LABORATORIUM MENTIS | 2023, VOL. 1, NO. 1

DOI: 10.52097/lm.8149

PIOTR BIŁGORAJSKI
John Paul II Catholic University of Lublin (Poland)
0000-0001-5139-3455
piotr.bilgorajski@kul.pl

Are thought experiments a reliable method of doing philosophy?[1]

Received: 24.11.2023 / Revised: 04.12.2023 / Accepted: 12.12.2023 / Available: 20.12.2023

Abstract:
In the paper I defend the practice of using thought experiments against the claim that it is not a serious way of philosophical argumentation. At the heart of the criticism leveled against thought experiments is the assumption that the products of imagination, due to their lack of grounding in reality, are fundamentally unreliable. Assuming the existence of an analogy between thought experiments and real experiments, I point out that there are criteria that define the framework of a good thought experiment.

Key words:
thought experiments, metaphilosophy, imagination, mental simulations

How to cite:
Biłgorajski, P. (2023). Are thought experiments a reliable method of doing philosophy? [English translation]. *Laboratorium Mentis*, *1*(1), 22–32. https://doi.org/10.52097/lm.8149

[1] The work was supported by the National Science Centre, Poland, under research project "Thought Experiments in Philosophy: Origin, Structure, Functions," no. UMO-2017/25/N/HS1/03019.

Let us start with a story. Imagine a professor of philosophy who built a significant portion of her academic career on presenting thought experiments that set the tone for many philosophical discussions over the years. However, one day, a group of researchers publicly announced that despite many attempts, they failed to reproduce the results of the thought experiments proposed by this philosopher. Based on this, they conclude that these thought experiments were likely fabricated. The renowned philosopher, faced with overwhelming evidence, admits to the fraud and, amid scandal, leaves the university, never to engage in philosophy again.

Is such a story possible? One might suspect that many would consider it a joke. It is commonly deemed implausible to fabricate data resulting from the execution of a thought experiment, an activity conducted solely in the realm of imagination. Yet, I believe everyone would also agree that if a similar story involved a well-known physicist or biologist accused of fabricating the results of their experiments, but this time real ones, no one would consider it a good joke. On the contrary, we would witness widespread and entirely justified outrage.

I would like to seriously address the question of why the first story, concerning the fabrication of thought experiments, seems amusing, while cases of fabricating real experiments are less so. What is funny about it? An explanation might lie in the popular theory of humor, which suggests that humor arises from perceived inconsistency. In this context, the story seems funny because the concept of a fabricated thought experiment contains an internal contradiction. The inconsistency lies in the fact that products of imagination cannot be forged; hence, thought experiments, occurring solely within the imagination, cannot be "cheated". In thought experiments, anything is allowed.

Such a conclusion aligns with the views of critics of thought experiments, such as Kathleen Wilkes (2003), who compared thought experiments to fantasy stories. In her view, such narratives belong to fiction rather than scientific work. On the other hand, the practice of engaging in philosophy seems to suggest otherwise. Thought

experiments are widely used not only in original philosophical works but also as a tool for popularizing philosophy. Are philosophers using a tool which produces outcomes they consider unreliable?

I will attempt to defend the reliability of thought experiments. I will start by presenting popular typologies of thought experiments and then propose a definition of a thought experiment built on the principle of analogy to real experiments. I will argue that a fabricated thought experiment is a specific case of an unsuccessful experiment and will identify the criteria for an unsuccessful thought experiment. If indeed a thought experiment can fail, there must be criteria which determine when a thought experiment is successful.

Why do philosophers use thought experiments?

One of the earliest typologies of thought experiments comes from Karl Popper, who distinguished thought experiments created with heuristic, critical (destructive), and apologetic (constructive) intentions (Popper, 2002). Heuristic thought experiments present a certain theory in an appealing way, making it easier for the audience to grasp. Such thought experiments can serve as illustrations of established theories or simplify a presentation of a theory's results for popularization purposes. Critical experiments are devised either against a particular theory or to challenge the assumptions and conclusions of other thought experiments. Thought experiments in this function are often presented as counterexamples to some general assertion. Apologetic experiments provide examples that confirm a given theory (Popper, 2002, p. 243).

Popper's typology can be complemented by Tamara Gendler's proposal, which categorizes thought experiments into three categories: (i) factual, (ii) conceptual, and (iii) evaluative. Factual thought experiments are those present in empirical sciences. For example, in Galileo's thought experiment, one might ask what would happen if two stones of different masses were dropped together. Conceptual

thought experiments can be found in metaphysics and epistemology, serving to verify whether a given concept applies to a described state of affairs. For instance, whether the concept of knowledge applies to the so-called Gettier problem. Evaluative thought experiments appear in ethics and aesthetics. Here, the audience is confronted, for example, with the trolley dilemma, and their task is to make a moral judgment on the presented situation.

The above typologies indicate the functions of thought experiments. But how do thought experiments carry out these functions? Chris Daly suggests that thought experiments can function as (1) "triggers," (2) insights into the world of Platonic ideas, (3) arguments, (4) variations of real experiments, and (5) mental models (Daly, 2010).

Thomas Kuhn is associated with the concept of thought experiments as triggers. According to Kuhn, thought experiments help to fit available data into new conceptual schemes, facilitating the detection of accumulated contradictions and anomalies (Kuhn, 1977). James Brown represents the Platonic approach to thought experiments, suggesting that thought experiments, through some form of intuition, provide access to a Platonic realm of necessary truths (Brown, 1991).

John Norton (2004) argues that thought experiments do not fundamentally differ from arguments; their main function is persuasion. Similarly, Daniel Dennett (2013) refers to thought experiments as "intuition pumps". In this context, thought experiments serve solely as tools of persuasion. However, if this holds for other arguments, a good thought experiment would be a good argument, and a bad thought experiment would be a bad argument. On the other hand, Roy Sorensen (1997) claims that there should not be a need to add the qualifier "thought" to experiments because the experiments commonly regarded as "thought experiments" are so similar to "real" experiments that the distinction becomes negligible.

Timothy Williamson (2007) believes that an essential feature of philosophical thought experiments is their modal character. This aspect is emphasized in the concept of thought experiments treated

as mental models (Nersessian, 2018). In this approach, thought experiments present situations that correspond to or prompt responses to the question: "What if?". The imagined situation serves as a model, a representation of a possible state of affairs, and the thought experiment involves simulating the behavior of that state of affairs. According to Nancy Nersessian, the process of simulation occurs in three steps. The first stage involves constructing a mental model representing a selected aspect of reality. Then, specific manipulations are performed on the presented model. Finally, the results obtained from the manipulations are used to infer about the modeled aspect of reality.

The simulation theory assumes that imagination is similar to perception in a way, but although its purpose is to represent reality, it is not, unlike perception, "controlled" by reality. Thus, it can be argued that such a view lacks a criterion for distinguishing valuable imaginings from the products of pure fantasy. In response, proponents of the simulation theory emphasize that the mechanism of imagination when perceiving fiction is no different from how it is used in everyday situations. In imagination, we can create different scenarios and test various solutions without taking the effort and risk of implementing them in reality. For example, before I do something, I can imagine the possible consequences of an action and decide based on that. Such a controlled use of imagination is so common that some researchers indicate that the ability to create mental simulations has evolutionary justification (Williamson, 2016).

These concepts have in common the assumption that there are certain structural similarities between a thought experiment and a real (scientific, empirical) experiment. A real experiment is a procedure that involves influencing a certain state of affairs to observe what will happen with the aim of confirming or refuting a scientific hypothesis. Similarly, a thought experiment, like a real one, is conducted for cognitive purposes and involves intentionally changing a state of affairs. However, in a real experiment, the material undergoing change is empirical and factual (currently existing), while in a thought experiment, it

is imaginative and counterfactual. In other words, in a real experiment, bringing about a certain state of affairs is equivalent to imagining the occurrence of that state of affairs in a thought experiment, and the counterpart of observing the result of an experiment is appropriate reasoning taking place in a "laboratory of the mind".

Do thought experiments deserve to be called experiments?

Roy Sorensen believes that what thought experiments and real ones have in common is "tinkering." This means that designing both real and thought experiments involves creating specific conditions, considering only the essential factors for the procedure (so-called ceteris paribus conditions). And just as the usual goal of an experiment is to reveal some anomaly, that is, a phenomenon that does not submit to explanation by the tested theory, thought experiments most often provide counterexamples to certain philosophical concepts.

Sorensen notes that thought and real experiments function similarly as reference points in ongoing discussions. Well-known thought experiments, such as the Gettier problem or trolley dilemmas, become the standard method for conducting analyses (in these cases, analyses concerning the concept of knowledge or the scope of applicability of certain ethical theories). Thought experiments are also subject to criticism and correction. The dynamics of disputes in philosophy show that the most common response to a proposed thought experiment is some counter-thought experiment.

Sorensen also points out differences between real and thought experiments, but in his opinion, they are not significant enough to nullify the similarities. Sorensen discusses some obvious characteristics of real experiments—like the fact they are usually carried out by research teams, where individuals responsible for designing the experiment differ from those executing them. In the case of philosophical experiments, there is no division of labor into design and execution stages, which

might be seen as an advantage. It also seems that results in thought experiments are not obtained randomly, as in some physics experiments. It would be difficult to expect the outcome of a thought experiment to be surprising to the person conducting it. However, thought experiments would be much less susceptible to chance events, such as equipment failure. For obvious reasons, the thought experimenter has greater control over the course of their reasoning.

Thought experiments—unlike real experiments—also do not require expensive and complicated research equipment. The philosophical equivalent of a physical laboratory could be the philosopher's mind, and the quality of such a "laboratory" would depend on the appropriate level of education and intelligence of the person conducting the thought experiment.

When can a thought experiment fail?

A thought experiment begins by presenting a possible, fictional situation and, if done correctly, will lead the recipient to a specific conclusion. A thought experiment proceeds in three stages: (1) presenting an imaginative (possible) situation ("Imagine a woman named Mary who does not know colors but knows all about the physics of colors..."), (2) the presented situation has a narrative character ("Mary sees a red rose and learns something new about the world..."), (3) the result of the presented narrative confirms or refutes a philosophical thesis ("So... Physicalism is false!"). This structured view of a thought experiment allows us to indicate how it can fail:

1. Unimaginability: A thought experiment can be challenged by pointing out that the situation presented is inconceivable (because it is inconsistent or described too generally).

2. Inconclusiveness: Even if the situation is imaginable, there are no good reasons to accept its outcome.

3. Lack of reference to the actual world: Even if the situation is imaginable, and there are good reasons to accept its outcome, it does not provide a basis for a claim about our world (Gendler, 2000, p. 22).

In the first sense, a thought experiment fails if the depicted state of affairs is unimaginable. The argument from unimaginability or inconceivability is a common criticism of some particularly extravagant thought experiments, such as those concerning the possibility of philosophical zombies (Chalmers, 1997). Indicating that zombies, beings physically identical to me but devoid of a first-person point of view, are inconceivable is a popular way of weakening such an argument. Therefore, if philosophical zombies are inconceivable, such a thought experiment could be considered unsuccessful.

In the second sense, an unsuccessful thought experiment would involve the experimenter being able to imagine the situation but inaccurately envisaging its course. In the well-known thought experiment by Frank Jackson in which, upon seeing a red rose for the first time, Mary learns something new about the world, leads to the conclusion that physicalism is false (Jackson, 1986). However, critics of this thought experiment might argue that the course of this experiment should be different. For instance, they might argue that, upon seeing the red rose, Mary exclaims that the rose looks exactly the way she thought it would—after all, Mary has all the knowledge about the physics of colors, and the color of the rose should not be new or surprising to her. A similar situation could occur in a real experiment if, for example, the experiment proceeded without disruptions but generated incorrect data (due to equipment damage or an error in the experiment's design).

An unsuccessful thought experiment in the third sense is one leading to a correct conclusion but failing to provide an answer to the question that prompted its conduct. In a real experiment, this might be a situation where the experiment yields correct results but fails to provide the data searched for by researchers. An example of an unsuccessful thought experiment might be Gottfried Leibniz's "Mill." Leibniz, formulating an argument against the idea that the human mind has a mechanical

nature, invites the reader to imagine the interior of a mechanical mind, where, just like in a mill, we would not be able to observe any mental phenomena but only "parts pushing one another, and never anything which would explain a perception" (Leibniz, 1720/2014, p. 17). We can imagine a mill, and we can agree with Leibniz that we won't see anything which would explain perceptions in it, whereas if we reject the mill-mind analogy, we will not agree with Leibniz's conclusion that mechanicism is false.

Conclusion

I started the paper with a story and the question of what is funny about it. If we consider thought experiments a reliable method of doing philosophy, the answer is nothing. The humor comes from adopting a particular conception of thought experiments, according to which thought experiments are a procedure in which everything is allowed. However, this conception seems inaccurate since, at each stage of a thought experiment, we can ask whether it was conducted correctly.

Roy Sorensen compared thought experiments to a compass (1992, p. 288). A compass is a simple, albeit useful tool for indicating direction. However, it is not a reliable device—for example, compass indications are unreliable around the North Pole. Few people know how the compass really works, although this is not an obstacle to the effective use of the device. Similarly, philosophical thought experiments conducted in imagination point to possible states of affairs. However, the ease of creating thought experiments—after all, telling "what if..." stories does not require special technical skills—does not translate into the reliability of the results obtained in this way. Therefore, being aware of the limitations of the products of imagination allows us to use thought experiments with more caution.

Thought experiments can resemble a compass in some respects and a magnifying glass in others. The fantastic stories given by philosophers are to meticulously test philosophical theories to see if the explanatory

power of these theories covers all possible situations. Suppose someone claims that knowledge is a true and justified belief. In that case, this view can be undermined by giving an example of someone having knowledge despite not meeting all the conditions given in the definition of knowledge. If someone claims that physicalism is true, so that everything, whatever exists, can be described in physical terms, this view can be challenged by giving an example of an object that cannot be described in physical terms. To provide a counter-example in philosophy is to present some imagined possible situation, that is, to propose a thought experiment.

Bibliography

Brown, J. (1991). *The Laboratory of Mind: Thought Experiments in the Natural Sciences*. London-New York: Routledge.

Chalmers, D. (1997). *The Conscious Mind*. Oxford University Press.

Daly, C. (2010). *Introduction to Philosophical Methods*. Peterborough: Broadview Press.

Dennett, D. (2013). *Intuition Pumps and Other Tools for Thinking*. New York: W.W. Norton & Company.

Gendler, T. (2000). *Thought Experiment: On the Power and Limits of Imaginary Cases*. New York: Garland Publishing.

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy, 83*(5), 291–295.

Kuhn, T. (1977). *The Essential Tension. Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Leibniz, G. (2014). In L. Strickland (Ed.), *Leibniz's Monadology: A New Translation and Guide*. Edinburgh, UK: Edinburgh University Press. Original work published in 1720.

Nersessian, N. (2018). Cognitive science, mental modeling, and thought experiments. In M. Stuart, Y. Fehige, J. Brown (Eds.), *The Routledge Companion to Thought Experiments* (pp. 309–326). London: Routledge.

Norton, J. (2004). On Thought Experiments: Is There More to the Argument?. *Philosophy of Science*, *71*(5), 1139–1151. https://doi.org/10.1086/425238

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

Wilkes, K. (2003). *Real People: Personal Identity without Thought Experiments*. Oxford: Clarendon Press.

Williamson, T. (2007). *The Philosophy of Philosophy*. Malden: Blackwell.

Williamson, T. (2016). Knowing by Imagining. In A. Kind, P. Kung (Eds.), *Knowledge Through Imagination* (pp. 113–123). Oxford: Oxford University Press.