Philosophia Scientiæ

Travaux d'histoire et de philosophie des sciences

21-3 | 2017

N'allez pas croire !

Philosophical Thought Experiments: The Case for Engel

Diego Marconi

Electronic version

URL: <http://journals.openedition.org/philosophiascientiae/1304>

DOI: 10.4000/philosophiascientiae.1304

ISSN: 1775-4283

Publisher

Éditions Kimé

Printed version

Date of publication: 30 October 2017

Number of pages: 111-124

ISBN: 978-2-84174-816-7

ISSN: 1281-2463

Electronic reference

Diego Marconi, "Philosophical Thought Experiments: The Case for Engel", *Philosophia Scientiæ* [Online], 21-3 | 2017, Online since 30 October 2019, connection on 31 March 2021. URL: <http://journals.openedition.org/philosophiascientiae/1304> ; DOI: <https://doi.org/10.4000/philosophiascientiae.1304>

Tous droits réservés

Philosophical Thought Experiments: The Case for Engel

Diego Marconi

Università di Torino (Italy)

Résumé : Je m'efforcerai de donner plus de poids aux doutes de Pascal Engel relatifs à deux affirmations de Timothy Williamson dans *Philosophy of Philosophy* : que les expériences de pensée philosophiques portent sur une possibilité métaphysique par opposition à une possibilité conceptuelle, et que le raisonnement contrefactuel joue un rôle crucial pour atteindre les conclusions modales pertinentes dans les arguments fondés sur des expériences de pensée. Dans le premier cas, je soutiendrai qu'à moins de comprendre la notion de conceptuel dans un sens psychologique, il est difficile de saisir la différence entre des expériences de pensée qui « portent sur des concepts » ou possibilités conceptuelles et des expériences de pensée qui « portent sur des êtres réels » ou possibilités métaphysiques. Dans le second cas, je soutiendrai qu'il n'est pas clair que les contrefactuels sont des prémisses indispensables dans des arguments fondés sur des expériences de pensée ou que le raisonnement contrefactuel prend l'avantage sur le raisonnement déductif ordinaire pour atteindre des conclusions modales.

Abstract: I will try to add some weight to Pascal Engel's doubts about two claims of Timothy Williamson's in *The Philosophy of Philosophy*: that philosophical thought experiments are about metaphysical possibility as opposed to conceptual possibility, and that counterfactual reasoning plays a crucial role in reaching modal conclusions relevant to arguments based on thought experiments. In the first case, I will argue that unless the notion of conceptual is understood psychologically, it is hard to tell the difference between thought experiments being "about concepts", or conceptual possibilities, and their being "about real things", or metaphysical possibilities. In the second case, I will argue that it not clear that counterfactuals are indispensable premises in arguments based on thought experiments, or that counterfactual reasoning has an epistemological edge over standard deductive reasoning in reaching modal conclusions.

In his [Williamson 2007], Timothy Williamson produced a battery of arguments against the widespread conception that philosophy is conceptual analysis, and philosophical knowledge—if there is any—is knowledge about concepts. One of his sub-claims is that philosophical thought experiments deal with metaphysical, not conceptual possibilities [Williamson 2007, 188, 205–207]; another is that modal knowledge—knowledge of metaphysical possibilities and necessities—does not require any special faculty of modal intuition; in fact, it is the sort of knowledge we often achieve by way of counterfactual reasoning, which is itself quite an ordinary business, and “deeply integrated into our empirical thought in general” [Williamson 2007, 141]. In his article on thought experiments “in or out of the armchair” [Engel 2011], Pascal Engel objects to the first claim: he argues that Williamson has not really shown that *no* thought experiments bear upon our concepts or about our understanding of concepts [Engel 2011, 158]. As to the second claim, Engel wonders, somewhat *en passant*, whether Williamson’s counterfactuality thesis, i.e.,

There is no more to our modal knowledge of possible facts than our capacity to handle and evaluate counterfactuals.
[Engel 2011, 151]

really casts any light on such issues as what kind of possibility claims are involved in thought experiments [Engel 2011, 146], whether such possibilities are genuine [Engel 2011, 146], and what kind of epistemic access we have to the situations described in philosophical thought experiments [Engel 2011, 151]. By and large, I am sympathetic with Engel’s criticism. I believe that (i) though some formulations of thought experiments (or of modal issues in general) may invite a metaphysical reading while others may invite a conceptual reading, the metaphysical vs. conceptual issue is hard to decide in general, as much depends on what is meant by “conceptual”. Furthermore, I believe that (ii) though bringing modal knowledge back to knowledge based on counterfactual reasoning may make modal knowledge less arcane, it hardly dissipates the difficulties that are involved in the evaluation of thought experiments or of modal claims in general. In genuinely problematic cases, we are not going to be decisively helped by choosing counterfactual formulations even where we can do so. In what follows, I will mainly be dealing with (ii). (i) involves vast issues—indeed, it is the topic of Williamson’s book as a whole—so I will limit myself to a few comments, much in the spirit of Engel’s remarks.

1 Conceptual vs. metaphysical

For Williamson, the notion of conceivable is a psychological notion. E.g.,

Although there are truths and falsities about conceivability and inconceivability, they concern our mental capacities.
[Williamson 2007, 135]

It is then to be expected that something may turn out to be both conceivable and impossible, for example, physically impossible:

We easily conceive particles violating what are in fact physical laws. [Williamson 2007, 135]

Indeed, it would be a sort of Darwinian miracle if our psychological abilities were attuned to deep natural regularities as understood by modern physics. However, this is not the only way “conceivable” was used in the analytic tradition. For example:

Of course, here “I can’t imagine the opposite [*ich kann mir das Gegenteil nicht vorstellen*]” doesn’t mean: my powers of imagination are unequal to the task. [Wittgenstein 1953, §251], [Wittgenstein 1974, 129]

Similarly, when Gilbert Ryle wrote:

Whatever I do or say, it is always conceivable, though nearly always false, that I did it, or said it, in complete absence of mind. [Ryle 1949, 91], [Engel 2011, 151]

he did not mean that I could somehow picture *myself* as acting or speaking “in complete absence of mind”. What he meant was that it is not obviously inconsistent to describe some human action as mechanistic, a mere reflex, or unconsciously executed. In this use, “conceivable” is very close to what other philosophers meant by “logically possible”, i.e., describable in language without contradiction [see e.g., Russell 1921, 159–160], [Schlick 1936].

“Conceivable” is an ordinary word that can be philosophically explicated in several different ways, as became clear in the recent debate concerning the controversial inference from conceivability to possibility.¹ None of the proposed explications turned out to be entirely satisfactory, this being perhaps the main reason why the issue is now somewhat moot.² Though “conceivable” is obviously related to “concept”, it is by no means clear that it is used synonymously with “conceptually possible”. Indeed, if “conceivable” is understood as “imageable”—such as we can form a mental picture of—then objects like Descartes’s chylagon would turn out to be conceptually impossible, which is certainly unintended by most philosophers who use the phrase “conceptually possible”. More generally, it seems clear that “concept” has both a psychological and a non-psychological sense. When we inquire which concepts can be attributed to our primate predecessors, or which pathologies tend to deprive us of the concepts of body parts, we are doing biology or neuropsychology and taking concepts to be categorizing functions in the human mind; whereas when we wonder whether some thing falls under the concept of chair, our question is undistinguishable from the question whether that thing is a chair—on the

1. See the papers in [Szabó Gendler & Hawthorne 2002] and the editors’ introduction, particularly [Szabó Gendler & Hawthorne 2002, 7–9].

2. For criticism of several notions of conceivability, see [Marconi 2009]. For more fine-grained distinctions within a two-dimensionalist framework [Chalmers 2002].

face of it, not a psychological question.³ Elsewhere, I argued that these are two different notions of concept, that they are irreducible to each other, and that failure to distinguish them is bound to generate confusion [Marconi 2015]. This is relevant to the issue of whether thought experiments are about concepts or about “real things”. Williamson argues that if a thought experiment showed that some state of affair is *conceptually* possible (e.g., that cases of justified true belief without knowledge are conceptually possible), that would be philosophically uninteresting, as it might still be the case that knowledge is identical with justified true belief of metaphysical necessity [Williamson 2007, 206]. This presupposes a difference between being conceptually possible and being possible *tout court*, or metaphysically possible. Engel, in turn, argues that at least some thought experiments appear to be about concepts: they test our common concept of *X* (e.g., of personal identity [Engel 2011, 158], or they “elicit in us our intuitions about two different concepts” of *Y* (e.g., of knowledge [Engel 2011, 160]), by having us imagine limit cases that make the application of our “common concept” of *X* problematic, or that discriminate between alternative concepts of *Y*. This presupposes that thought experiments are either about concepts or about “real things”, not about both. Both Williamson’s and Engel’s views are plausible if the psychological notion of concept is involved. But is it? Did philosophers that meant their thought experiments to be about concepts, or to whom such a view was attributed, have the psychological notion in mind? If not, are there good reasons to take their thought experiments to be about concepts *rather than* about “real things”?

Traditionally, Wittgenstein’s thought experiments are regarded as typically about concepts. Engel agrees with the tradition [Engel 2011, 149]. However, is such a reading obligatory? Let us consider a representative example:

[...] what if they piled the timber in heaps of arbitrary, varying height and then sold it at a price proportionate to the area covered by the piles?

And what if they justified this with the words: “Of course, if you buy more timber, you must pay more”?

How could I shew them that—as I should say—you don’t really buy more wood if you buy a pile covering a bigger area?—I should, for instance, take a pile which was small by their ideas and, by laying the logs around, change it into a “big” one. This might convince them—but perhaps they would say: “Yes, now it’s a lot of wood and costs more”—and that would be the end of the matter. We should presumably say in this case: they simply do not mean the same by “a lot of wood” and “a little wood” as we do; and they have a quite different system of payment from us. [Wittgenstein 1978, I, §§ 149–150]

3. Frege’s notion of concept in [Frege 1891] exemplifies the non-psychological notion.

This has been taken to be about the concepts of selling and buying, measuring, “more of” and “less of”, etc. [e.g., Baker & Hacker 1985, 328–329], [Glock 1996, 49]. But why not about sales themselves—those very transactions? Start with the assumption that in sales, the amount paid is proportional to certain stable features of what is being sold, as measured by some standard (say, a scale in the case of wood). Now, we show the wood sellers that the area a pile of wood covers is not a stable feature of that pile of wood. The wood sellers may be implicitly committed to our conception of what a sale is, in which case they would retract their method of payment (“this might convince them”). However, they might also reject the stability requirement (“*Now* [my it.] it’s a lot of wood and costs more”), thereby showing that they disagree with us about what sales are. They entertain a partially different theory of sales.

One might object that this account clashes with Wittgenstein’s conclusive comment, “They simply do not mean the same by ‘a lot of wood’ and ‘a little wood’ as we do”. This seems to indicate that the thought experiment reveals *semantic* differences between the wood sellers and us, and semantic properties are Wittgenstein’s *explicatum* for conceptual features (a concept is just a word together with its “deep” semantic properties, i.e., norms governing the use of it). However, he also comments that the wood sellers “have a quite different system of payment from us”, i.e., a different institution of selling: their sales are not like our sales, though they are similar in certain respects. When seen in this light, the wood sellers thought experiment could be read as showing that *sales* are not necessarily what we had assumed them to be.

I cannot see any strong reason for seeing it one way or the other. Wittgenstein’s thought experiments have been interpreted (correctly, I believe) as part of an anti-essentialist strategy [see e.g., Stroud 1965]. They are aimed at showing that what we tend to see as metaphysical (“logical”) necessities are inextricably intertwined with practices, institutions, and (ultimately) facts of nature: if these were different we might think differently and have different practices and institutions [Wittgenstein 1953, II, xii]. Seeming necessities are ultimately contingent: they are motivated, though not necessitated, by the natural world being the way it is and by certain basic patterns of behavior and practices having become so entrenched that, for us, they resemble nature in being spontaneous, automatic, and unquestioned [cf. Wittgenstein 1967, §355]. This conception should not be seen as entailing naturalization, let alone psychologization of concepts, for two reasons. First of all, Wittgenstein is generally hostile to any attempt at reducing social or semantic facts to natural facts in the strict sense (e.g., to physical or biological facts). Secondly, though concepts are indeed *our* concepts, they are not mental abilities ultimately grounded in the biology of the human brain, but structural ingredients of our *Weltbild*: they determine how we happen to think and believe we cannot but think, not how we *must* think of natural necessity. So, even if one (legitimately) takes the wood sellers thought experiment, as well as others, to be about concepts, there is no reason to take “concept” in the psychological sense.

2 Thought experiment and counterfactual conditionals

Suppose we grant that, epistemologically, evaluating modal claims (such as “It is possible that A ”) is on a par with evaluating “counterfactual” claims of the form “If A were the case, then B would be the case”.⁴ Further suppose we agree that counterfactual reasoning is quite an ordinary business and that we often take its conclusions to be reliable. Does it follow that controversial modal claims, or philosophical thought experiments, are really less controversial than they have been taken to be? Concerning thought experiments in particular, how crucial is the role of counterfactual reasoning in arguments based on thought experiments?

2.0.1 Gettier’s thought experiment

Let me start with the second issue. Williamson provides a detailed analysis of one (and only one) kind of thought experiments, namely Gettier-like thought experiments. According to his analysis, such thought experiments are aimed to refute a *target claim*: the claim that necessarily, one knows that p if and only if one has a justified true belief that p .

$$(\mathbf{TC}) \quad \Box \forall x \forall p (K(x, p) \equiv JTB(x, p))$$

The refutation is effected by showing that there are possible cases of justified true belief that are not cases of knowledge:

$$(\mathbf{TEc}) \quad \Diamond \exists x \exists p (JTB(x, p) \ \& \ \sim K(x, p))$$

(**TEc**), in turn, is presented as the conclusion of an argument whose premises are (a) the possibility of Gettier cases, i.e., the claim that some Gettier stories might be true;⁵ (b) the necessary implication from there being Gettier cases to there being cases of justified true belief that are not cases of knowledge, i.e.,

$$(\mathbf{Imp}) \quad \Box \forall x \forall p (GC(x, p) \rightarrow (JTB(x, p) \ \& \ \sim K(x, p)))$$

From such premises, (**TEc**) follows by standard modal logic and the target claim is refuted.

4. I use (just this time) scare quotes around “counterfactual” to remind the reader that Williamson uses the word for conditionals with a subjunctive antecedent in general, whether or not they involve a presupposition that the antecedent is false. See [Williamson 2007, 137].

5. As Williamson sees it, a Gettier case should be described as a situation in which a true belief is validly derived from false premises, *not* as a case of justified true belief without knowledge.

This looks like a fair presentation of the argumentative structure of Gettier's criticism of the "classical" definition of knowledge. Is counterfactual reasoning involved? So far, not at all. Yet Williamson's conclusion of his account of thought experiments is that:

Paradigm thought experiments in philosophy [surely including Gettier's] are simply valid arguments about counterfactual possibilities. [Williamson 2007, 207]

So, how do counterfactual possibilities—as opposed to simple possibilities—get into the picture?

Williamson reasons as follows. The strict implication (**Imp**) is both unnecessarily strong and not "the best representation of the verdict on the Gettier case". We don't want (and don't need) *every* Gettier case to be a case of justified true belief without knowledge. There might be bizarre circumstances—possible worlds—in which the Gettier mechanism of inferring true beliefs from false premises doesn't work as expected; e.g., the inference itself might cast doubt on the premise, thereby preventing the true conclusion. We want to exclude such bizarre circumstances ("far away" worlds); this is achieved by weakening (**Imp**) to

(**Counter**) $\exists x \exists p GC(x, p) \Box \rightarrow \forall x \forall p (GC(x, p) \rightarrow (JTB(x, p) \& \sim K(x, p)))$

I.e., if there *were* Gettier cases, then Gettier cases would be cases of justified true belief without knowledge. This rules out bizarre circumstances: only "close enough" worlds matter.⁶ (**TEc**) still follows, assuming standard principles of counterfactual logic. Thus, a better reconstruction of Gettier-like arguments against (**TC**) gives pride of place to a counterfactual premise.

Is this enough to justify the emphasis on the role of counterfactual reasoning in thought experiments? After all, Williamson admits that the switch from (**Imp**) to (**Counter**) may be just the first step down a slippery slope of successive attempts at guaranteeing that Gettier cases are indeed cases of justified true belief without knowledge—something we cannot absolutely guarantee anyway [Williamson 2007, 204]. It is then hard to argue that the argumentative structure that relies on (**Counter**) is definitely better than the alternatives as an account of Gettier's argument, particularly, better than the original structure that relies on (**Imp**) and does not involve any counterfactuals. But then the motivation for regarding the argumentative import of Gettier's thought experiment as essentially relying on counterfactual reasoning is thin.

6. As Williamson points out, however, even this does not guarantee that every Gettier case will be a case of justified true belief without knowledge, e.g., if *the actual world* contains an abnormal Gettier case that is not a case of justified true belief without knowledge (as it is not a case of justified true belief to begin with) [Williamson 2007, 200].

2.0.2 Another thought experiment

As we saw, Williamson claims that “paradigm thought experiments” are about counterfactual possibilities. Let us see if his analysis extends to examples other than Gettier’s. As Engel points out, it is natural to conjecture that it should, as thought experiments “involve possibilities”—they are based on imagining possible situations—and “are about possibilities and necessities”, i.e., they refute allegedly necessary claims by exhibiting contradicting possibilities [Engel 2011, 157]. The first point is unassailable; the second, as we shall see, is not equally secure.

I will choose to consider Searle’s “Chinese room” thought experiment, which I take to be no less “paradigmatic” than Gettier’s, though more controversial and somewhat harder to analyze. Gettier’s argument is straightforward: the target claim is clear, the envisaged situations may be awkward but their possibility is rarely challenged (as both Williamson and Engel note, there are even *actual* Gettier cases) [Williamson 2007, 192], [Engel 2011, 154], and most people—or most Western philosophers—tend to agree that, given that Gettier cases are possible, the conclusion (**TEc**) follows and the target claim is refuted. With the Chinese room, things are messier. The thought experiment is meant to refute a thesis that Searle characterizes as the “strong AI” claim and formulates as follows:

Appropriately programmed computers [...] can be literally said to *understand* and have other cognitive states. [Searle 1980, 417]

We might then be tempted to attribute to Searle’s argument the same structure that appeared to fit Gettier’s:

- *S1. (TargetClaim)* Necessarily, appropriately programmed devices understand natural language.⁷
- *S2.* Necessarily, to be a programmed device is to be a symbol manipulating system.⁸
- *S3.* We can imagine a symbol manipulating system (the man in the room) that does not understand the natural language symbols it manipulates.
- *S4.* Hence, the “strong AI” claim *S1* is false.

However, this is not at all what Searle intends to show and believes his thought experiment to have shown. He doesn’t want to show that programmed devices *may or may not* understand language; he wants to show that *no* such device *could ever* understand language, for any such device is like the man in the room in every relevant respect (i.e., a mere symbol manipulator), and—obviously—the man in the room does not understand the language whose symbols he is manipulating. I.e., he wants to show that

(**ChR**) Necessarily, no programmed devices understand natural language.

7. I will adopt this simplification of the “strong AI” claim as in fact Searle only deals with language understanding, not with cognitive states in general.

8. This is to be read as a biconditional.

If (**TargetClaim**) is $\Box\forall x (PDx \rightarrow Ux)$, what Searle wants to prove is not its contradictory as in the proposed reconstruction, i.e.,

$$\Diamond\exists x(PDx \ \& \ \sim Ux)$$

but rather its contrary,

$$(\mathbf{ChR}) \quad \sim \Diamond\exists x(PDx \ \& \ Ux).$$

It is then natural to reinterpret the target claim—the thesis of “strong AI” that Searle is attacking—as directly contradicted by (**ChR**), i.e., as

$$(\mathbf{TargetClaim}^*) \quad \Diamond\exists x(PDx \ \& \ Ux)$$

which, by the way, much better suits actual views Searle may have had in mind.

But now, what is the overall structure of Searle’s argument? Even assuming that the Chinese room is a possible situation, how can he show that the man in the room is representative of every possible computing device? What he needs is an additional premise, namely that necessarily, no symbol manipulating device has cognitive powers that exceed the man in the room’s: if the man in the room doesn’t understand, then no symbol manipulating system does. The structure of Searle’s argument can then be reconstructed as follows:

- $S'1. \Diamond\exists x(PDx \ \& \ Ux)$ [**TargetClaim***]
- $S'2. \Box\forall x(PDx \rightarrow SMDx)$ [being a programmed device entails being a symbol manipulating system]
- $S'3. \Diamond\exists x((SMDx \ \& \ \sim Ux) \ \& \ \Box\forall y(SMDy \rightarrow (Uy \rightarrow Ux)))$
[Hypothesis]
- $S'4. \Diamond\exists x((SMDx \ \& \ \sim Ux) \ \& \ \Box\forall y(SMDy \rightarrow \sim Uy))$ [from $S'3$]
- $S'5. \Box\forall y(SMDy \rightarrow \sim Uy)$ [from $S'4$, assuming $S5$]
- $S'6. \sim \Diamond\exists x(PDx \ \& \ Ux)$ [from $S'2, S'5$].

This structure is quite different from the structure of Gettier’s argument. First, the target claim $S'1$ is not a necessitated conditional but a possible existentially quantified conjunction, so that what needs to be proved to refute it is the necessitated conditional $S'5$. Secondly, the premise $S'3$ does not just assert the “verdict” of the Chinese room thought experiment but the crucial paradigmatic character of the Chinese room system. The possibility of a symbol manipulating system that does not understand language does not by itself prove that no other such system could, unless it is also assumed that no symbol manipulating system can exceed the man in the room’s understanding abilities. Otherwise, it could be retorted that though the man in the room is indeed unable to understand, other, differently programmed devices might not share such inability. Indeed, this is what many partisans of AI objected to Searle in the original debate (see e.g., Dennett’s, Fodor’s, Lycan’s, and Pylyshyn’s contributions in [Searle 1980]).⁹

9. More precisely, these critics objected to either $S'2$ or $S'3$ in the argument above depending on whether they understood “symbol manipulating system” in a

Is there any room for counterfactual reasoning in Searle's argument so reconstructed? Paralleling Williamson's considerations, we might want to rule out bizarre worlds in which—say—the man in the room's physical manipulations of written scribbles systematically affect his mental states in such a way that he does understand the symbols he is handling. I.e., we might want to weaken the thought experiment's "verdict" (that does not explicitly appear in the above reconstruction) from

$$(\mathbf{ImpChR}) \quad \Box \forall x (CRx \rightarrow (SMDx \ \& \ \sim Ux))$$

To

$$(\mathbf{CounterChR}) \quad \exists x CRx \ \Box \rightarrow \forall x (CRx \rightarrow (SMDx \ \& \ \sim Ux))$$

i.e., "If there were Chinese rooms, then a Chinese room would be a symbol manipulating device that doesn't understand language". However, this would not affect the crucial, highly controversial assumption that no mere symbol manipulator could exceed the man in the room's understanding powers. We surely don't want to restrict the operation of mere symbol manipulators to normal, close-to-home circumstances: clearly, Searle wants his result to concern *any* (mere) symbol manipulator. Indeed, he comes very close to making the negative connection between being a symbol manipulator and understanding into an *a priori* necessary connection:

As soon as we knew that [a robot's] behavior was the result of a formal program [...] we would abandon the assumption of intentionality. [Searle 1980, 421]

None of this contradicts Williamson's claim that counterfactuals have a role to play in philosophical thought experiments. Even in the Chinese room case, the experiment's "verdict" can arguably take counterfactual form. However, as the Chinese room also shows, the verdict alone may not provide a refutation of the target thesis. To that effect, further assumptions may be needed that cannot be immediately drawn from examining the imagined situation (though they may be suggested or made vivid by it, as Searle clearly hoped they would). It seems that, at least in the Chinese room case, for such assumptions counterfactual weakening is not in order.

2.1 Counterfactual reasoning and the rejection of thought experiments

Gettier-like situations are clearly possible. Indeed, there are actual, hence possible Gettier cases. Other thought experiments confront us with more

narrow sense (in which case they rejected *S'2*, holding that not every programmed device must be a *mere* symbol manipulating system) or in a more inclusive sense (in which case they rejected the second conjunct of *S'3*, holding that not every symbol manipulating system must share the man in the room's limitations).

problematic situations. For example, could there be such a planet as Twin Earth? Since the beginning, it was pointed out that no substance molecularly very different from water could have every superficial property of water [Kuhn 1989, 27]. It could be added that Oscar₁ and Oscar₂ cannot be identical “molecule by molecule”, as Oscar₁ is about 65% H₂O whereas Oscar₂ is (in the same proportion) XYZ. Thus, the situation we are asked to imagine is scientifically impossible: no planet that conforms to the laws of nature could be like Twin Earth, but even if there were such a planet, it could not contain a molecular twin of a human being of Earth.

Similarly with the Chinese room. As far as we know, it is not possible that, “after a while”, a human being “gets so good at following the instructions for manipulating Chinese symbols and the programmers get so good at writing the instructions that [...] [the man’s] answers to the questions are absolutely indistinguishable from those of native Chinese speakers” [Searle 1980, 418]; see Richard Hofstadter’s objections in [Searle 1980, 433–434]. Remember, the man is supposed to identify strings of symbols and relate them to other strings by their form alone: the level of performance Searle imagines far exceeds the limitations of human memory. There can’t be any Chinese room that fits Searle’s description.

Prima facie, showing that a situation that a thought experiment asks us to imagine is physically or biologically impossible looks like a good reason to forget about the thought experiment, for the assumption that the imagined situation is a genuine possibility is bound to be crucial in any argument based on it. Why is it that, on the contrary, we tend to discount impossibility objections? An interesting question. Williamson raises a related issue: it may happen that the situation we are asked to imagine in a thought experiment just lacks the features that are required to reach the desired “verdict”, i.e., to function as a counterexample to the target claim. For example, it is a pseudo-Gettier case: not a case in which a true belief is validly derived from false premises. The reason of such inadequacy may be trivial, and the thought experiment easy to repair. In such cases, “the similarity of the new thought experiment to the old one is evidence that the old one was not far wrong” [Williamson 2007, 201]. Perhaps we reason in the same way with allegedly impossible situations, such as Twin Earth or the Chinese room: we feel that though the situations we are asked to imagine appear to be impossible, “something pretty much like that is going to work”. Well, is it? Perhaps in some cases, not necessarily in every case. With Twin Earth, perhaps aluminum and molybdenum will work.¹⁰ With the Chinese room, we seem to be forced to replace the man in the room with some superhuman being that would be,

10. Actually, there would be problems even there. Both aluminium and molybdenum are present in the human body in tiny proportions (aluminium, 0.000015%, molybdenum, $4.5 \times 10^{-8}\%$). However, molybdenum is essential to human life (aluminium is not), so if on Twin Earth molybdenum were everywhere replaced by aluminium, but not vice versa, there would be no Oscar₂. Notice that trouble is bound to arise in any case, for molybdenum’s essentiality for human life depends on

however, sufficiently like us to allow plausible conclusions about his/her not understanding Chinese. Ultimately, this amounts to a partial suspension of natural laws; but when natural laws are suspended, everything—or anyway, a lot—goes.¹¹ Hence, either we choose to appeal to a notion of possibility such that scientific impossibility does not entail impossibility *tout court*, or we should declare the Chinese room a failed thought experiment.

Be that as it may, when we reach the conclusion that a situation we are asked to imagine is impossible what we reach—if our argument is sound—is modal knowledge. Is it reached by way of counterfactual reasoning? It certainly could be so presented. For example, in the Twin Earth case we might reach the impossibility conclusion by evaluating the counterfactual conditional (**TwE**):

(**TwE**) If Twin Earth existed,¹² no human being of Earth would be molecularly identical to any human being of Twin Earth.

Alternatively, we could assume a description of Twin Earth (e.g., as per fn. 12) and, given some chemical and biological knowledge, derive a contradiction from the hypothesis that some human being of Earth is molecularly identical with some human being of Twin Earth. Assuming a certain amount of scientific knowledge has roughly the same effect as ruling out “bizarre” worlds, as we do by adopting the counterfactual formulation (**TwE**). Is there any epistemological advantage of counterfactual reasoning over standard deductive reasoning? As Williamson points out [Williamson 2007, 143], counterfactual reasoning uses imagination, constrained by perception and knowledge of how nature works; it may use folk physics, perhaps not expressible in propositional form but “stored in the form of some analogue mechanism” such as a neural network [Williamson 2007, 145]. Indeed. However, nothing of this seems relevant in the present case. It seems that, whether we are trying to establish the truth of (**TwE**) or to prove by deductive argument that the Twin Earth hypothesis is inconsistent with the properties attributed to the Oscars, we reason in essentially the same way, using the same bits of scientific knowledge. We make no use of perception or folk science, we do not execute any simulation, and even imagination appears to play, at most, an auxiliary, heuristic role. Thus, I can’t see any epistemological advantage of counterfactual reasoning in this case.

Of course, I am well aware that this conclusion does not generalize to all modal issues that arise in connection with thought experiments, let alone to all instances of modal knowledge. There may be thought experiments

its presence in certain compounds. Even if it could be replaced by aluminium, the effects would not be the same.

11. A point that was forcefully made by Hilary Putnam [Putnam 1990, 69–70] while reaching the conclusion that there are no necessary and sufficient conditions for being water in all possible worlds, *including* worlds that obey different natural laws.

12. This abbreviates the statement “If there were a planet that is exactly like the Earth, except that water is everywhere replaced by a superficially similar but molecularly quite different liquid”.

(Mary comes to mind) where imagination and simulation play a major role. Nevertheless, Williamson's own generalization that the evaluation of counterfactuals is crucial in arguments based on "paradigm" thought experiments seems to me ungrounded.

3 Conclusion

I have been trying to add substance to Pascal Engel's doubts about two claims of Timothy Williamson's in *The Philosophy of Philosophy* [Williamson 2007]: that philosophical thought experiments are about metaphysical possibility as opposed to conceptual possibility, and that counterfactual reasoning plays a crucial role in reaching modal conclusions relevant to arguments based on thought experiments. In the first case, I argued that unless the notion of conceptual is understood psychologically, it is hard to tell the difference between thought experiments' being "about concepts", or conceptual possibilities, and their being "about real things", or metaphysical possibilities. In the second case, I argued that it is far from clear that counterfactuals are indispensable premises in arguments based on thought experiments, and that it is not generally true that counterfactual reasoning has an epistemological edge over standard deductive reasoning in reaching modal conclusions (e.g., the conclusion that the situation a thought experiment asks us to imagine is impossible). While I believe Engel would agree with me concerning the role and import of counterfactuals, he might disagree about the possibility of a distinction between conceptual and metaphysical interpretations of thought experiments even where the notion of conceptual is not understood as psychological; this is for him to tell.

I realize that the article I chose as my point of departure is a tiny fraction of Pascal's immense philosophical production. Nevertheless, even here he fully displays the acumen, subtlety and critical spirit for which he is well known in the analytic community (and beyond). I am glad to have picked this particular intersection between his philosophical interests and mine (one among many) to celebrate his birthday.

Bibliography

- BAKER, Gordon P. & HACKER, P. M. S. [1985], *Wittgenstein: Rules, Grammar and Necessity: Volume 2 of an Analytical Commentary on the Philosophical Investigations, Essays and Exegesis*, Oxford: Blackwell.
- CHALMERS, David [2002], Does conceivability entail possibility?, dans *Conceivability and Possibility*, édité par T. Szabó Gendler & J. Hawthorne, Oxford: Clarendon Press, 145–200.

- ENGEL, Pascal [2011], Philosophical thought experiments: in or out of the armchair?, dans *Thought Experiments in Methodological and Historical Contexts*, édité par K. Ierodiakonou & S. Roux, Leiden: Brill, *History of science and medicine library*, t. 19, 143–164, doi: 10.1163/ej.9789004201767.i-233.27.
- FREGE, Gottlob [1891], *Funktion und Begriff*, Jena: H. Pohle.
- GLOCK, Hans-Johann [1996], *A Wittgenstein Dictionary*, Oxford: Blackwell.
- KUHN, Thomas [1989], Possible worlds in history of science, dans *Possible Worlds in Humanities, Arts and Sciences*, édité par A. Sture, Berlin: De Gruyter, 9–32.
- MARCONI, Diego [2009], Wittgenstein and necessary facts, dans *Wittgenstein: Mind, Meaning and Metaphilosophy*, édité par P. Frascaola, D. Marconi & A. Voltolini, London: Palgrave Macmillan, 140–166.
- [2015], Concepts: too heavy a burden, dans *Mind, Language and Action*, édité par S. Moyal-Sharrock, V. Munz & A. Coliva, Berlin; Munich; Boston: De Gruyter, 497–521.
- PUTNAM, Hilary [1990], Is water necessarily H₂O?, dans *Realism with a Human Face*, Cambridge, Ma: Harvard University Press, 54–79.
- RUSSELL, Bertrand [1921], *The Analysis of Mind*, London: Allen & Unwin.
- RYLE, Gilbert [1949], *The Concept of Mind*, London: Hutchinson.
- SCHLICK, Moritz [1936], Meaning and verification, *The Philosophical Review*, 45(4), 339–369, doi: 10.2307/2180487.
- SEARLE, John R. [1980], Minds, brains, and programs, *Behavioral and Brain Sciences*, 3(3), 417–424, doi: 10.1017/S0140525X00005756.
- STROUD, Barry [1965], Wittgenstein and logical necessity, *The Philosophical Review*, 74(4), 504–518, doi: 10.2307/2183126.
- SZABÓ GENDLER, Tamar & HAWTHORNE, John [2002], *Conceivability and Possibility*, Oxford: Clarendon Press.
- WILLIAMSON, Timothy [2007], *The Philosophy of Philosophy*, Malden: Blackwell, doi: 10.2307/21831262007.
- WITTGENSTEIN, Ludwig [1953], *Philosophische Untersuchungen*, Oxford: Blackwell, 4^e éd., Chichester: Wiley-Blackwell, 2009.
- [1967], *Zettel*, Oxford: Blackwell.
- [1974], *Philosophical Grammar*, Oxford: Blackwell.
- [1978], *Remarks on the Foundations of Mathematics*, Oxford: Blackwell, 3^e éd.