, insofar as, for obvious reasons, interpersonal variation ceases to undermine the reliability of intuitions. Moreover, once the subject matter of philosophical inquiry ceases to be a mind-independent entity, such as Platonic forms, natural kinds, or concepts in the Fregean sense, the causal links between intuitions and target concepts become intelligible again and worth investigating, thereby guaranteeing the possibility of placing philosophical analysis within a wholly naturalistic framework.

[259] *Ibid* (14).
[260] *Ibid* (15).
[261] Goldman finds important to distinguish here between "a theory presupposed by a concept and a theory *about* a concept, i.e., a general account of the concept's content", and to point out that only the latter is likely to lead one's intuitions astray. *Ibid* (15).

3. Epistemological considerations

3.1 *On what there* might *be*

One of the most intriguing features of our species, I believe, is that it can be credited with a remarkable ability to transcend the actual in thought by reflecting on, or else mentally entertaining, *counterfactual* state of affairs. On a first, rough, approximation, a counterfactual state of affairs can be profitably understood for our present purposes as a mental representation of a non-actual *way* in which the world has been, is, or will be. Besides being able to reflect on similar scenarios, moreover, we display a natural tendency toward formulating *judgements* concerning the possible consequences of similar scenarios.

At the moment, for instance, I am sitting in front of my computer, in Pesaro, writing about thought experiments and sipping from a cup of green tea, but I can easily picture myself strolling about the winding little alleys of Urbino with my friend Claudio, smoking a cigar, and reasoning about the several different ways in which we could have missed to be friends. If in an early afternoon of a windy New York autumn, for instance, I had not entered a symbolic logic class held in one of the austere buildings of Columbia university, I would have not had the pleasure of making his acquaintance. At the same time though, I might reason further, my not entering *that* classroom on *that* specific afternoon, certainly could not have prevented me from meeting Claudio on some other departmental or more mundane occasion.

Examples similar to the above, I believe, drove the American philosopher Roderick Chisholm to formulate the following epistemological consideration:

> "We seem to have knowledge of what *might* have happened, of what *would* happen if certain conditions were realized, of what tendencies, faculties, or potentialities and object *could* manifest in suitable environments. And this, most of us would be inclined to say, is valid and significant, even though the possible events to which it seems to pertain may never become actual"[262].

Our ability to produce similar representations, it seems, appears quite early in our developmental history, to the point that it has been used by cognitive scientists in order to investigate the way in

[262] Chisholm (1946: 289).

which young children acquire *causal* knowledge[263]. As a matter of fact, it seems difficult to deny that this highly creative mental skill plays a fundamental role in our *explanatory abilities* in general. Our ability to account for the occurring a given phaenomenon, it can indeed be argued, seems to be largely dependent on a corresponding capacity to produce and mentally manipulate relevant counterfactual scenarios involving physical, biological, or psychological variables.

"If that stupid tree had moved out of the way", I might exclaim after a car accident, "I would not have crashed my new Lamborghini!"; "if there had not been so much abuse of psychotropic substances during the seventies", some hard-headed philosopher of science might lament, "there would not be so many modal realists around today!"; "if you had made different encounters in your life", I might tell Claudio, "you would not now think of yourself as an ontologist!".

Indeed, if we envisioned scientific abstraction in general as instantiating similar patterns of counterfactual reasoning (as in asking, for instance, what *would* happen to a body if it *were* to move along a frictionless plane), I don't think it would be an exaggeration to maintain that the afore mentioned human ability fuels the engine of most human *discovery*.

As I hope the examples considered in the previous chapters have contributed to make clear, all thought-experimental thinking exploits this general ability to contemplate counterfactual scenarios and to make judgements concerning their most likely or unlikely consequences. As a matter of fact, I take the *raison d'être* of every thought experiment to lay in the fairly common pre-theoretical acknowledgement of the fact that reflecting on counterfactual scenarios has often proved capable of enhancing our understanding of reality in a way that would not have been possible had we confined our reflection to the actual world only. For this reason, I think that thought experiments are best seen as a sort of epistemic tools by means of which we investigate a realm of *possibilities* in order to assess *truths* concerning, or to test our *beliefs* about, what is *actual*.

In the present chapter I will focus on *philosophical* thought experiments. "Philosophers", it has been written, "characteristically ask not just whether things *are* some way but whether they *could have been* otherwise"[264]. As a matter of fact philosophical thought experiments present themselves as inherently *modal* ways of reasoning, which purport to give us access to some kind of modal truths. While in fact inspecting the world tells us what *is* the case, the more or less bizarre counterfactual scenarios evoked by thought experiments, as we have seen, purport to tell us something about what *might* or *must* be the case. It seems therefore natural to expect that a

[263] See Sobel (2004). According to some, children would *learn* to *detect* new causal relations by considering counterfactual alternatives. Along the same line of reasoning, one might argue, adults *discover* new causal relations as well as other truths about the world.

[264] Williamson (2007: 134). My emphases.

critical assessment of their powers and limits will be largely dependent on a previous assessment of the various different kinds of possibilities that thought experiments invite us to consider.

According to a widely agreed upon standard definition, the *modality* of a sentence *S* is the way in which its truth value holds. Starting at least from Aristotle[265], we regard a given proposition as necessary if it *must* be true, and as possible if it *might* be true. Whenever we *modalize*, that is, we make a judgement or entertain a thought concerning the *way* in which a certain proposition *holds* or a certain state of affairs *obtains*. I think it is important to realize that, whenever we thus try to assess the modal status of a proposition, we do so by more or less consciously appealing to other propositions with respect to which we deem that proposition to be necessary or possible. Possibility and necessity, in other words, are normally understood as essentially *relative* notions. We seem to have a pre-theoretical understanding of this fact, which appears to be deeply rooted in our familiarity with natural language.

This is not an isolated phenomenon. A similar one is observable in the way in which we use *quantifiers* in ordinary discourse. Every competent user of a natural language, that is, shows an intuitive understanding of the fact that quantifiers rarely occur unrestricted in human discourse and that their scopes normally depend on the context of utterance of the sentences within which they appear. When confronted with sentences which display an identical syntactical form, such as, for instance, 'there is no God' and 'there is no beer', we more or less consciously accommodate to the relevant scope of the quantifier at hand.

Something similar, as I am trying to argue, holds in the case of ordinary modal thought and talk. Suppose, for instance, that some crazy philosopher told us that it is *possible* that the brain of a man be transplanted overnight into the skull of another man, or that, for all we know, it is *possible* for any of us to have a *doppelgänger* in a world spatio-temporally inaccessible from our own, or else that, according to the way in which we use the verb 'know', it is *possible* for an epistemic agent to be justified in believing something true without thereby knowing it. When confronted with similar situations, I believe, our normal reaction would be a demand for clarification, which we would presumably express by uttering the words: "What do you mean by 'possible'?" Or similar words to the same effect.

A way to approach the same matter from a somehow different angle is the following. Suppose that we found ourselves at the philosophy department of the University of Urbino and we were asked to assess the truth or falsity of the two intentionally ambiguous following claims, concerning the usual hero of our examples:

(1) Claudio cannot have travelled from Empoli to the department in under 10 minutes

[265] Aristotle 1015b.

and

(2) Claudio cannot have travelled from Andromeda to the department in under 10 minutes[266].

Now, it seems reasonable to maintain that, while answering that sentence (1) is true we would normally *mean* that travelling from Empoli to the department in under 10 minutes is not possible *relative to* the present state of human (especially Italian) technology, by providing the same answer to sentence (2) we would rather *mean* that travelling from Andromeda to the department in under 10 minutes is not possible *relative to* the known laws of physics.

Some would take this to reflect the fact that there is no such thing as an *absolute* modality. Truths, it seems, cannot be meaningfully said to be necessary or possible *simpliciter*, but they are usually treated as being necessary or possible *relative to something*. "Claims about what is possible", it has indeed been argued, "bear an implicit relativization to a set of facts which are held constant"[267]. With respect to sentences (1) and (2), for instance, as we just saw, the two relevant sets could be plausibly taken to be the set of all facts concerning human technology, in the case of (1), and the set of all known physical laws, in the case of (2). Joseph Melia has tried to generalize the same insight by claiming that, whenever we formulate modal judgements, "we take a certain collection of truths as given – call these the φ truths – and then define the notion of φ-possibility as "compatible with the φ truths" "[268].

What now matters for our present purposes is the fact that, contrary to the state of affairs described in our previous examples, the characterization of what Melia calls the set of φ truths relevant to the counterfactual scenario evoked by a given thought experiment does not seem to be an immediately obvious task. It is in this sense that Sören Häggqvist has seen in their *modal* nature the "Achilles heel" of both scientific and philosophical thought experiments, that which makes them significantly different from ordinary or laboratory experiments[269].

In the first chapter, I tried to draw attention on and account for the mysterious puzzlement that every thought experiment seems bound to generate in its potential audience. A significant part of that puzzlement, I would now like to add, seems to be due to the fact that, in the case of thought experiments, the sense in which a given scenario is said to be *possible* is not always transparent. In fact, as we shall presently see, the way in which most thought experiments are typically

[266] This is an adaptation of two examples due to Kit Fine. See Fine (2002: 254).
[267] Kitcher (1984, 26).
[268] Melia (2003: 17).
[269] See Häggqvist (1996: 117).

presented usually leaves open the possibility of associating different sets of truths with the same thought experiment.

This last aspect is crucial for our purposes, insofar as the purported outcome of any single thought experiment, as I will try to show in the next section, will carry a very different epistemic weight according to the way in which we decide to interpret the modal notion featuring in it. The possibility of drawing significant conclusions from a given thought-experimental scenario, in other words, appears to be largely dependent on our previous determining the kind of possibility appealed to by the thought experiment. This means, in particular, that the *conclusiveness* of the thought experiment itself, namely its success in providing compelling reasons for rejecting the polemical target it is designed to attack, cannot be credited to its argumentative structure only, but relies rather on a host of crucial, yet usually not explicitly stated assumptions concerning what we hold or do not hold to be possible.

3.2 *Varieties of modality*

The plausibility of the last claim of the previous section, I believe, can be corroborated by means of an example. A good candidate for the job could be the case of *Twin Earth*, the fictional entity bred by Hilary Putnam's fervid imagination, which we considered in the first part of last chapter. As we already know from that occasion, Twin Earth is, or so we are asked to imagine, *exactly* like Earth but for *one* fundamental aspect: the liquid called 'water' on Twin Earth is not H_2O but has a different chemical formula, abbreviated by Putnam himself as XYZ. Once confronted with a similar scenario, I believe, it is perfectly natural to ask the thought experimenter: "In which *sense* are you taking the existence of this planet of yours to be *possible*?".

Suppose his answer pointed to the fact that his mental creation undeniably constitutes a perfectly intelligible *logical* possibility. As a matter of fact, if we standardly construe a logical possibility roughly as a coherently describable state of affairs, i.e. a state of affairs which does not involve any explicit or implicit contradiction, then Twin Earth certainly does look as a fairly plausible candidate to this kind of possibility. Indeed, the thought experimenter might draw our attention to the fact that there is nothing self-contradictory in imagining a universe in which different physical laws hold, to the effect that a substance functionally equivalent to the one we call 'water' in our universe might have a different chemical constitution. Nonetheless, we might object, insofar as Putnam's thought experiment aims at establishing a semantical thesis concerning *our* world, and not some other logically possible world, this sense of 'possible' does not seem to be the relevant one.

This kind of objection, in particular, has been meticulously articulated by Kathleen Wilkes[270], whose views are worth mentioning. With respect to the *logical* possibility that water had a different chemical constitution from the one it actually has, Wilkes observes:

> "Even if by reliance on the willing suspension of disbelief we are prepared to say that logical possibilities such as these *are* imaginable […] we also know that most or all of these things *could not happen*: are, in short, impossible. The point is this: what is fine in literary fantasy (where the ambition is to entertain) is not necessarily enough to 'establish a phenomenon' (from which the ambition is to draw conclusions)"[271].

According to Wilkes, then, we cannot base thought experiments on mere logical possibilities, insofar as the imaginability of these "fairy stories"[272], as she calls them, would not be "enough of a basis upon which to build conclusions about what we would say if such things *did* happen"[273]. According to the same author, it is indeed crucial to distinguish between simply framing, or entertaining, a *mental picture*, on the one hand, and what she, borrowing terminology from Brown[274], calls *establishing a phenomenon in thought*, on the other.

This last point is very important for our purposes. In order for us to be able to draw reliable conclusions from a thought experimental scenario, Wilkes maintains, that scenario must be capable of 'establishing a phenomenon' in thought. To accomplish this latter task, in particular, we must be able to specify the *background* conditions relevant to the thought experiment at hand. These conditions constitute what we might call the decisive *dark side* of every thought experiment, in that they grant us both that our scenario is adequately described and that, as a consequence, significant conclusions can be drawn from it.

Now, maintains Wilkes, a phenomenon cannot be said to have been established in thought unless it is *theoretically possible*, i.e. compatible with our best current scientific theories[275]. "The notion of imaginability that is needed for genuine thought experiments", she writes, "will presuppose attention to the relevant backing theories"[276]. In particular, Wilkes stresses the importance of the difference between what she calls imagining the '*that*' and imagining the '*how*' of a given thought-experimental scenario. "If the 'how' cannot be imagined", she very reasonably contends,

---

[270] See Wilkes (1988: Ch. 1).
[271] *Ibid* (18).
[272] *Ibid* (43).
[273] *Ibid* (21).
[274] See *Ibid* (8).
[275] See *Ibid* (18).
[276] *Ibid* (21).

"the 'that' thought-experimental conclusion becomes decidedly meagre"[277]. Indeed, the scenarios that most thought experiments appeal to, being too vaguely sketched, encourage us to disregard the fact that, in order for the situation they describe to occur in our world, many other concomitant changes would also have to occur, and we cannot be said to have established any phenomenon unless we have assessed the theoretical possibility of these latter changes as well[278]. Wilkes considerations seem therefore to rule out as epistemically barren the option of interpreting as *logical* possibility the kind of modality relevant to Putnam's thought experiment. The set of φ truths relevant to the potential conclusiveness of the Twin Earth thought experiment then, to couch her opinions in Melia's words, should not be taken to coincide with the set of all know *logical* truths.

A more promising strategy, it might seem at first, would rather be that of opting for the set of all known *physical* truths. We could try to maintain, that is, that Twin Earth is possible with respect to the presently known laws of physics. Unfortunately, this way of construing the relevant modal notion presents a serious shortcoming, insofar as it trades heuristic unfruitfulness for incoherence. As a matter of fact, as it has been pointed out[279], an appeal to the physical possibility of Twin Earth would have the very unwelcome effect of rendering the thought-experimental scenario itself self-inconsistent. Here is how Sören Häggqvist envisions the problem:

> "If Twin Earth contains no H_2O molecules, the [Twin Earthlings] don't contain any either. But we contain plenty. Hence the [Twin Earthlings] cannot be "molecular copies" of us. This objection […] is frequently mentioned in passing in the literature, but has never, as far as I know, been take seriously. I am not sure what to make of this lassitude, since *the objection seems quite cogent*"[280].

As we already know, Putnam's thought experiment purports to show that it is possible for two epistemic agents to be in the exact same mental or brain state and yet for the terms they use to differ in extension. It follows that a fundamental requirement of every Twin Earth example is that the two agents be *exactly* identical. In order for the argument to go through, that is, our hypothetical *doppelgänger* on Twin Earth has to be an *identical* copy of ourselves. Now the problem is that this possibility seems to be ruled out by the though experiment itself, in that the

[277] *Ibid* (34). She raises this problem with respect to the thought-experimental challenges put up by artificial intelligence to the very notion of *personhood*, but the point, I believe, applies equally well to the case we are considering.
[278] See *Ibid* (31).
[279] See Häggqvist (1996: 169; 2009b: 69) and Crane (1996 [1991]: 290).
[280] Häggqvist (1996: 169). My emphasis.

chemical constitution of our bodies comprises molecules of water, a substance which, *ex hypothesis*, is completely absent on Twin Earth. Hence our *doppelgänger* cannot, by stipulation, be an exact copy of ourselves, and this obviously invalidates the conclusiveness of Putnam's thought experiment.

The above considerations, I think, provide at least *prima facie* support for the view, advanced earlier, according to which a single philosophical thought experiment can neither be regarded as informative, nor as conclusive, unless one is already willing, for largely independent reasons, to accept specific assumptions concerning the 'proper' way of construing the modal notion it appeals to. Not all the readings of the relevant modal notion, as I tried to show, allow to draw from the thought-experimental scenario the conclusions it was designed to establish.

3.3 *The old riddle of counterfactuals*

A further characteristic feature of philosophical thought experiments, in particular, is especially relevant for our purposes. In the first chapter I observed that it may not be possible, nor even desirable, to single out a set of necessary and sufficient conditions that an argumentative strategy has to satisfy in order to be rightly considered, or to qualify as, a fully-fledged thought experiment. Nonetheless, even a quick look at their argumentative structure, naturally suggests the following observation. All philosophical thought experiments display a very peculiar syntactical feature, namely they make extensive and systematic use of a particular kind of subjunctive conditional sentences, usually referred to as *counterfactual* conditionals (from now on *counterfactuals*)[281]. A counterfactual is a sentence of the form

"if it *were* the case that *P*, then it *would* be the case that *Q*",

usually formalized as,

$P \mathrel{\square\!\!\rightarrow} Q$,

where *P* and *Q* are propositional variables ranging over state of affairs, and where, in particular, the state of affairs represented by *P* is either not actual or taken to be such by the utterer of the sentence. A counterfactual, in other words, postulates the obtaining of a certain hypothetical state

[281] For early introductions to the epistemic problems raised by counterfactuals, see in particular Chisholm (1946) and Goodman (1983 [1947]). For a recent survey of further developments on the topic, see Bennett (2003).

of affairs and then draws a conclusion concerning what *would* follow if that state of affairs *were* to actually occur. Assuming, for instance, that I were reasonably certain of the earthly origins of my friend Claudio, then a sentence of the form

'If Claudio came from Andromeda, we would certainly still be friends'

would be a counterfactual.

Notoriously, counterfactuals raise a perplexing *verification* problem, which has long puzzled epistemologists in that it seems to seriously undermine the possibility to *justify* our beliefs in (or claims to knowledge of) their consequents. This is due to the fact that, as Nelson Goodman famously pointed out in his seminal treatment of the topic[282], counterfactual conditionals are typically not used in natural language, nor in science for that matter, as truth-functional compounds. Another way to put this, is to observe that the logical behaviour of the connective '□→' is generally not regarded as truth-functional.

As a matter of fact, insofar as a material conditional with a *false* antecedent is always true, and insofar as the antecedent of every counterfactual is, by hypothesis, *always* false, it follows that both a counterfactual and its opposite, if considered as truth-functional compounds, would always come out true. This can easily be showed by recalling our previous example. If treated as a truth-functional compound, for instance, my original sentence

'If Claudio came from Andromeda, we would certainly still be friends'

would turn out to be logically equivalent to its opposite, namely

'If Claudio came from Andromeda, we would certainly *not* be friends anymore'.

Obviously, this would utterly fail to capture the intended *meaning* of the original counterfactual[283]. Indeed, by uttering the previous sentence I meant to convey my willingness to count Claudio amongst my friends, even if it should turn out that the several odd traits of his personality are to be blamed on the fact that he did not grow up on this planet!

[282] Goodman (1983 [1947]).

[283] In a sense, this can be seen as a special case of a difficulty which characterizes the analysis of conditionals in general. As it is well known, insofar as a material conditional with a false antecedent or with a true consequent is always true, its purely logical treatment often yields unexpected and unwanted truths. Both material conditionals "If 7 is even, then Claudio comes from Andromeda", and "if Claudio comes from Andromeda, then 7 is odd", for instance, are logically true.

Moreover, such a treatment would make counterfactuals in general very little *informative*, to say the least. On the contrary, as their ubiquitous use in both scientific and non scientific discourse attests, by uttering counterfactuals we certainly *do* intend to convey some rather specific piece of information. A counterfactual sentence, in particular, claims something about the world by holding that a certain kind of connection between its antecedent and its consequent *obtains*. The verification problem mentioned above, at this point, is precisely that of providing a reliable criterion for assessing *when* such a connection obtains, or, in other words, of defining the circumstances under which a given counterfactual holds[284].It should be clear from what I said so far that this last task cannot be seen as a purely *a priori*, nor as a purely *a posteriori* business.

It is not purely *a priori* because, as we have just seen, the connection between the antecedent and the consequent of a given counterfactual, if it obtains at all, does not obtain as a matter of pure logic. The principle on the basis of which we might hold ourselves justified in inferring the consequent of a counterfactual from its antecedent, in other words, cannot possibly be a law of logic, but must draw its epistemic value (if any) from some other level of normativity[285].

On the other hand, the problem we are confronted with is not purely *a posteriori* either. Indeed, insofar as a counterfactual conditional concerns what *would* have ensued if a certain state of affairs *had* occurred which in fact *did not* occur, we obviously cannot simply look at the way the world *is* in order to determine its truth value[286]. Determining its truth value, that is, cannot possibly be a matter of empirical investigation only.

At this point, some might feel, it would be fair to ask: How do the above considerations, pertaining to the semantics of a particular kind of conditional sentences, namely counterfactuals, bear on the general topic of our present work, namely philosophical thought experiments? I believe the answer to be the following. Insofar as counterfactuals, as we ha seen, seem to play a fundamental role in their internal structure, and insofar as, when first confronted with a well designed philosophical thought experiment, we are often willing to regard its premises as providing some kind of support for its conclusion, it seems reasonable to maintain that the potentially persuasive force of philosophical thought experiments should not be regarded as strictly logical in nature.

[284] See Goodman (1983 [1947]: 4).
[285] See *Ibid* (8).
[286] I take this consideration to be implicit in Chisholm characterization of counterfactuals, according to which: "A subjunctive conditional is one such that we can know that the antecedent in some sense implies the consequent without knowing the truth value of either". Chisholm (1946: 295).

3.4 *Regimenting thought experiments*

Taking a closer look at the typical inferential structure of most philosophical thought experiments will help us reach a better grasp of their modal nature, while at the same time taking us a step further in our investigation of their general epistemic status. In order to do this, I will take as a starting point the two attempts at regimentation of thought experiments I considered at the end of the first chapter, namely the schemas put forward by Roy Sorensen and Sören Häggqvist.

As a matter of fact, I think that these two authors have made some of the most valuable efforts in this direction to be found in the literature, or at least of the most profitable from the point o view of our present purposes. It can indeed hardly be denied, I think, that the general frameworks put forward by Sorensen and Häggqvist both originate from a previous acknowledgement of the unavoidably *modal* nature of all thought experimental thinking. Both proposals, moreover, place a strong emphasis on that very peculiar syntactical feature of thought experiments, namely counterfactuals, which, as we have seen, plays a crucial role in our understanding of their proper functioning[287].

According to both Sorensen and Häggqvist, a significant number of thought experiments share a common purpose, namely they can be seen as procedures aimed at causing *justified belief revision* by *rejecting* a given target statement. This means, in particular, that their function ought to be envisioned as a fundamentally *negative* one. To be relevant for us is the fact that, in order to perform this function, thought experiments appeal to *counterfactual scenarios* and present these scenarios as *possible*. It is important to realize that this last move is motivated by the existence of a very common *desideratum* that we normally tend to impose on our theories. In order, that is, to accept a given philosophical theory as a plausible explanation of the phenomenon or set of phenomena we are investigating, we want that theory to be, as it is often put, 'counterfactual supporting'. This means, roughly, that our theory should be able to yield *true* sentences of the following form: 'if it were the case that A, then it would be the case that C', or alternatively, 'if A were to occur, then C would also occur'. Another way to put this is to maintain that we normally envision our theories as more or less implicitly committed to certain modal consequences.

The initial purpose of Sorensen's schema, as I explained above, was that of charting the general logical structure under which, according to his opinion, most thought experiments could be subsumed. According to his Kuhnian 'cleansing model' of armchair inquiry, it might be useful to

[287] This latter feature, in particular, also characterizes the approach to philosophical thought experiments recently adopted by Timothy Williamson, whose views I won't discuss here. See Williamson (2007).

recall, thought experiments would be best thought of as a sort of *detectors*, aimed at revealing inconsistencies in our system of beliefs. For this reason, the schema he finds most appropriate to formalize their argumentative structure is that of a *paradox*, standardly construed as a set of individually plausible yet jointly inconsistent propositions.

Thought experiments, in his view, would aim at revising our beliefs by persuading us to reject a purported modal consequence of a given target statement. In the case of what Sorensen calls *necessity refuters*, let me recall, the relevant schema would be the following[288]:

6. S
7. $S \supset \Box I$
8. $(I \wedge C) \Box\!\!\rightarrow W$
9. $\neg \Diamond W$
10. $\Diamond C$

The same idea of regimenting thought experiments by envisioning their fundamental structure as a group of sentences forming an inconsistent set, i.e. a paradox, as we have seen above, has been insightfully elaborated further by Sören Häggqvist, who, as we saw, has recently proposed the following alternative schema[289], which he has dubbed schema (α):

$\Diamond C$
$T \supset (C \Box\!\!\rightarrow W)$
$C \Box\!\!\rightarrow \neg W$
$\therefore \neg T$

Despite their relative simplicity, the two schemas recalled above might run the risk of appearing rather abstract to the reader. For this reason, I believe it might be useful for our present purposes to see what these schemas look like once applied to single thought experiments. In what follows, I will therefore consider two specific applications first suggested by Sorensen and Häggqvist themselves. Also, in order to take a first step toward testing their generality, I will subsequently move on, in the remainder of this section, to a further application of my own, the legitimacy of which, I believe, would probably be granted by the same authors.

[288] For details as to how the schema has to be interpreted see section 1.4.4 above.
[289] See the end of section 1.4.5 above for details concerning interpretation.

By subsuming under his general schema Gettier's thought experiment, which we considered in the previous chapter, Sorensen ends up with the following argument[290] (for the sake of clarity, each sentence will be followed by the formalization provided by the author, enclosed in square brackets):

1. Knowledge is justified true belief. [S].
2. *If* knowledge were justified true belief, *then*, *necessarily*, all justified true belief that *p* would be knowledge that *p*. [S $\supset$ $\Box$ I].
3. *If* all justified true belief that *p* were knowledge that *p*, *and* Smith had justified true belief that *p* because of luck, *then* Smith would have knowledge that *p* because of luck. [(I $\wedge$ C) $\Box\rightarrow$ W].
4. It is *not possible* to have knowledge because of luck. [$\neg$ $\Diamond$ W].
5. It is *possible* that Smith has justified true belief because of luck. [$\Diamond$ C].

In a similar fashion, Häggqvist has put his schema to work by applying it to another thought experiment we are already familiar with from last chapter. It is the brilliant science-fictional scenario by means of which Hilary Putnam proposed the view on meaning now commonly known as *semantic externalism*. As I noticed above, Putnam's thought experiment gave birth to *Twin Earth*, one of the most heuristically fruitful yet theoretically controversial fictional entities of recent analytic philosophy. Once subsumed under the argument schema he proposes, according to Häggqvist, the Twin Earth thought experiment takes the following form[291]:

- It is *possible* that Twin Earthians be "neuron for neuron" copies of Earthlings. [$\Diamond$ C].
- *If* psychological or brain states totally determine the extension of terms, *then* a Twin Earthian "neuron for neuron" copy of an Earthling would refer to water by the term 'water'. [T$\supset$ (C $\Box\rightarrow$ W)].
- *If* there were a Twin Earthian "neuron for neuron" copy of an Earthling, he would *not* refer to water by the term 'water'. [C $\Box\rightarrow$ $\neg$W].
- Therefore, psychological or brain states do *not* completely determine the extension of terms. [$\therefore$ $\neg$T].

[290] See Sorensen (1992: 137). I have slightly modified Sorensen's own example.
[291] See Häggqvist (2009: 68). Again, I have slightly modified the original version in order to make it terminologically consistent with the exposition of Putnam's thought experiment that I gave in the previous chapter. Even in this case, moreover, I have added after each sentence its corresponding formalization, enclosed in square brackets.

As I anticipated above, a first step toward testing both the plausibility and the generality of the schemas proposed by Sorensen and Häggqvist might be that of trying to see how they fare once applied to one of the philosophical thought experiments we considered in the first chapter. In order to do this, I will show what Frank Jackson's *knowledge argument*, featuring the miserable prisoner scientist Mary, might look like once subsumed respectively under the schema proposed by the former and under the one proposed by the latter author.

It will be useful to recall that Jackson's thought experiment was originally aimed at rejecting the metaphysical thesis commonly known as *physicalism* by providing evidence for the non-reducibility of some introspectively accessible features of our mental lives, usually referred to as *qualia*. Insofar as physicalism is commonly taken to claim that everything that exists must either be physical or supervene on the physical, acknowledging the existence of entities such as qualia, according to many, would constitute a powerful counterexample to it. Of course, I am not suggesting that the regimentation I am proposing should be seen as the only possible one. On the contrary, it seems plausible to maintain that there might be several alternative ways of making a single thought experiment fit any of the two schemas I am considering. This last assumption, of course, is built into the very nature of these schemes. Here is a version I believe Sorensen would subscribe to:

1. Physicalism is true. [S].
2. *If* physicalism were true, *then*, *necessarily*, knowing everything physical there is to know, would amount to knowing everything there is to know. [$S \supset \Box I$].
3. *If* knowing everything physical there is to know amounted to knowing everything there is to know, *and* Mary managed to learn everything physical there is to know from inside a black and white room, *then*, once released form her room, Mary would not learn anything new. [$(I \wedge C) \Box\!\!\rightarrow W$].
4. It is *not possible* that Mary, upon release from her room, does not learn anything new. [$\neg \Diamond W$].
5. It is *possible* that Mary manages to learn everything physical there is to know from inside a black and white room . [$\Diamond C$].

And here, in turn, is a plausible way in which, if I understand him correctly, Häggqvist would envision the same thought experiment:

- It is *possible* that Mary manages to learn everything physical there is to know from inside a black and white room. [$\Diamond C$].

- *If* physicalism is true, *then*, *if* Mary managed to learn everything physical there is to know from inside a black and white room, *then*, once released form her room, Mary would not learn anything new. [$T \supset (C \; \Box\!\!\rightarrow W)$].
- *If* Mary managed to learn everything physical there is to know from inside a black and white room, *then*, once released form her room, Mary would learn something new. [$C \; \Box\!\!\rightarrow \neg W$]
- *Therefore*, physicalism is false. [$\therefore \neg T$].

Both regimentations, I believe, provide very plausible reconstructions of the "essence", so to speak, of Jackson's reasoning in his much celebrated thought experiment, insofar as they seem to capture well both its underlying logical structure and its remarkable persuasive force. Although the generality of any regimentation attempt can never be established conclusively, I think that their apparent plausibility can be reasonably taken as *prima facie* evidence for the purported generality of their corresponding argument schemas, i.e. for the expected consistent application of these schemas to other thought experiments.

With respect to the unavoidably modal nature of all thought experiments, in particular, to which I drew attention in the previous sections, I think it is important to observe that, by their own nature, the general schemas considered above, once applied to single thought experiments, do not tell us what is the *kind* of modality relevant to the counterfactual scenarios at issue. The two schemas, in other words, are completely neutral with respect to the class of modal truths appealed to by the thought experiment to which they are applied. While, on the one hand, this constitutes a further confirmation of their generality, it seems to provide, on the other, further support to the hypothesis according to which the epistemic weight of a given thought experiment is not dependent solely on its logical structure.

3.5 *The model at work*

Attempts at subsuming other thought experiments under the two schemas, moreover, as I would like to argue, suggest that Häggqvist's regimentation proposal ought to be preferred over Sorensen one for at least three reasons.

For one thing, I think that Häggqvist's schema has the advantage of making more visible the similarity between *thought* experiments and *ordinary* experiments advocated by Sorensen's Machian stance. As a matter of fact, as we saw in the first chapter, Häggqvist explicitly presents

the schema for thought experiments simply as a modalized version of the schema which, according to his opinion, characterizes ordinary, laboratory experiments.

A second reason for preferring his regimentation, I believe, is that it seems to do a better job at representing the dialectical progression which actually takes place in the performing of most thought experiments, thereby rendering the regimentation itself less artificial and more fit to be applied to actual cases. Indeed, contrary to what Sorensen's schema suggests, a thought experiment usually starts by postulating the possibility of a counterfactual state of affairs, rather than drawing modal implications from a given theory. The latter move is indeed usually made only once the audience has already granted the possibility of the scenario.

The third reason is to be found in what I take to be Häggqvist's most valuable contribution to the study of thought experiments inferential structure. I have already pointed out that his regimentation develops further Sorensen's fundamental intuition, according to which the effectiveness of a successful thought experiment would originate from its power to generate a paradox. Starting from this assumption, as we have seen, Häggqvist has been able to provide a very profitable regimentation of the possible kinds of argumentative strategies available to the potential critic of the thought experiment. The three further schemas that he associates to the one intended to reflect the thought experimenter's intentions, as a matter of fact, prove themselves very useful in the difficult task of accounting for the often very intricate debate sparked off by the purported outcome of the thought experiment at hand.

A good way to appreciate the usefulness of Häggqvist's contribution, I think, could be that of applying schemas (β), (γ), and (δ), that I introduced in the first chapter[292], to the same thought experiment against which, in the last section, we tested the plausibility of schema (α), namely Jackson's knowledge argument. The three schemas, let me recall, are the following:

(β) T	(γ) T	(δ) T
$\Diamond C$	$\Diamond C$	$T \supset (C \Box\!\!\rightarrow W)$
$T \supset (C \Box\!\!\rightarrow W)$	$C \Box\!\!\rightarrow \neg W$	$C \Box\!\!\rightarrow \neg W$
$\therefore \neg(C \Box\!\!\rightarrow \neg W)$	$\therefore \neg (T \supset (C \Box\!\!\rightarrow W))$	$\therefore \neg \Diamond C$

Granting then that schema (α) itself, as we saw in the last section, appropriately reflects Jackson's original intentions, schema (β), i.e. the so-called *biting the bullet* strategy, would consist in defending physicalism by refusing to assent to the conditional $C \Box\!\!\rightarrow \neg W$. This would amount to denying that Mary, once released from her room, would actually *learn* anything new.

[292] See section 1.4.5 above.

It is interesting to notice that this is precisely the stance that Daniel Dennett has taken at this regard[293]. Jackson's argument, in his view, would simply be a bad thought experiment, "an intuition pump", according to his famous expression, "that actually encourages us to misunderstand his premises"[294]. Its purportedly "obvious" outcome, Dennett laments, is merely a consequence of the fact that, when imagining the counterfactual scenario depicted by the thought experiment, we usually, and understandably, fail to imagine *exactly* what the thought experimenter is asking us to imagine, namely that Mary has "*all* the physical information". This failure, in his view, is due to the fact that the thought experimenter directions, being "preposterously immense"[295], lay far beyond the reaches of our imagination. Nonetheless, so goes the gist of Dennett's objection, if those directions were followed correctly, it would not at all be absurd to maintain that Mary, upon release from her room, would not in fact learn anything new.

The application of schema (γ), in turn, which Häggqvist dubs the *irrelevance* strategy, would consist in denying the nested conditional $T \supset (C \; \Box\!\!\rightarrow W)$. With respect to our case, this move would amount to the claim that physicalism, once properly construed, is not committed to predict that, if the counterfactual situation envisioned by the thought experiment were to occur, Mary would not learn anything new.

As a matter of fact, as soon as 1984[296] Terence Horgan claimed that Jackson's attack on physicalism misses his target, in that it rests upon "a subtle equivocation"[297] between two different senses of the phrase 'physical information'. In order to dispel the confusion to which the knowledge argument falls pray, Horgan distinguished two relevant senses of the phrase 'physical information' such that it would not be legitimate to infer, as Jackson does, from its possession in the first sense to its possession in the second sense. His distinction reads as follows:

> "Let *S* be a sentence that expresses information about processes of a certain specific kind, such as human perceptual processes. We shall say that *S* expresses *explicitly physical information* just in case *S* belongs to, or follows from, a theoretically adequate physical account of those processes. And we shall say that *S* expresses *ontologically physical information* just in case (i) all the entities referred to or quantified over in *S* are physical

[293] See Dennett (1991: 398-406).
[294] *Ibid* (398).
[295] *Ibid* (399).
[296] See Horgan (2004 [1984]).
[297] *Ibid* (306).

entities, and (ii) all the properties and relations expressed by the predicates in *S* are physical properties and relations"[298].

While *explicitly* physical information, according to Horgan, has to be expressed in physicalistic language, *ontologically* physical information, in order to be such, does not need to satisfy this requirement, and can therefore be expressed in other languages. In other words, there may be sentences which express *ontologically* physical information while not expressing at the same time *explicitly* physical information.

Now, maintains Horgan, insofar as it does not mean to rule out the information-conveying power of languages other than the one of physical theories, *physicalism* is obviously a thesis concerning *ontologically* physical information. It follows, in his view, that it is plausible to hold that Mary, at the moment of her first color experience, *does* obtain new knowledge. Nonetheless, she is not likely to express her new knowledge by means of a similarity judgement such as 'Seeing ripe tomatoes is like seeing bright sunsets', because she has probably already learned that from her studies. Rather, she will express it by using an indexical term, as in 'Seeing ripe tomatoes has *this* property'[299], where 'this' designates a phenomenal feature of the perceptual experience she's undergoing. This last sentence, in particular, may very well express *ontologically* physical information, insofar as the referent of the phrase 'this property' could be a physical property which Mary is now experiencing from a first-person perspective. The above argument then, according to Horgan, would show that physicalism, once properly construed, is compatible with Mary's learning something new upon release. "The information is new", he writes, "not because the quale she experiences is a non-physical property, but because she is now acquainted with this property from the experiential perspective"[300].

The application of schema (δ), finally, which Häggqvist dubs *impossibility* strategy, would consist in defending the target thesis by rejecting the plausibility of the counterfactual scenario itself, i.e by denying assent to the modal claim according to which the situation envision by the thought experiment would be possible.

To my knowledge, no one explicitly took this stance with respect to Jackson's knowledge argument. Nonetheless, I believe that a similar line of reasoning can be said to be implicit in Kathleen Wilkes reaction to Putnam's Twin Earth thought experiment, considered above[301]. Indeed, as we saw, the question as to whether philosophical thought experiments can actually contribute to our understanding of reality, according to Wilkes, depends crucially on their ability

[298] *Ibid* (304).
[299] See *Ibid* (305).
[300] *Ibid* (306).
[301] See section 3.2 above.

to 'establish a phenomenon in thought', to use her own expression. Following her lead, then, it would certainly be possible to question Jackson's counterfactual scenario and to maintain that, insofar as the background of such scenario is not adequately described, the purported outcome of his thought experiment is bound to be inconclusive.

3.6 *Defeasible reasoning*

The considerations of the previous section, I believe, suggest that Häggqvist's schema (α) can be regarded as a profitable regimentation attempt, in that it adequately reflects the typical inferential structure of most philosophical thought experiments. If one is willing to grant this, it seems quite natural to ask what kind of inference (α) instantiates.

For the present purposes, we can provisionally think of an *inference* as the mental act of deriving a sentence, or a proposition, from another sentence, or proposition, according to a rule. With respect to the inference instantiated by (α), as we already argued in section 3.3 above, the rule which allows us to derive the conclusion from its premises is not a deductive one. The argument instantiated by our schema, in other words, is not deductively valid. It might be useful to recall here that an inference is standardly said to be deductively *valid* just in case it is not possible for its premises to be all *true*, without its conclusion also being *true*. Now, insofar as the kind the reasoning underlying a philosophical thought experiment, as our examples show, does not meet this requirement, and insofar as its premises seem nonetheless to provide compelling reasons for accepting its conclusion, it seems plausible to assume that (α) formalizes a *non deductive inference*.

As a matter of fact, a striking and long acknowledged feature of human reasoning in general is that most of the inferences our beliefs stem from are clearly non deductive ones, meaning that the truth of their premises are usually taken to provide some sort of support for their conclusions, i.e. to have some sort of confirmation-theoretic import, without at the same time *guaranteeing* their truth.

One obvious case is that of *induction*. If, based on my having never seen an honest politician, I conclude that *all* politicians are either dishonest men or become such after going into politics, my conclusion is justified to a certain extent. Nonetheless, it could certainly be false, without any of its premises being false as well.

An equally, perhaps even more ubiquitous case is that of *abduction*, or, as it is often called, *inference to the best explanation*, which Gilbert Harman referred to as the most basic form of

non deductive inference[302]. We can think, for instance, of the major role it plays in attributing mental states to other people. Suppose I were inclined to think that his Tuscan origins would *best explain* the fact that Claudio is unable to correctly pronounce the letter 'c'. My conclusion, which is probably correct, could be justified (to some small extent), but it could certainly not be said to follow *deductively* from my premises. Indeed, Claudio's *belief* that girls like Tuscan accent, for instance, and his corresponding decision to fake it, could *explain* his (apparent) inability to correctly pronounce the letter 'c' equally well. Understanding natural language, in general, can also be thought of as a matter of inferring, on the basis of some background knowledge, to (what we take to be) the best explanation of why some speaker uttered what she did. Similar considerations, it seems, hold for the case of a medical diagnosis, where a certain disease might be held by a physician, again on the basis of some more or less tacit background knowledge, to be the best explanation for the symptoms of her patient.

Non deductive inferential patterns are the subject matter of a relatively new and burgeoning area of research in both epistemology and logic, which focuses on the study of what has come to be referred to generally as *defeasible reasoning*. According to a widely agreed upon characterization, a piece of reasoning is usually said to be *defeasible* when the corresponding argument is rationally compelling but not deductively valid[303]. An argument of this sort, in turn, is itself regarded as *defeasible*. A *good* defeasible argument, according to the same characterization, will be one in which the premises are taken to provide evidential support for the conclusion, even if it is possible for such premises to be true and for the conclusion to be false.

Broadly speaking, the study of what has come to be referred to under the label of *defeasible reasoning* can be traced back to Aristotle's *Topics* and the *Posterior Analytics*, where the Greek philosopher set out to inquire into various forms of *dialectical* reasoning. Its most recent rediscovery, in the early sixties, is usually associated with the names of Roderick Chisholm[304] and John Pollock[305], whose epistemological work was explicitly intended to develop further the insights of their illustrious ancient predecessor. Subsequently, by the early eighties, researchers in artificial intelligence, grappling with the task of implementing reasoning in AI systems, made large use of the notion and developed it further, devising highly sophisticated formal frameworks[306].

---

[302] See Harman (1965).
[303] See, for instance, Koons (2009).
[304] See Chisholm (1957, 1966). Roderick Chisholm is indeed credited by John Pollock for being the first epistemologist to use, in 1957, the term "defeasible". According to Pollock, Chisholm borrowed the term from the philosopher of law Herbert L. A. Hart. See Pollock (1987: 482).
[305] See Pollock (1970, 1974, 1987, 1995).
[306] See Reiter (1980), McCarthy (1980) and McDermott and Doyle (1980).

John Pollock has recently lamented the fact that, according to a long and well established tradition in Western philosophy, a *good* piece of reasoning, in order to be such, is required to be deductively valid[307]. This status quo, according to his opinion, is to be blamed for the fact that the exploration of other forms of non deductive reasoning, such as inference to the best explanation or analogical reasoning for instance, on which we heavily rely in our everyday cognitive lives, has remained very rudimentary until quite recent times. What, starting from the sixties, brought to the gradual dismissal of the previously dominant tradition, maintains the same author, was "the recognition that many familiar kinds of reasoning are not deductively valid, but clearly confer *justification* on their conclusions"[308]. Various sources of knowledge, in particular, such as perception, memory, or testimony are often taken as *prima facie* reasons for accepting the purported truth of some hypothesis. As a basic example of this kind of reasoning, he mentions the case of *perception*:

> "Most of our knowledge of the world derives from some sort of perception. But clearly, perception is fallible. For instance, I may believe that the wall is grey on the basis of its looking grey to me. But it may actually be white, and it only looks grey because it is dimly illuminated. In this example, my evidence (the wall's looking grey) makes it reasonable for me to conclude that the wall is grey, but further evidence could force me to retract that conclusion. Such a conclusion is said to be justified *defeasibly*, and the considerations that would make it unjustified are *defeaters*"[309].

Another interesting example is to be found in the common phenomenon he calls *temporal projection*:

> "Suppose you are standing in a courtyard between two clock towers, and I ask you whether the clocks agree. You look at one, noting that it reads "2:45", and than you turn to the other and note that it reads "2:45", so you report that they do. But note that you are making an *assumption*. You could not look at the two clocks at the same instant, so you are assuming that the time reported by the first clock did not change dramatically in the short interval it took you to turn and look at the second clock. Of course, there is no *logical* guarantee that this is so. Things change. […] Thus a *defeasible* presumption of stability must be a primitive part of our reasoning about the world"[310].

[307] See Pollock (1987: 481).
[308] *Ibid*.
[309] *Ibid* (482).
[310] *Ibid* (483). My emphases.

The two examples above introduce us to a fundamental feature of defeasible reasoning, which, as we shall see, is very important for our present purposes. Contrary to deductive inferences, the reason schemes instantiated by defeasible inferences can have what Pollock calls *defeaters*, i.e. pieces of information that can mandate the retraction of their conclusion[311].

While epistemological approaches to defeasible reasoning are mainly concerned with the transmission of *warrant* from the premises to the conclusion of a given defeasible argument, the logical approaches focus primarily on the relationship between propositions. Approaching the matter from this angle allows us to introduce a further feature of defeasible reasoning, although it would probably be more appropriate to speak of a different side of the same feature. As a matter of fact, both *deductive* and *defeasible* logic are aimed at studying a certain *consequence relation* between propositions. Now, while in the former case the consequence relation is said to be *monotonic*, in the latter case it is usually called *non-monotonic*.

An inference rule is said to be non-monotonic if it allows us to infer certain conclusions from a subset of a set *S* of premises which cannot be inferred from set *S* as a whole. Another way to put this is to maintain that, and this is the crucial point, the relation of support between premises and conclusion can be defeated by additional information. This phenomenon is particularly visible in the case of induction. Indeed, the conclusion of an inductive inference, as it is well known, can be invalidated by adding further information to the base of our induction.

In the light of what I have said so far, I would like to consider the following proposal. As others have already pointed out[312], scientific thought experiments might be plausibly looked at as very peculiar and potentially fruitful forms of defeasible, i.e. non-monotonic reasoning. In what follows, I would like to suggest that the same point applies as well, mutatis mutandis, to *philosophical* thought experiments. Acknowledging the defeasible nature of the inference underlying a philosophical thought experiment, as a matter of fact, seems to me a plausible and very natural way of extending the approach first put forward by Sorensen and Häggqvist, the usefulness of which I have already tried to defend in the previous sections.

Nonetheless, in order to show the plausibility of this suggestion, allow me to go back to one of our previous favourite examples, namely the case of one of the most famous (fictional) scientist ever, Mary. In the previous section I claimed that Daniel Dennett's early reaction to Jackson's

[311] "What distinguishes defeasible arguments from deductive arguments", writes Pollock, "is that *the addition* of information can mandate the retraction of the conclusion of a defeasible argument without mandating the retraction for any of the earlier conclusions from which the retracted conclusion was inferred. By contrast, you cannot retract the conclusion of a deductive argument without also retracting some of the premises from which it was inferred". See *Ibid* (485).

[312] See Fano (2007).

knowledge argument could be seen as an instance of Häggqvist's schema (β). This schema, let me recall, has been dubbed by its creator the *biting the bullet* strategy for resisting the conclusion of a thought experiment, insofar as it consists in the stubborn rejection of its central modal intuition. As applied to our case, therefore, Dennett's move, as we saw above, was that of denying that Mary, once released from her room, would actually *learn* anything new.

To be interesting from the present point of view is the way in which Dennett intends to overturn the purported 'obviousness' of the thought-experimental conclusion. He does this by means of a personal narrative contribution to Jackson's original thought experiment, i.e. by adding further information to the counterfactual situation it invites us to reflect upon. Jackson's original story, Dennett suggests, could be "legitimately" continued as follows:

> "And so, one day, Mary's captors decided it was time for her to see colors. As a trick, they prepared a bright blue banana to present at her first color experience ever. Mary took one look at it and said "Hey! You tried to trick me! Bananas are yellow, but this one is blue!" Her captors were dumfounded. How did she do it? "Simple", she replied. "You have to remember that I know *everything* – absolutely everything – that could ever be known about the physical causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object or a blue object (or a green object, etc.) would make on my nervous system. So I already knew exactly what *thoughts* I would have (because, after all, the "mere disposition" to think about this or that is not one of your qualia, is it?)"[313].

Adding new information then, and enriching thereby the counterfactual state of affairs contemplated by Jackson's thought experiment[314], as the above passage shows, can have the power to undercut the persuasive force of its conclusion, thereby mandating its retraction. As a matter of fact, in the wake of Dennett's astute narrative intervention, the original outcome of the thought experiment has not only ceased to be experienced as rationally *compelling*, it has also handed this crucial psychological feature over to an opposite claim, to the effect that *now* it seems just 'obvious' that Mary, upon release, would *not* actually learn anything new.

---

[313] Dennett (1991: 399-400). Elsewhere, Dennett explicitly acknowledges the importance of philosophical thought experiments by claiming that "the reflection on the history of philosophy shows that the great intuition pumps have been the major movers all along". See Dennett (1984: 17). Nonetheless, it must be pointed out that Dennett's use of the term 'thought experiment' (or 'intuition pump' as he has dubbed them), while including the examples I considered in the previous chapter, encompasses also arguments that would not fall under the characterization adopted in the present work.

[314] What Dennett calls "turning the knobs" of our "intuition pumps". *Ibid* (74).

Nothing seems to prevent us, at this point, from suspecting that adding yet further information to Dennett's new story might end up reinforcing Jackson's original conclusion, or a different one for that matter, and that this process could be potentially carried on indefinitely. Appropriate further additions of information to the original counterfactual scenario, in other words, seem to constitute, for the case of a given philosophical thought experiment, what John Pollock calls *defeaters* of a given argument. Indeed, it seems safe to forecast that Dennett's kind of reaction to Jackson's knowledge argument is very likely to prove effective on other thought experiments as well, and this fact, I believe, lands further support to my earlier proposal, according to which philosophical thought experiments are best seen as not yet fully explored instances of defeasible, non-monotonic reasoning.

3.7 *Vindicating philosophical thought experiments*

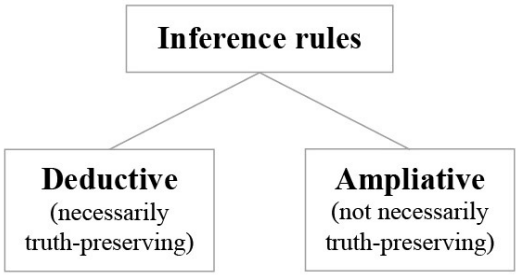
If a general moral were to be drawn from the previous section, as well as from the whole of this work, I believe that it would be the following. If, by whatever means, one comes to acknowledge the undeniably massive role that many known forms of non-monotonic reasoning play in both the more practical and the more theoretical intellectual endeavours of our species, as well as the many unexpected ways in which they have valuably contributed and still contribute to the growth of human knowledge, then Pollock's observation, according to which "defeasible reasoning is the *norm*, and deductive reasoning is the *exception*"[315], is likely to appear to his eyes as little more than an obvious fact, which is barely worth mentioning. And yet, it would hardly be an exaggeration to maintain that defeasible reasoning has often been carelessly overlooked by a large part of our philosophical tradition. As a matter of historical fact, philosophers interested in the study of human reasoning, over the centuries, have focused their attention mainly on far more logically well-behaved forms of inference than the ones we find at work in our actual reasoning.

While reflecting on the inferential structure of philosophical thought experiments, my attention was drawn to the work of the Italian logician and philosopher of mathematics Carlo Cellucci. In recent years, Cellucci developed a philosophical view concerning the nature of mathematical method and the growth of mathematical knowledge, which he has called *the open world view* and which might be interesting in its own right[316], but the details of which do not need concern us here. To be relevant from our point of view is rather the fact that, in order to advocate his

[315] Pollock (1987: 518). My emphases.
[316] See, in particular, Cellucci (2000).

views further, Cellucci has recently devised a *classification* of inference rules which he has presented as an alternative to what he holds to be the currently dominant one. This classification, as I would like to argue, contains an insight that I find to be highly relevant to our present purposes[317].

According to what Cellucci calls the *standard classification* of inference rules[318], all inferences can be conveniently divided into the two subclasses of *deductive* inferences and *ampliative* ones. While *deductive* inferences, on the one hand, are of course expected to be necessarily truth-preserving, i.e. to transfer (without exceptions) the truth of their premises to the truth of their conclusions, *ampliative* inferences, on the other hand, which are potentially able to contribute to our knowledge by generating new information which is not already contained in their premises, are considered as *not* necessarily truth-preserving. The situation then, according to the standard classification, is the one presented in the following tree diagram[319]:



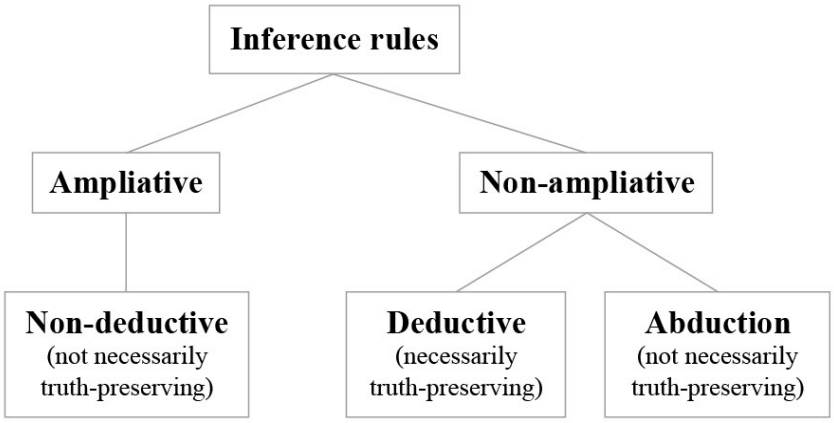
Cellucci finds the general view which underlies the standard classification patently inadequate. In fact, according to the above classification, *every* inference rule is bound to be either *deductive* or *ampliative*. On the contrary, he maintains, we normally make large use of defeasible inferential patterns that present themselves as *neither* deductive *nor* ampliative. One such inferential pattern, in his view, would constitute "one of the most important means of obtaining hypotheses"[320], namely *abduction*.

In order to defend this position, Cellucci argues as follows[321]. Once formalized, abduction responds to the following schema: $B \rightarrow A$, A, $\therefore B$. Insofar as this schema, as it is well known, is not deductively valid, any piece of reasoning which instantiates it will not necessarily be truth preserving, i.e. it will not always transfer truth from its premises to its conclusions. According to

[317] See Cellucci (2011).
[318] Which he takes to be well represented in Hintikka and Sandu (2007).
[319] The diagram is a slight modification of the one which appears in Cellucci (2011: 124).
[320] *Ibid* (124).
[321] See *Ibid*.

the standard classification then, we should expect abduction to be ampliative, but this, Cellucci argues, does not seem to be the case either, in that its conclusion (B) would be already contained in one of its premises ($B \rightarrow A$) and hence, strictly speaking, our inference would *not* contain in itself any amount of new information. New information, maintains indeed Cellucci, "is *not* generated by [abduction], but rather by the process that yields its major premise $B \rightarrow A$, thus it is generated *before* [abduction]"[322].

Insofar then as he takes it to be neither deductive nor ampliative, abductive reasoning constitutes, in his view, a clear counterexample to the standard classification, which stands therefore in serious need of radical revision. In particular, while according to the standard classification, as we have seen above, the main distinction to be drawn amongst inference rules would be the one between *truth-preserving* and *not truth-preserving* ones, "the only adequate classification of inference rules", according to Cellucci, ought to be given in terms of *ampliativity*[323]. Accordingly, he proposes to start by dividing inference rules in *ampliative* and *non-ampliative* ones. These latter are then further divided into *deductive* ones, which are truth preserving, and *abduction*, which is not. The following, therefore, is the tree diagram of the new classification we end up with by following his instructions:



I mentioned earlier that the new classification devised by Cellucci contained an insight that I found highly relevant for our present purposes, and I think I am now in the position to show wherein it lies. In order to do this, I need to state beforehand that I do not intend to discuss here whether *abduction* actually has or indeed fails to have the particular features that Cellucci ascribes to it. The interest of his classification for our present purposes comes rather from the

[322] *Ibid* (125). My emphases. Cellucci attributes the view according to which abduction would be non-ampliative to Peirce, according to whom "quite new conceptions cannot be obtained from abduction". See Cellucci (2011: 126).
[323] See *Ibid* (126). "That non-deductive rules are not necessarily truth preserving", he writes, "is a *consequence* of their being ampliative. On the other hand, deductive rules are truth preserving *because* they sacrifice ampliativity". *Ibid*. My emphases.

fact that it makes room for *other* kinds of non-ampliative and not necessarily truth-preserving inference rules *besides* the strictly deductive ones. When first confronted with Cellucci's proposal, it suddenly occurred to me that, regardless of the tenability of his views on abduction, this last insight deserved to be developed further. Indeed, while trying to come to grips with his classification, my attention was powerfully caught by the observation which he decided to end its presentation with. "It remains an open question", he writes, "whether, *in addition* to [abduction], there are *other* interesting inference rules that are both non-ampliative and not necessarily truth preserving"[324].

Now, as I hope the considerations I developed in the present work have contributed to make clear, I believe that the kind of reasoning which takes place in most philosophical thought experiments is best seen as instantiating inference rules of this latter sort. While in fact *non-ampliativity*, on the one hand, seems implicit in what I, endorsing Sorensen's *cleansing model* of armchair inquiry, have indicated above as the primary, ideal function of a philosophical thought experiment, namely that of making us more rational by detecting inconsistencies hidden in our system of beliefs, Häggqvist schema (α), on the other hand, which I take to reflect the typical inferential structure of most philosophical thought experiments, is not, strictly speaking, a deductively valid inference, and hence *not necessarily truth preserving*.

I take these two features to be shared, in particular, by all the philosophical thought experiments considered in the previous chapter. As an example, we can recall the one due to Edmund Gettier[325], whose form, as regimented under schema (α), we have not considered so far. Gettier's epistemological thought experiment, let me recall, was specifically designed in order to reject the classical tripartite analysis of knowledge, according to which knowledge could be defined in terms of justified true belief. Being a justified true belief, in other words, according to the same view, would be a *necessary* and *sufficient* condition for being knowledge. Once subsumed under schema (α) then, Gettier's examples become inferences of the following form:

- It is *possible* that Smith has justified true belief because of luck. [$\Diamond$ C].
- *If* knowledge is justified true belief, *then*, *if* Smith had justified true belief because of luck, *then* he would have knowledge. [T$\supset$ (C $\Box\!\!\rightarrow$ W)].
- *If* Smith had justified true belief because of luck, *then* he would *not* have knowledge. [C $\Box\!\!\rightarrow \neg$ W]
- *Therefore*, knowledge is *not* justified true belief. [$\therefore \neg$T].

[324] *Ibid* (127). My emphases.
[325] See 2.2.5 above.

As I already mentioned above, I take the *non-ampliativity* of this inference to be implicated by the cleansing model that I endorse, according to which philosophical thought experiments aim at generating justified belief revision by detecting inconsistencies contained in our system of beliefs. With respect to the present case, in particular, it seems indeed plausible to maintain that the counterfactual scenario appealed to by Gettier's thought experiment has not the immediate effect of generating any *new* beliefs about knowledge, but rather that of providing evidence for the fact that our *current* belief, according to which knowledge would be justified true belief, is not in fact consistent with the rest of our currently held beliefs, and must therefore be revised.
As to our second feature then, namely the fact of being *not necessarily truth-preserving*, I believe that it will be sufficient to point out that, insofar as the argumentative core of Gettier's brilliant thought experiment seems to be fully captured by the logical form of schema (α), we seem justified in denying to the kind of reasoning which takes place in it the power to transfer (without exceptions) the truth of its premises to the truth of its conclusions.

I would like to conclude this chapter with the following consideration. In his much discussed *prolegomena* to a pragmatic theory of cognitive evaluation, Stephen Stich put down his Jamesian pragmatist *credo* in the following words:

> "*There are no intrinsic epistemic virtues*. Rather, for the pragmatist, cognitive mechanisms or processes are to be viewed as tools or policies and evaluated in much the same way that we evaluate other tools or policies. One system of cognitive mechanisms is preferable to another if, in using it, we are more likely to achieve those things that we intrinsically value"[326].

Sharing a similar pragmatic spirit, Cellucci holds that any attempt at justifying inference rules cannot avoid taking into account their role in knowledge. In order to do this, he contrasts the *validation* of an inference rule with what he, following Feigl, calls its *vindication*[327]. Contrary to the former, which aims at proving that any given inference rule can be derived from other inference rules, "to *vindicate* an inference rule", he writes, "is to demonstrate that it is appropriate to a certain end"[328]. According to his opinion, which he argues for at some length, some basic inference rules cannot be validated but only vindicated[329]. In particular he maintains that, while the ultimate *non-deductive* rules, on the one hand, would be the ones appropriate to

[326] Stich (1990: 24).
[327] He borrows this distinction from Feigl (1971).
[328] Cellucci (2011: 133).
[329] "There must be both some ultimate deductive and non-deductive rules", he writes, "that cannot be validated. For these one can only seek vindication". *Ibid* (137).

the end of discovering new hypotheses, the ultimate *deductive* and *abductive* rules, on the other, would be the ones appropriate to the end of making explicit the content (or part of it) that is implicit in their premises[330]. If my proposal is on the right track, then this should also be regarded as the primary purpose of those fascinating cognitive tools that go under the name of philosophical thought experiments.

[330] See *Ibid* (134-137).

Conclusions

"*We can improve our conceptual scheme, our philosophy, bit by bit while continuing to depend on it for support; but we cannot detach ourselves from it and compare it objectively with an unconceptualized reality*"[331]

W.V.O. Quine

The time has now come to take stock of our previous considerations and to make room for them into as organic a picture of their subject matter as they allow. In the introduction to the present work, I assimilated my study of philosophical thought experiments to the investigations of an inquisitive naturalist, trying to come to grips with a very puzzling feature of human cognitive life. I am well aware of the fact that this simile is unavoidably bound to sound intolerably unpalatable to some hard-headed rationalist. Oddly enough, over a century an a half after the publishing of Charles Darwin's *On the Origin of Species*, there still reigns, it seems, amongst many philosophers, an insurmountable resistance towards acknowledging the harmless fact that, insofar as human *beings* are part of nature, human *reasoning* itself can be considered as natural a phenomenon as any other, the investigation of which falls squarely within the limits of natural science.

I find it symptomatic of the present state of affairs, for instance, that a naturalistic minded philosopher such as Hilary Kornblith might have found it necessary to stress the fact that "a view of philosophy as empirically informed", as the closing lines of his book on *Knowledge and its Place in Nature* soothingly read, "does not take philosophy away from philosophers"[332]. Philosophy, in particular, according to the view he articulates in his book, may and should be

[331] Quine (1953: 79).
[332] Kornblith (2002: 176).

regarded as an intellectual activity aimed at *empirically informed theory construction*[333]. As I hope my previous considerations have indirectly contributed to make clear, I share vary similar views concerning the proper role of philosophical inquiry. With respect, in particular, to the apparently profound dilemma of whether philosophical activity ought to be envisioned as 'internal' or as 'external' to natural science, I must confess that, despite the best of intentions, I find it too idle a question to be seriously addressed.

As an understandable consequence of this general attitude towards philosophy, I am far from willing to abjure my heretical simile, and in fact I intend to exploit it even further, in order to summarize the partial conclusions that I have arrived at concerning the nature of philosophical thought experiments. So bear with me for the last few pages! For expository reasons, I propose to start by drawing a very loose, an yet, as we shall presently see, remarkably profitable analogy between the single functional components of living organisms, namely *organs*, and a large family of functional components of human mental life, namely *rational processes*. In the last chapter of the present work, after all, I claimed that a philosophical thought experiment would be best seen and treated as a cognitive *tool*, and the ancient Greek word for instrument, *organon*, seems to suggest that our analogy is not too far-fetched. According to this analogy then, when I say that a thought experiment could be fruitfully regarded as a rational *instrument*, intended to perform specific tasks, I mean it much in the same sense in which ancient Aristotelian commentators considered the logical work of their teacher as an *organon*, i. e. an instrument philosophers might use in order to cope with their specific philosophical problems.

Suppose then that one lucky day our imaginary naturalist were to discover a new species on a remote island. Suppose further that, at the moment of dissecting the body of the unlucky animal, and to his great surprise, he were to find within it a mysterious *organ*, whose peculiar aspect were to strike him as slightly familiar and yet at the same time different from anything that he had ever encountered before in his previous anatomical explorations. At this point, in compliance with a long established methodology amongst his fellow anatomists, he might start 'interrogating nature' by asking three different, though closely related, questions. The first of these questions would concern the organ's *anatomy*, and could be readily addressed by asking: *What is its structure?* What does the organ look like? The second question, in turn, would concern its *physiology*, that is its inner functioning, and could therefore be addressed by simply asking: *What does it do?* How does it function? The third and final question would concern its *purpose*, and would be normally addressed by asking: *Why does it do it?* To what end?

One of the reasons that make me regard the above analogy as particularly illuminating for the present purposes, is that a promising way to look at philosophical analysis, I believe, is to

[333] Kornblith (2002: 177).

envision it as a non invasive attempt at dissecting human rationality. Accordingly, in what follows, I will try to state tautly what I take to be the main results of the present work by following the same tripartite structure. I will therefore start by providing an answer to the first question, concerning the *anatomy* of our mysterious "organ", namely: What is the structure of a philosophical thought experiment? After considering various proposals, I have come to believe that, insofar as a number of *premises* intended to support a specific *conclusion* are clearly distinguishable in it, a philosophical thought experiment can be considered as either being or being reducible to a peculiar kind of *argument*. A characteristic feature of these arguments, I have further contended, is that, insofar as they rely crucially on the positing of *counterfactual* scenarios, they are inherently *modal*, and raise specific epistemic problems related to the justification of counterfactual conditionals. In particular, as I have tried to argue, I believe it is reasonable to maintain that the typical logical structure of most philosophical thought experiments is the one indicated by Roy Sorensen and Sören Häggqvist in their respective work, namely the structure of a paradox, i.e. a set of individually plausible yet jointly inconsistent sentences. The formal regimentation that best reflects the dialectical progression of a philosophical thought experiment, I have also suggested, seems to be the one put forward by Sören Häggqvist and referred to above as schema (α).

The regimentation attempt just mentioned, allows us now to provide an answer to the second question, concerning the *physiology* of our epistemic tool, namely: How does a philosophical thought experiment function? What does it do? As I have tried to show by expounding and analysing the inner workings of Häggqvist's schema, most philosophical thought experiments are specifically designed by their creators to play a fundamentally *negative* role. Their proper function is that of rejecting a given target statement, or in general a target theory, by disproving one of its modal consequences. They perform this function, in particular, by exploiting our psychological ability to *contemplate* counterfactual scenarios and to make reliable *judgements* concerning their most likely or unlikely consequences. Insofar as we expect our best theories to be 'counterfactual supporting', i.e. we normally take them to be more or less implicitly committed to certain modal consequences, a philosophical thought experiment will typically appeal to a counterfactual scenario in order to show that the modal consequences the theory would be committed to if the particular state of affairs were to actually occur, fall short of matching our current intuitions. These intuitions, in particular, as I tried to show, once couched within a naturalistic framework, become a fairly uncontroversial component of philosophical inquiry. To the extent that they are not taken to provide some sort of unintelligible epistemic access to a realm of eternal *a priori* truths, that is, and to the extent that they are placed within a reliabilist epistemological framework, intuitions help us making our folk-theories explicit. In this

sense, philosophical thought experiments afford a fallible, but nonetheless valuable instrument to explore relations amongst intuitions not immediately detectable by introspection, indirectly showing us the limits of our introspective powers. No one interested in understanding the nature of human knowledge, I believe, would seriously consider for a minute the possibility of getting entirely rid of *perception* on the basis that it has often been proved to be *fallible*. The same inquirer would probably consider far more reasonable the painstaking effort of trying to single out specific conditions under which our perception is most likely to err. Something similar, I think, applies to our *intuitions*.

We are now left with our third and final question, concerning the *purpose* of our tool, namely: What is the purpose, or at least the primary purpose of a philosophical thought experiment? Why does it perform the function that we have just described? Following a suggestion advanced by Roy Sorensen, which I find to be reminiscent of Thomas Kuhn's previous reflection on the same topic, I have tried to show that most philosophical thought experiments are primarily aimed at generating *justified belief revision* by detecting inconsistencies contained in our more or less conscious system of beliefs. The same idea can be put by saying that, according to the construal I have tried to defend, they are instruments which may enable us to decide whether certain beliefs we 'live by' are consistent with the rest of our beliefs and may therefore be used as a compass to navigate safely the theoretical world. Their purpose, that is, is not just that of eliciting intuitions (intuition pumps), but to keep them under control. They do this by increasing the level of reflective equilibrium amongst intuitions themselves. Reflective equilibrium amongst intuitions, of course, would be a necessary but not sufficient condition that any philosophical theory has to satisfy in order to play an explanatory role, and hence to produce knowledge. The missing sufficient condition, I contend, is provided by empirical science. It follows that the main purpose of most philosophical thought experiments is that of making us more *rational* and in so doing they conform to what Sorensen calls the *cleansing model* of armchair inquiry. An important side effect of their main purpose, I contend further, is that of enhancing our understanding of reality. As a matter of fact, while on the one hand I believe that thought experiments have to do with *understanding* more than with *knowledge*, I try to argue, on the other, that *understanding* itself has a lot more to do with *knowledge* than we usually think. It is a plain fact, I maintain, that our theoretical notions usually reflect deeper ontological commitments, i. e. that they normally reveal our expectations as to how the world must be, and by so doing they certainly do contribute significantly to our understanding of that world.

So, one might ask, what is the proper subject matter of thought experiments? What are these cognitive instruments produced by our imagination really talking about? The *world* or the *words*? It has recently been suggested that "broadly speaking, views about philosophical analysis

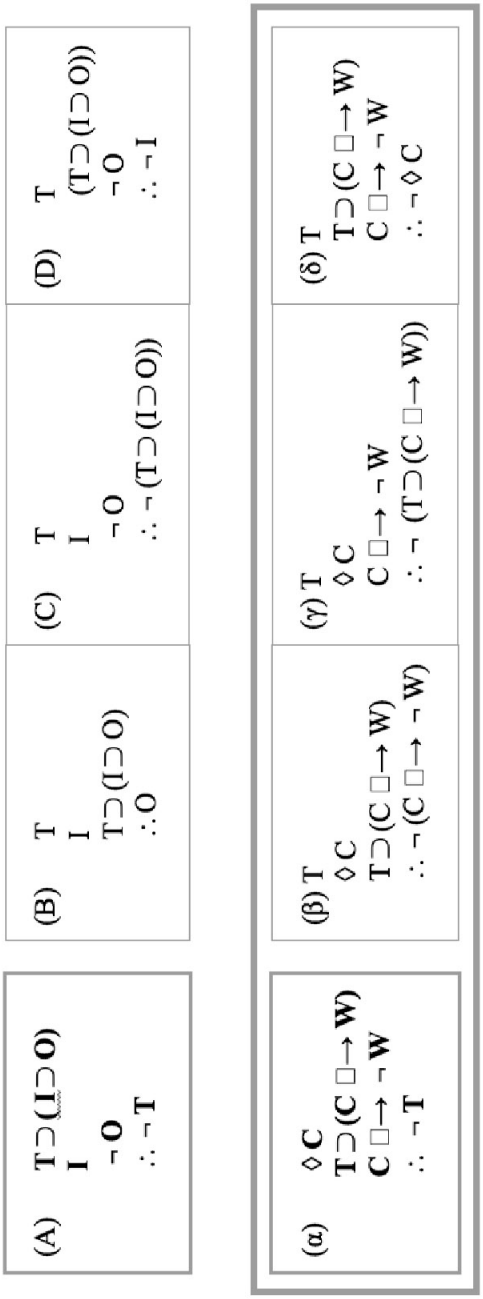
may be divided into those that take the targets of such analysis to be in-the-head psychological entities versus outside-the-head nonpsychological entities"[334]. According to a similar way to make the same point, the only two mutually exclusive answers to the question above, as applied to philosophical inquiry more generally, would be the following: Either (1) philosophy deals with the world, or (2) philosophy deals with the conceptual apparatus relying on which we deal with the world. Supporters of (1), in particular, usually hold that (2) introduces an unnecessary and highly artificial filter between us and the world, thereby launching a serious explanatory regress, and risking to fall pray of radical *skepticism* (of which, in their view, *idealism* would be one of the most unpalatable consequences). Their position, in other words, may be effectively epitomized by the following, intentionally provocative question: Why can't we just talk about things, instead of talking about *ways* of talking about things?

As I hope my work makes clear, I strongly believe that this way of looking at things dramatically oversimplifies matters. In other words, I think that Quine's "Neurathian" suggestion, which I quoted at the opening of the present conclusions, is on the right track, especially when trying to tackle the many difficult epistemological questions raised by philosophical thought experiments. In the same spirit, I think that any effort to show the fundamental epistemic role played by our conceptual framework does not necessarily amount to giving up the very idea of a mind-independent world. In other words, to hold that between an epistemic agent and reality or the world there is a conceptual framework, does not automatically commit one, I believe, to the thesis that the world does not exist (idealism), nor to the idea that an epistemic agent is hopelessly trapped within the boundaries of his own conceptual framework (solipsism). The limits of our linguistic resources, pace Wittgenstein, are *not* the limits of our worlds.

Let me conclude by saying that, as I have tried to argue at the end of the third chapter above, I am confident that new light might be shed in the future on the nature of philosophical thought experiments by a careful attempt at properly locating them against the wider background of a burgeoning and relatively new area of research, namely the study of so called *defeasible* or *non-monotonic reasoning*. The possibility of seeing a philosophical thought experiment as a very peculiar kind of *defeasible inference*, I particular, seems very promising as a starting point to begin charting, both from a logical and an epistemological point of view, that vast and still largely unknown territory in which *normative* and *descriptive* aspects of human cognition meet. It really seems to me like there is a lot of work ahead for our naturalist.

[334] Goldman and Pust (1998: 183).

Appendix



The argument schemas reproduced in the table are drawn from Häggqvist (1996, 2009b).

Bibliography

Antonelli, A. (2004) 'La Logica del Ragionamento Plausibile', in Floridi, L. (ed.) *Linee di Ricerca*, SWIF: 226-252, Sito Web Italiano per la Filosofia – ISSN 1126-4780 – www.swif.it/biblioteca/lr.

- Antonelli, A. (2005) *Grounded Consequence for Defeasible Logic*, Cambridge University Press.

- Antonelli, A. (2010) 'Non-monotonic Logic', *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, Edward N. Zalta (ed.).

Aristotle, *Metaphysics*, ed. by W. D. Ross, Clarendon Press, 1953.

Audi, R. (1998) *Epistemology. A Contemporary Introduction to the Theory of Knowledge*, Routledge.

Ayer, A. J. (1952 [1946]) *Language, Truth and Logic*, second edition, Dover Publications.

Bealer, G. (1993) 'The Incoherence of Empiricism', in Wagner, S. J. and Wagner, R. (eds.) *Naturalism: A Critical Appraisal*, University of Notre Dame Press: 163-96.

- (1998) 'Intuition and the Autonomy of Philosophy', in DePaul, M. and Ramsey, W. (eds.) (1998): 201-39.

- (2002) 'Modal Epistemology and the Rationalist Renaissance', in Gendler, T. and Hawthorne, J. *Conceivability and Possibility*, Clarendon Press: 71-125.

Bennett, J. (2003) *A Philosophical Guide to Conditionals*, Oxford University Press.

Bishop, M. A. (1999) "Why Thought experiments Are Not Arguments", *Philosophy of Science* 66, 4: 534-541.

Boghossian, P. and Peacocke, C. (eds.) (2000) *New Essays on the A Priori*, Oxford University Press.

BonJour, L. (1994) 'Against Naturalized Epistemology', *Midwest Studies in Philosophy* XIX: 283-300.

- (1998) *In defense of pure reason: a rationalist account of a priori justification*, Cambridge University Press.

Borghini, A. (2009) *Che Cos'è la Possibilità*, Carocci.

Braddon-Mitchell, D. and Jackson, F. (1996) *Philosophy of Mind and Cognition*, Blackwell Publishing.

Brown, J. R. (1986) 'The Structure of Thought Experiments', *International Studies in the Philosophy of Science: the Dubrovnik Papers* 1: 1-15.

- (1991a) "Thought Experiments: A Platonic Account", in Horowitz, T. and Massey, G. J. (eds.) *Thought Experiments in Science and Philosophy*, Rowman & Littlefield Publishers.

- (1991b) *Laboratory of the Mind: Thought Experiments in the Natural Sciences*, Routledge.

- (1993) "Why Empiricism Won't Work", *Proceedings of the Philosophy of Science Association* 2: 271-279.

- (2004a) "Why Thought Experiments Transcend Experience", in Hitchcock, C. (ed.) *Contemporary Debates in the Philosophy of Science*, Blackwell: 23-43.

- (2004b) "Peeking into Plato's Heaven", *Philosophy of Science* 71: 1126-1138.

Buckwalter, W. and Stich, S. 'Gender and Philosophical Intuition', forthcoming in Knobe, J. and Nichols, S. (eds.) *Experimental Philosophy*, Vol. 2, Oxford University Press.

Buzzoni, M. (2004) *Esperimento ed esperimento mentale*, Franco Angeli.

Carnap, R. (1971 [1934]) *The Logical Syntax of Language*, Routledge & Kegan Paul.

Cellucci, C. (2000) 'The Growth of Mathematical Knowledge: An Open World View', in Grosholtz, E. and Berger, H. (eds.) *The Growth of Mathematical Knowledge*, Kluwer: 153-176.

- (2011) 'Classifying and Justifying Inference Rules', in Cellucci, C. Grosholtz, E. and Ippoliti, I. (eds.) *Logic and Knowledge*, Cambridge Scholars Publishing: 123-143.

Chisholm, R. M. (1946) 'The Contrary-To-Fact Conditional', *Mind* 55: 219-307.

- (1955) 'Law Statements and Counterfactual Inference', *Analysis* 15: 97-105.

- (1957) *Perceiving*, Princeton University Press.

- (1966) *Theory of Knowledge*, Prentice-Hall.

Cohen, M. (2005) *Wittgenstein's Beetle and Other Classic Thought Experiments*, Blackwell.

Cohnitz, D. and Häggqvist, S. (2009) 'The Role of Intuitions in Philosophy', *Studia Philosophica Estonica* 2.2: 1-14.

Crane, T. (1991) 'All the Difference in the World', *The Philosophical Quarterly* 41, 162: 1-26. Reprinted in Pessin, A. and Goldberg, S. (eds.) (1996) *The Twin Earth Chronicles. Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'"*, M. E. Sharpe: 284- 304.

Daniels, N. (2011) "Reflective Equilibrium", *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Zalta, E. N. (ed.) URL = <http://plato.stanford.edu/archives/spr2011/entries/reflective-equilibrium/>.

Dawkins, R. (2006 [1976]) *The Selfish Gene*, third edition, Oxford University Press.

Dennett, D. (1984) *Elbow Room. The Varieties of Free Will Worth Wanting*, MIT Press.

- (1991) *Consciousness Explained*, Little Brown.

DePaul, M. and Ramsey, W. (eds.) (1998) *Rethinking Intuition. The Psychology of Intuition and Its Role in Philosophical Inquiry*, Rowman & Littlefield Publishers.

Descartes, R. (1641) *Meditations, Objections, and Replies*, edited and translated by Ariew, R. and Cress, D., Hackett Publishing Company, 2006.

Duhem, P. (1914) *The Aim and Structure of Physical Theory*, transl. by Wiener, P., Princeton University Press, 1954.

Fano, V. (2005) *Comprendere la scienza. Un'introduzione all'epistemologia delle scienze naturali*, Liguori.

- (2007) "Un fattore epistemico inaffidabile nella scoperta scientifica. Einstein che insegue un raggio di luce", *Protagora* 35:15-31, volume monografico, *Albert Einstein filosofo e metodologo*, edited by Fano, V., Minazzi, F. and Tassani, I.

Feigl, H. (1971) 'De Principiis Non Disputandum…? On the Meaning and the Limits of Justification', in Black, M. (ed.) *Philosophical Analysis. A collection of Essays*, Books for Libraries: 113-147.

Fine, K. (2002) 'The Varieties of Necessity', in Gendler, T. and Hawthorne, J. (eds.) *Conceivability and Possibility*, Clarendon Press.

Fodor, J. (1964) 'On Knowing What We Would Say', *The Philosophical Review* 73, 2: 198-212.

Foley, R. (1998) 'Rationality and Intellectual Self-Trust', in DePaul, M. and Ramsey, W. (eds.) *Rethinking Intuition. The Psychology of Intuition and Its Role in Philosophical Inquiry*, Rowman & Littlefield Publishers: 241-56.

Gettier, E. (1963) 'Is Justified True Belief Knowledge?', *Analysis* 23: 121-123.

Goldman, A. I. (1967) 'A causal Theory of Knowing', *Journal of Philosophy* 64: 357-72.

- (1975) ‘Innate Knowledge’, in Stich, S. P. (ed.) *Innate ideas*, University of California Press.

- (1976) ‘Discrimination and Perceptual Knowledge’, *Journal of Philosophy* 72: 771-91.

- (1980) ‘What is Justified Belief’, in Pappas, G. (ed.) *Justification and Knowledge*, Reidel.

- (1986) *Epistemology and Cognition*, Harvard University Press.

- (2007) ‘Philosophical Intuitions: their Target, their Source, and their Epistemic Status’, *Grazer Philosophische Studien* 74: 1-26.

Goldman, A. and Pust, J. (1998) ‘Philosophical Theory and Intuitional Evidence’, in DePaul, M. and Ramsey, W. (eds.) (1998): 179-197.

Gooding, D. (1992) ‘What is *Experimental* about Thought Experiments?’, *PSA*: *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Volume Two: Symposia and Invited Papers: 280-290.

- (1992b) ‘The Procedural Turn; or, Why Do Thought Experiments Work?’, in Giere, R. (ed.) *Cognitive Models of Science*, *Minnesota Studies in the Philosophy of Science*, 15, University of Minnesota Press: 45-76.

Goodman, N. (1947) ‘The Problem of Counterfactual Conditionals’, *The Journal of Philosophy* 44, 5: 113-128. Reprinted in Goodman, N. (1983): 3-27.

- (1983) *Fact, Fiction, and Forecast*, Fourth Edition, Harvard University Press.

Gopnik, A. and Schwitzgebel E. (1998) ‘Whose Concepts Are They, Anyway? The Role of Philosophical Intuition in Empirical Psychology’, in DePaul, M. and Ramsey, W. (eds.) (1998): 75-91.

Gutting, G. (1998) ‘“Rethinking Intuition”: A Historical and Metaphilosophical Introduction’, in DePaul, M. and Ramsey, W. (eds.) (1998): 3-13.

- (2009) *What Philosophers Know. Case Studies in Recent Analytic Philosophy*, Cambridge University Press.

Häggqvist, S. (1996) *Thought Experiments in Philosophy*, Almqvist & Wiksell International.

- (2007) 'The A Priori Thesis: A Critical Assessment', *Croatian Journal of Philosophy* 19: 47-61.

- (2009a) 'Modal Knowledge and the Form of Thought Experiments', in Kompa, N., Nimtz, C. and Suhm, C. (eds.) *The A Priori and its Role in Philosophy*, Mentis Verlag.

- (2009b) 'A Model for Thought Experiments', *Canadian Journal of Philosophy* 39, 1: 55-76.

Haldane, J.B.S. (1927) *Possible Worlds and Other Papers*, Chatto & Windus.

Harman, G. (1965) 'The Inference to The Best Explanation', *The Philosophical Review* 74, 1: 88-95.

Heil, J. (2004) *Philosophy of Mind*, *A Contemporary Introduction*, Routledge.

Hempel, C. G. (1966) *Philosophy of Natural Science*, Prentice-Hall.

Hintikka, J. (1999) 'The Emperor's New Intuitions', *The Journal of Philosophy* 96, 3: 127-147.

Hintikka, J., and Sandu, G. (2007) 'What Is Logic?', in Jacquette, D. (ed.) *Philosophy of Logic*, North-Holland: 13-39.

Horgan, T. (1984) 'Jackson on Physical Information and *Qualia*', *Philosophical Quarterly* 34: 147-152. Reprinted in Ludlow, P., Nagasawa, Y. and Stoljar, D. (eds.) 2004: 302-308.

Horowitz, T. and Massey, G. J. (eds.) (1991) *Thought Experiments in Science and Philosophy*, Rowman & Littlefield Publishers.

Ierodiakonou, K., and Roux, S. (eds.) (2011) *Thought Experiments in Methodological and Historical Contexts*, Brill.

Jackson, F. (1982) 'Epiphenomenal Qualia', *Philosophical Quarterly* 32: 127-36.

- (1986) 'What Mary Didn't Know', *Journal of Philosophy*, 83: 291-95.

- (1994) 'Armchair Metaphysics', in Michaelis, M. and O'Leary-Hawthorne, J. (eds.) *Philosophy in Mind, The Place of Philosophy in the Study of Mind*, Kluwer Academic Publishers: 23-42.

- (1998) *From Metaphysics to Ethics. A Defence of Conceptual Analysis*, Oxford University Press.

Kahneman, D., Slovic, P., and Tversky, A. (1982) *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kim, J. (1988) 'What is "Naturalized Epistemology"?', *Philosophical Perspectives* 2: 381-405.

Kitcher, P. (1980) 'A priori Knowledge', *Philosophical Review* 89: 3-23.

- (1984) *The Nature of Mathematical Knowledge*, Oxford University Press.

- (1992) 'The Naturalists Return', *The Philosophical Review* 101, 1: 53-114.

- (1998) 'Kant's *A Priori* Framework', in Kitcher, Patricia (ed.) *Kant's Critique of Pure Reason. Critical Essays*, Rowman & Littlefield Publishers.

Knobe, J. and Nichols, S. (eds.) (2008) *Experimental Philosophy*, Oxford University Press.

Koons, R. (2009) 'Defeasible Reasoning', *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*, Edward N. Zalta (ed.).

Kornblith, H. (1998) 'The Role of Intuition in Philosophical Inquiry: An Account with No Unnatural Ingredients', in DePaul, M. and Ramsey, W. (eds.) (1998):129-41.

- (2002) *Knowledge and Its Place in Nature*, Oxford University Press.

Koyré, A. ([1960] 1968) "Galileo's Treatise *De motu gravium*: The Use and Abuse of Imaginary Experiment", in Koyré, A. (ed.) *Metaphysics and Measurement*, Chapman and Hall. Originally published in *Revue d'histoire des sciences* 13: 197-245.

Kripke, S. (1980) *Naming and Necessity*, Harvard University Press.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*, The University of Chicago Press.

- (1964) 'A function for Thought Experiments', in *The Essential Tension*, University of Chicago Press, 1977.

Kusch, M. (1995) *Psychologism. A Case Study in the Sociology of Philosophical Knowledge*, Routledge.

- (2011) "Psychologism", *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*, Edward N. Zalta (ed.).

Lakoff, G. and Johnson, M. (1980) *Metaphors we live by*, The University of Chicago Press.

Lewis, D. (1973) *Counterfactuals*, Blackwell.

Locke, J. (1690) *An Essay Concerning Human Understanding*, Abridged and Edited by Pringle-Pattison, A. S., Humanities Press, 1978.

Loux, M. J. (2006) *Metaphysics. A Contemporary Introduction*, Routledge.

- (ed.) (2008) *Metaphysics. Contemporary Readings*, Routledge.

Ludlow, P., Nagasawa, Y., Stoljar, D. (eds.) (2004) *There's Something About Mary. Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, MIT Press.

Ludwig, K. (2007) 'The Epistemology of Thought Experiments: First Person versus Third Person Approaches', *Midwest Studies in Philosophy of Science* 31: 128-159.

Lycan, W. G. (2008) *Philosophy of Language. A Contemporary Introduction*, Routledge.

Mach, E. (1960 [1883]), *The Science of Mechanics*, trans J. McCormack, 6th edn, Open Court.

- (1976 [1905]), "On Thought Experiments", in *Knowledge and Error*, trans. by McCormack, J., Reidel: 134-147.

Maffie, J. (1997) '"Just-so" stories about "inner cognitive Africa": some doubts about Sorensen's evolutionary epistemology of thought experiments', *Biology and Philosophy* 85: 207-224.

Margolis, E. and Laurence, S. (2011) "Concepts", *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Edward N. Zalta (ed.).

McCarthy, J. (1980) 'Circumscription – A Form of Non-monotonic Reasoning', *Artificial Intelligence* 13: 27-39, 171-172.

McDermott, D. and Doyle, J. (1980) 'Non-monotonic Logic I', *Artificial Intelligence* 13: 41-72.

Melia, J. (2003) *Modality*, McGill-Queen's University Press.

Moue, A. S., Masavetas, K. A. and Karayianni, H. (2006) 'Tracing the Development of Thought Experiments in the Philosophy of Natural Sciences', *Journal for General Philosophy and Science / Zeitschrift für Allgemeine Wissenschaftstheorie* 37, 1: 61-75.

Nagel, T. (1974) 'What Is It Like to Be a Bat?' *Philosophical Review* 83: 435-450.

Newton, I. (1934 [1689]) *Philosophiae Naturalis Principia Mathematica*, trans. by Motte, A. (1729) rev. Florian Cajori, University of California Press.

Nisbett, R. E. and Ross, L. (1980) *Human Inference: Strategies and Shortcomings of Social Judgement*, Prentice-Hall.

Norton, J. (1991) "Thought Experiments in Einstein's Work", in Horowitz and Massey (1991): 129-148.

- (1993) “Seeing the Laws of Nature”, *Metascience* 3 (new series): 33-38.

- (1996) "Are Thought Experiments Just What You Thought?" *Canadian Journal of Philosophy* 26: 333-366.

- (2004a) "On Thought Experiments: Is There More to the Argument?" *Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association, Philosophy of Science* 71: 1139-1151.

- (2004b) "Why Thought Experiments Do Not Transcend Empiricism", in Hitchcock, C. (ed.) *Contemporary Debates in the Philosophy of Science*, Blackwell: 44-66.

Papineau, D. (2009) “Naturalism”, *The Stanford Encyclopedia of Philosophy (Spring 2009 Edition)*, Edward N. Zalta (ed.).

Peirce, Ch. S. (1868) ‘Questions Concerning Certain Faculties Claimed for Man’, *Journal of Speculative Philosophy* 2: 103-114.

Pessin, A. and Goldberg, S. (eds.) (1996) *The Twin Earth Chronicles. Twenty Years of Reflection on Hilary Putnam’s “The Meaning of ‘Meaning’”*, M. E. Sharpe.

Pollock, J. L. (1970) ‘The Structure of Epistemic Justification’, *American Philosophical Quarterly* (Monograph Series) 4: 62-78.

- (1974) *Knowledge and Justification*, Princeton University Press.

- (1987) ‘Defeasible Reasoning’, *Cognitive Science* 11: 481-518.

- (1995) *Cognitive Carpentry*, MIT Press.

Putnam, H. (1975) "The Meaning of 'Meaning'", *Minnesota Studies in the Philosophy of Science* VII:131-93. Reprinted in Pessin, A. and Goldberg, S. (eds.) (1996): 3-52.

Quine, W.V. (1947) 'The problem of interpreting modal logic', *Journal of Symbolic Logic* 12, 2:43-48.

- (1953) 'Three grades of modal involvement', *Proceedings of the Xth International Congress of Philosophy* (Brussels): 156-174.

- (1980 [1954]) *From a Logical point of View*, Harvard University Press.

- (1960) *Word and Object*, MIT Press.

Ramsey, W. (1992) 'Prototypes and Conceptual Analysis', *Topoi* 11: 59-70. Reprinted in DePaul, M. and Ramsey, W. (eds.) (1998): 161-177.

Reichenbach, H. (1938) *Experience and Prediction: An Analysis of the Foundations of Science*, University of Chicago Press.

Reiter, R. (1980) 'A Logic for Default Reasoning', *Artificial Intelligence* 13: 81-132.

Rescher, N. (2001) *Philosophical Reasoning: A study in the Methodology of Philosophizing*, Blackwell Publishers Inc.

- (2005) *What if? Thought Experimentation in Philosophy*, Transaction Publishers.

Rey, G. (1983) 'Concepts and Stereotypes', *Cognition* 15: 237-62.

Rorty, R. M. (ed.) (1992 [1967]) *The Linguistic Turn. Essays in Philosophical Method*, The University of Chicago Press.

Rosenberg, A. (2005) *Philosophy of Science*, Second edition, Routledge.

Rosch, E. and Mervis, C. B. (1975) 'Family Resemblances: Studies in the Internal Structure of Categories', *Cognitive Psychology* 7: 573-605. Reprinted in DePaul, M. and Ramsey, W. (eds.) (1998): 17-44.

Rosch, E. (1978) 'Principles of Categorization', in Rosch, E. and Lloyd, B. (eds.) *Cognition and Categorization*, Lawrence Erlbaum Associates: 27-48.

Russell, B. (1912) *The Problems of Philosophy*, Williams and Norgate.

- (2005 [1914]) *Our knowledge of the external world*, Routledge.

Rynasiewicz, R. (2011) "Newton's Views on Space, Time, and Motion", *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, Edward N. Zalta (ed.).

Smith, E. and Medin, D. (1981) *Concepts and Categories*, MIT Press.

Sobel, D. M. (2004) 'Exploring the Coherence of Young Children's Explanatory Abilities: Evidence from Generating Counterfactuals', *British Journal of Developmental Psychology* 22: 37-58.

Sorensen, R. (1992) *Thought Experiments*, Oxford University Press.

- (1992b) 'Thought Experiments And The Epistemology Of Laws', *Canadian Journal of Philosophy* 22: 15-44.

Sosa, E. (1998) 'Minimal Intuition', in DePaul, M. and Ramsey, W. (eds.) (1998): 257-69.

Stich, S. (1988) 'Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity', *Synthese* 74: 391-413. Reprinted in DePaul, M. and Ramsey, W. (eds.) (1998): 95-112.

- (1990) *The Fragmentation of Reason*, MIT Press.

Stoljar, D. (2009) "Physicalism", *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*, Edward N. Zalta (ed.).

Strawson, P. F. (1959) *Individuals. An Essay in Descriptive Metaphysics*, Routledge, 2003.

Szabó Gendler, T. (1998) 'Galileo and the Indispensability of Scientific Thought Experiment', *The British Journal for the Philosophy of Science* 49, 3: 397-424.

- (2000) *Thought Experiment: On The Power and Limits of Imaginary Cases*, Routledge.

- (2004) 'Thought Experiments Rethought – and Reperceived', *Philosophy of Science* 71, 5: 1152-1163.

- (2007) 'Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium', *Midwest Studies in Philosophy of Science* 31: 68-89.

- (2011) *Intuition, Imagination, and Philosophical Methodology*, Oxford University Press.

Szabó Gendler, T. and Hawthorne, J. (2005) 'The Real Guide to Fake Barns: A Catalogue of Gifts For Your Epistemic Enemies', *Philosophical Studies*, 124: 331-352.

Weatherson, B. (2003) 'What Good Are Counterexamples?', *Philosophical Studies* 115, 1: 1-31.

Weinberg, J. M., Nichols, S., Stich, S. (2001) 'Normativity and Epistemic Intuitions', *Philosophical Topics* 29, 1 & 2: 429-460.

Wilkes, K. V. (1988) *Real People. Personal Identity without Thought Experiments*, Clarendon Press.

Williamson, T. (2004) 'Philosophical 'Intuitions' and Scepticism about Judgement', *Dialectica* 58, 1: 109-153.

- (2007) *The Philosophy of Philosophy*, Blackwell Publishing.

Witt-Hansen, J. (1976) 'H.C. Örsted, Immanuel Kant, and the Thought Experiment', *Danish Yearbook of Philosophy* 13: 48-65.

Wittgenstein, L. (2009 [1953]), *Philosophical Investigations*, fourth edition, Blackwell Publishing.