

## Contents

### Research Articles

Michael Agerbo Mørch – Atle Ottesen Søvik: <i>A Systematic Account of the Argumentative Role of Thought Experiments</i> .....	2
William Hunt: <i>Free Will: A consensus gentium Argument</i> .....	22
Miguel López-Astorga: <i>Chrysippus' Conditional Captured from a Non-Axiomatic Computer Program</i> .....	48
Vladimir Vujošević: <i>The Alethic Status of Contradictions in Fictional Discourse</i> .....	60
Brian Garrett – Jeremiah Joven Joaquin: <i>Lampert on the Fixity of the Past</i> .....	90

## A Systematic Account of the Argumentative Role of Thought Experiments

Michael Agerbo Mørch\* – Atle Ottesen Søvik\*\*

Received: 10 October 2023 / Revised: 10 November 2023 / Accepted: 13 November 2023

*Abstract:* What is the role of thought experiments in scientific exploration? Can they provide us with new knowledge about the world? In a recent article, Lorenzo Sartori argues that thought experiments function like ordinary (material) experiments: Both material experiments and thought experiments are made in a specific context, which must then be extrapolated and generalized to say something true about the world. This article discusses and criticizes Sartori’s proposal. It suggests a new theoretical framework for understanding thought experiments, their argumentative role, and how they provide new knowledge about the world. The framework presented is a coherentist framework, where coherence has three aspects: consistency, cohesiveness, and comprehensiveness. The proposal is that the argumentative role of thought experiments is to demonstrate the presence or absence of consistency, cohesiveness, and comprehensiveness, thereby strengthening a theory, weakening a theory, or showing one theory to be better than another. This is the way thought experiments

---

\* Fjellhaug International University College


 <https://orcid.org/0000-0002-5712-208X>

 Leifsgade 33.6, 2300 Copenhagen S, Denmark

 [mam@dbi.edu](mailto:mam@dbi.edu)

\*\* Norwegian School of Theology, Religion and Society

 <https://orcid.org/0000-0002-8616-7105>

 MF Norwegian School of Theology, P.O. Box 5144 Majorstua, 0302 Oslo, Norway

 [Atle.O.Sovik@mf.no](mailto:Atle.O.Sovik@mf.no)



provides new knowledge about the world, since the way we learn something new about the world is by discovering which theories about the world are most coherent.

*Keywords:* Coherence theory; Lorenzo Sartori; Scientific epistemology; Thought experiments.

## 1. Introduction

A thought experiment is an imagined scenario, often presented in the form of a narrative, conducted in mind with functions similar to scientific experiments but without collecting new empirical data from the world. But what is the role of thought experiments in scientific exploration? Can they provide us with new knowledge about the world? These are old and fascinating questions that are still discussed in the philosophical literature.

In a recent article, Lorenzo Sartori (Sartori 2023) argues that the discussion on this topic lacks an overarching theoretical framework. He provides a useful classification of positions but claims that no one has succeeded in giving a clear answer on how thought experiments provide us with new knowledge about the world. He then presents his own theory, which is that thought experiments function like ordinary (material) experiments, which we can see by distinguishing between internal and external validity. Both material experiments and thought experiments are made in a specific context, which must then be extrapolated and generalized to say something true about the world.

In this article, we first present an overview of the debate and Sartori's position before criticizing it. We propose an alternative theoretical framework for understanding thought experiments, their argumentative role, and how they provide new knowledge about the world. The framework we present is a coherentist framework, where coherence has three aspects: consistency, cohesiveness, and comprehensiveness. The proposal is that the argumentative role of thought experiments is to demonstrate the presence or absence of consistency, cohesiveness, and comprehensiveness, thereby strengthening a theory, weakening a theory, or showing one theory to be better than another. We argue that this is the way thought experiments provides new knowledge about the world, since the way we learn something

new about the world is by discovering which theories (broadly understood) about the world are most coherent.

## 2. An overview of the debate and Sartori's position

Lorenzo Sartori provides a useful overview of the philosophical debate on thought experiments, based on the following question from Thomas Kuhn: Do thought experiments provide us with new knowledge about the empirical world? If so, how do they do that when no observation is involved? If not, why not? (Sartori 2023, 2; Kuhn 1977).

Sartori uses Kuhn's question to categorize various positions into a yes-camp and a no-camp. Many answer yes because thought experiments have been so important in the history of science, for example, with Galileo, Newton, Einstein, and others. (Sartori 2023, 2) Sartori presents three different answers to how thought experiments provide us with new knowledge – Platonism, objectualism, and structuralism (Sartori 2023, 3).

Platonism is represented by James Brown (Brown 2004). He argues that thought experiments allow us to “see” abstract laws and structures that apply to the world through a priori intuitions. Objectualism is represented by Tamar Gendler and Nenad Miscevic (Gendler 2004; Miscevic 1992). They envision thought experiments as objects or images that allow us to see the world in a different way than through propositions. Structuralism is represented by Nancy Nersessian (Nersessian 1992). Her view is that thought experiments are a type of simulative model-based reasoning that reveals structural analogs to reality (Sartori 2023, 3).

Sartori raises objections to the three positions. Platonism is mysterious in its answer to how thought experiments work. Objectualism fails to explain why we gain new or different knowledge by thinking about objects instead of propositions. The problem with structuralism is that structural analogies can be either wrong or right. But then it seems that thought experiments do not help unless you already know the structures of reality (Sartori 2023, 3-4).

Other philosophers have argued that thought experiments do not provide us with new knowledge about the world (the no-camp). According to Daniel Dennett, thought experiments do not teach us anything new about

the world but rather stimulate our intuition (“intuition pumps”) (Dennett 1996). Ian Hawking believes that thought experiments can reveal inconsistency but do not tell us anything new about the world (Hacking 1993; Sartori 2023, 4). But, one could ask, why have thought experiments been so important in the history of science if they do not teach us anything new about the world?

A more detailed answer in the no-camp regarding why thought experiments do not teach us anything new comes from John Norton. He says we acquire knowledge through observation and logic, and since thought experiments do not provide us with new observations, their contribution must be logical. Thought experiments are like pictorially presented arguments. We gain *knowledge* from thought experiments because they are arguments with empirical knowledge embedded in the premises, but it is not *new* knowledge because the knowledge was already implied in the premises (Sartori 2023, 5; Norton 2004).

Sartori objects to Norton that some thought experiments have non-empirical and even impossible premises (such as running as fast as light, riding in an elevator with no gravity, etc.) (Sartori 2023, 5). Another position Sartori discusses is that of Rawad El Skaf (El Skaf 2018). He builds on Hacking but says that thought experiments reveal inconsistencies within or between theories (Sartori 2023, 5). Against this, Sartori argues that not all thought experiments are about inconsistencies. Examples of thought experiments that are not are Maxwell's demon, Newton's rotating spheres in an empty universe, or Einstein's elevator. Thought experiments like these seem to say something not only about old theories but also something new about the world (Sartori 2023, 6).

Sartori's overview is a helpful systematization, where one could easily insert other theorists based on whether they believe thought experiments provide us with new knowledge about the world and what role they believe thought experiments play in science. For example, while many have appealed to intuition as evidence for learning something new about the world, others reject it (Cappelen 2012).

After Sartori's presentation of various alternatives, his summary is that the yes-side provides vague answers, while the no-side does not explain the significance and success of thought experiments (Sartori 2023, 6). He believes

much of the disagreement is due to lack of a common theoretical framework (Sartori 2023, 7). Sartori then presents his own proposal, which is to think of thought experiments as ordinary experiments in science (material experiments) but to distinguish between internal and external validity (Sartori 2023, 7).

The distinction between internal and external validity comes from Donald Campbell (Campbell 1957) and deals with material experiments (Sartori 2023, 8). Internal validity is about whether the specific experiment in the specific setting is correct, while external validity is about whether one correctly extrapolates the results outside the specific setting of the experiment. It is important to distinguish between these two forms of validity since specific experiments are conducted under a set of assumptions that do not necessarily apply to other contexts (Sartori 2023, 8).

Sartori then applies this distinction to thought experiments. Internal validity for thought experiments is to understand thought experiments as a “game of make-believe” (Walton 1990), while external validity is to interpret thought experiments as an accurate representation of the world (Sartori 2023, 9 and 16). As an example, one can think of Galileo, Newton, and Einstein first conducting a thought experiment in their minds, describing a specific context and specific assumptions, and then deducing a general statement for all contexts from it (Sartori 2023, 9-11). When generalizing from both material experiments and thought experiments, one must make a series of assumptions. This process is the same in both material experiments and thought experiments and is best understood as a transition from internal to external validity (Sartori 2023, 11-12).

Regarding the external validity of thought experiments, it means asking whether the result of a thought experiment provides a true representation of the world (Sartori 2023, 19). Then one must check if the representation of the world is actually correct, but this is the same in material experiments as well (Sartori 2023, 23-25). The way thought experiments tell us something true about the world is then similar to material experiments (Sartori 2023, 27).

According to Sartori, not all thought experiments fit perfectly into this model, for example, if a thought experiment only points out an inconsistency in another theory (Sartori 2023, 25). Sartori has no recipe for

determining the external validity of a thought experiment (Sartori 2023, 25), but he believes there are no universal criteria for it (Sartori 2023, 26).

In contrast to Sartori, we believe there are some universal criteria that can and should be used when establishing the validity of thought experiments. The coherence criterion with its various aspects can be used to explain how we establish validity and how thought experiments work. It provides an alternative answer compared to Sartori regarding how thought experiments teach us something new about the world. Thought experiments can have both a destructive and a constructive function by weakening some theories and strengthening others. They provide new data even if these are not observational data, and they clarify connections or lack of connections in our theories of the world, thereby teaching us which understandings of the world are most likely true.

A central insight from coherence theory is that the way we learn something new about the world is by discovering which theories about the world are most coherent. Observations are just one of many types of data that we must combine in the most coherent way possible to discover how the world is. These claims will be further explained and defended in the next section.

### 3. An alternative understanding of thought experiments

In 1973, Nicholas Rescher defined the concept of coherence as having three aspects: consistency, cohesiveness, and comprehensiveness (Rescher 1973, 169). *Consistency* means that the elements of a theory cannot contradict each other. *Cohesiveness* means that the elements of the theory are connected. The more connections and the more precise and fine-grained they are, the more cohesive the theory. Connections should be thought of as including any kind of connection (spatial, temporal, logical, causal, etc.): describing relations between elements in a theory makes it more cohesive. *Comprehensiveness* is a measure of how many elements a theory manages to integrate consistently. The ideal is an integration of an optimal number of relevant elements.

The previous paragraph speaks of coherence between the *elements* of a theory. More precisely, the elements of a theory are *data*, where the concept of data is understood broadly to include any truth candidate, i.e., anything

somebody has reason to hold as true (Rescher 1973, 39-40), including scientific laws.<sup>1</sup> There are many advantages to having a broad definition of data, as opposed to a narrow understanding of data as, for example, strictly empirical observations. It clarifies the relation between data and theory, since all data are interpreted, and might be interpreted differently in the future in light of new theories. There are in fact many elements of theories that are not empirical observations, and many empirical observations can be interpreted in different ways that are not consistent with each other, such as the different interpretations of quantum mechanics. This broad understanding of data makes good sense of actual scientific praxis.<sup>2</sup>

All of our experiences with the world are interpretations of how the world is and happens in our mind. There is no experience and no access to the world that is not given to us as content in our minds. If you say to someone “Do not tell me how you think the world is, but how it actually is,” this is an impossible order, since nobody can say other than how they think the world is (Rescher, 2010, p. 5). Our understanding of all situations is interpreted and can be thought of as theories about the situation in a broad sense of theory, where a theory is meant to be a true understanding of how things are related. Our understanding of the world is constantly revised in light of observation and thinking. To learn something new about the world, means that new understanding of the world in your mind has replaced an old understanding.

We now proceed to present a theoretical framework for understanding the argumentative role of thought experiments. While Sartori focuses on thought experiments in natural science, our account is meant to cover both natural sciences and the humanities. From now on, the term “science” is used broadly to include the humanities. We focus on the argumentative function of thought experiments and how they can teach us something new,

---

<sup>1</sup> This means that no data are “raw.” All data is interpreted, and laws are also data, because they are fallible truth candidates. But in a coherentist understanding, the data are placed in a theoretical framework, which means that they are related to each other, so that the theory has both data and structure. The theory is expressed in language, and there are rules for how things should be related in the theory.

<sup>2</sup> For a more extensive discussion of this notion of data, see Puntel (2008, 11 et passim).



but we acknowledge that thought experiments can have many functions beyond that, for example illustrative, pedagogical, or heuristic functions (Cohnitz 2000; Corcilius 2018, 69).

Scientific work progresses by *strengthening* theories, *weakening* theories, or *comparing* theories to show that one is better than another. One can *strengthen* a theory by demonstrating or increasing the *presence* of consistency, cohesiveness, or comprehensiveness. One can *weaken* a theory by demonstrating *absence* of consistency, cohesiveness, or comprehensiveness. One can *compare* two theories by showing that one is *more* consistent, cohesive and/or comprehensive than the other.

We argue that all thought experiments used in science have the function of demonstrating either the presence or absence or a relatively better score of consistency, cohesiveness, or comprehensiveness. In the following, we substantiate this claim by testing it with regard to some examples. We comment on how the examples fit the theory by showing that they are examples of goals 1, 2, 3 (strengthening, weakening, or comparing) or means A, B, C (consistency, cohesiveness, or comprehensiveness) in our theory.

Galileo made a famous thought experiment to show that bodies fall to the ground at the same speed regardless of their weight (unless hindered by other forces such as air resistance). Aristotle had claimed that a heavier object will fall faster than a lighter object, but Galileo then suggested the following thought experiment: Imagine that we combine a heavy object A with a light object B and drop the combined object to the ground. Now the lighter object B should make the heavier object A fall more slowly if Aristotle is right. But the combined objects A+B can also be seen as one heavier object C, which should now fall faster than both A and B. Aristotle's theory implies that A should fall both faster and slower in this scenario, which is inconsistent (Palmieri 2018; Brown 1991, 1-3).

In this thought experiment, we see how Galileo demonstrates the presence of an inconsistency in Aristotle's theory. Galileo's alternative theory—that objects fall at the same speed regardless of their weight—does not have this inconsistency. In comparison, then, Galileo's theory is more coherent than Aristotle's, and thus a better theory. Galileo compares Aristotle's theory with his own (cf. goal 3) by means of demonstrating the presence of inconsistency in Aristotle's theory and consistency in his own theory (means

A). He thus weakens Aristotle's theory (goal 2) and shows his own theory to be better in comparison (goal 3).

Galileo is famous for introducing Galilean relativity, which says that the laws of nature are the same for all observers regardless of whether they are standing still or moving at uniform speed. He defended the heliocentric worldview, but understandably people had problems believing that the earth should be moving through space at a very high speed. After all, we experience standing still and seeing the sun move—would we not have noticed if we were moving at more than 100 000 km/h?

Galileo answers with another thought experiment: Imagine sitting below deck in a boat with the curtains pulled. In this scenario you would not know whether you were moving at a uniform speed or sitting still in still water. We can conclude from the thought experiment that if the earth moves at a uniform speed, we will not notice the difference between standing still and moving at high speed. If, in addition, the earth rotates around itself, it will seem like the sun is rising and setting.

This thought experiment has the function of demonstrating the cohesiveness of a theory (goal 1, means B). The theory that the earth orbits the sun seems unable to explain several data, like our experience of standing still and watching the sun move. Galileo uses the thought experiment of the boat to demonstrate how these data are nevertheless coherently connected since we would not notice any difference between the earth standing still or the earth moving at uniform speed.

Galilean relativity seems to imply that there is no objective answer to who is moving and who is standing still. Newton famously disagreed, arguing that there is an absolute space making it true that some objects are actually standing still while others are moving. He introduced the famous thought experiment of the bucket to argue this point. Imagine a bucket of water, hanging by a twisted cord, and then released. First the surface of the water is flat, but when the bucket starts spinning, the surface of the water turns concave in shape. Even if the water is immobile relative to the spinning bucket, we know from the shape of the water that the bucket is in fact spinning and not hanging still. According to Newton, this cannot be explained if motion and immobility are considered relative matters. Instead, we need the concept of an absolute space to explain the difference between

the two scenarios of bucket spinning and bucket hanging still (Brown 1991, 8-10).

With this thought experiment, Newton introduces a datum to a specific discussion and argues that it can be explained by his own theory, but not by Galilean relativity. By means of showing his own theory more comprehensive (means C), he tries to weaken Galilean relativity (goal 2) and strengthen his own theory (goal 1) to show it to be comparably better (goal 3).

Note the broad use of the concept of data. Data are truth candidates (Rescher 1973, 39-40). When scientists make an observation in a traditional scientific experiment, the data are interpreted (e.g., that the dots on the screen are in fact Higgs' boson). They are thus truth candidates and can be wrong. Thought experiments also deliver truth candidates that can be wrong (e.g., that there could be a zombie like humans in all respects, but without consciousness). Some data from thought experiments are new in the sense of being truth candidates nobody has thought about, like philosophical zombies, twin earths, etc. Other data from thought experiments are based on empirical data that are not new (like Newton's bucket), but they are used in a new context where they are relevant in deciding what is most coherent. In searching for the truth, researchers must take data (in the sense of truth candidates) and combine them as coherently as possible, and some of the data must then also be rejected as false.

Einstein later expanded Galilean relativity into his own theory of special relativity. This theory is based on two fundamental principles. The first is the principle of Galilean relativity, that the laws of nature are the same for all observers in uniform motion. The second and new principle is that all observers measure the same speed of light in a vacuum. According to Einstein, he was led to this insight in his youth, pondering various thought experiments of himself moving at the speed of light.<sup>3</sup>

Einstein imagines sitting on a train at the speed of light, looking into a mirror. Would he see nothing in the mirror? That would contradict Galilean relativity, which says that you cannot know whether you are standing still or moving from data inside your own frame of reference. Light should

---

<sup>3</sup> It is not important for our purposes what historically preceded what—we are only interested in the argumentative function of Einstein's thought experiments.

instead be measured as moving at the same speed regardless of your motion relative to light. From the insight that everyone measures the same speed of light, Einstein drew the consequences that measurements of time, distances and simultaneity are relative, as shown by various thought experiments involving light, trains and embankments.

Here is a thought experiment providing us with a new datum (truth candidate: light speed is the same for all), from which further new data can be deduced by means of thought experiments—for example that simultaneity is relative. To integrate these new data, Einstein needed to develop more concepts for describing relations more precisely, such as distinguishing between proper time and coordinate time or between rest mass and relativistic mass. This makes the theory more cohesive by showing more precise connections between the data, which then strengthens the theory.

Of course, Einstein's theory of relativity has been confirmed by empirical observations and would have been weaker without those observations. Hypothetically, observations could be made that would contradict the theory, but possibly the theory could also be adjusted to fit new observations. The point is that both thought experiments and empirical experiments can strengthen and weaken theories and are open to different interpretations.

We find that all common examples of thought experiments are easy to fit into our model. So far, we have focused on natural science, but in what follows we add some more examples for support, many coming from other disciplines than natural science, since our theoretical framework is meant to work for thought experiments in all disciplines of natural science and the humanities. The examples are categorized as examples of the means of consistency, cohesiveness, and comprehensiveness.

We start with consistency. Many thought experiments are created to show that a theory is inconsistent, thus *weakening* the theory. Since consistency is an either/or issue, if the thought experiment is successful the theory (in its present form) is destroyed, but it may be rescued later by introducing new distinctions or clarifications. Unless such repairs are ad hoc, the thought experiment which first points out inconsistency can help to improve the theory by making it more cohesive.<sup>4</sup> Here are some examples.

---

<sup>4</sup> An ad-hoc repair means adding a claim where the only reason for believing the claim to be true is that it would solve the problem. The repair is not ad-hoc if we

Bertrand paradoxes, such as first presented by Joseph Bertrand in *Calcul des probabilités* from 1889, suggest that all understandings of probability are inconsistent. Here is an example offered later by Bas van Fraassen: A factory produces cubes with side lengths between 0 and 1 meter. The probability that a randomly selected cube should have a side length of less than  $\frac{1}{2}$  meter seems to be  $\frac{1}{2}$ . But the probability that a randomly selected cube should have a face area of less than  $\frac{1}{4}$  square meter seems to be  $\frac{1}{4}$ . The problem is that we then get two different probabilities describing the same event, since a cube with a side length of  $\frac{1}{2}$  meter also has a face area of  $\frac{1}{4}$  square meters (van Fraassen 1989, 303).<sup>5</sup> When a thought experiment thus points to inconsistency in all theories, the thought experiment can be understood as a new datum that a new or any theory must integrate. In this case, the truth candidate is that all theories of probability are inconsistent, and thus a coherent theory of probability must be able to reinterpret Bertrand paradoxes or show why they are wrong and can be discarded.

Sometimes a thought experiment is created to defend a theory against the critique of inconsistency. The thought experiment can then support the view that the theory is consistent after all. One example is from the philosophy of time, in which different views are presented in modern philosophy. The Platonic view says that time itself can move even though everything else in the universe stands still, while the Aristotelian view says that if everything else in the universe stands still, time stands still too. The critic of the Platonic view challenges the Platonists to explain how it could make sense to imagine that time moves even though everything else stands still. Sydney Shoemaker took on the challenge of demonstrating how the Platonic view could be consistent: Imagine people living in three zones—A, B and C—where each of the zones sometimes experiences a local freeze—everything stops moving for an hour. This happens every other year in A, every three years in B, and every five years in C. For the people who experience the freeze, it just feels like going from one second to the next,

---

have other reasons to believe that the claim is true. This means that the coherence is very low in ad-hoc repairs, and that is why they are not a good thing.

<sup>5</sup> Van Fraassen uses 2 cm cubes, but we found the example easier to understand using 1 meter.

but after every freeze period, there is a red glow on things for a short while. The people in the different zones know about the freezes in the other zones. The inhabitants realize that every thirty years, all three zones should experience a freeze at the same time, and they do experience the usual red glow at all places. They conclude that they have probably had an hour of global freeze, meaning that one hour has passed, even if nothing has moved (Shoemaker 1969).

A unique type of demonstration of consistency is to show that the alternative is inconsistent such that the theory is necessarily correct. Sometimes this consistency can be proved by a thought experiment. The most well-known example stems from Descartes, who describes the possibility that an evil demon deceives our perceptions. But the demon cannot deceive us when it comes to the question whether we think, since we need thought in order to be deceived. You cannot be inconsistent in thinking that thoughts exist, since even being wrong requires that thoughts exist. In conclusion, we can know for sure that thoughts exist (Descartes 1641/1986, 12-15).

With these examples concerning consistency, we now proceed to the second aspect of coherence—cohesiveness. To recapitulate, cohesiveness refers to the connections between the data in a theory. The more connections, the better, since connections increase the plausibility that the data are relevant and needed in a theory. Thought experiments can be created to show a lack of relevant connections or clarify existing connections in a theory. In the following, we look at some examples.

In *Reasons and Persons* from 1984, Derek Parfit discusses the condition for personal identity over time. Is it physical continuity or is it psychological connectedness and continuity, or maybe different combinations of these? Parfit creates some thought experiments connected to teleporting and to a possible split between brain and body halves (Parfit 1984, ch. 10). Take the latter first: Imagine that you are in an accident. You are heavily injured, but the doctors manage to save half your brain and half your body. You have a lot of memories in the remaining part of the brain, and it is connected to a new brain hemisphere. The surviving half of your body is then successfully sewn together with a new half body. You therefore think that you survived the accident and that you are still yourself. But then the doctors

inform you that they also managed to save the other halves of your brain and body, and that these parts are now sewn together with new halves, and that they also have memories of the past. Now, suddenly, there are two persons with physical and psychological coherence and continuity with the former person, but which of these is you? What should be the reason that only one of them is you? Is it the case that you survived first, but then ceased to exist when two new persons appeared—but how could a double success be a failure? Or can we say that two persons can be identical to one person—but how can one be identical to two?

Another example of absence of cohesiveness is the well-known trolley problem (Foot 1967, 4): A person has tied five persons to a rail track and a runaway trolley is approaching, about to kill them. You can make the trolley change tracks by pulling a lever, but there is another person tied to that track who will then be killed instead. Should you pull the lever? Most people say “yes.” But what if a trolley is about to run over five persons, and you are standing on the bridge with a big man leaning over to see—is it then acceptable to push the big man over the bridge to stop the trolley and save five persons? This time most people would say “no,” and then the challenge is to explain the morally relevant difference in the two cases. The challenge here is to unite two moral intuitions with an overall principle that explains them both. This is lack of cohesiveness because we lack an explanation for why two descriptively similar events nevertheless are morally different.

While many thought experiments show connections lacking between data, thought experiments can also show how data are connected (as shown by Shoemaker above). Sometimes you have elements that you wish to connect or to give a specific justification. John Rawls, for example, wants to connect social democracy and justice by showing how social democracy yields a just society, and he does so through a thought experiment where people design a society behind a veil of ignorance. He argues that people would choose to create a kind of social democracy if they had to make a society where they had to live afterwards, not knowing what position or role they would have in the society. This is then meant to show that such a way of organizing society is fair (Rawls 1999, 118-123). Thomas Hobbes wants to connect the use of violence by the king with an ethical justification

of it, and he does so through the thought experiment that a social contract is written where the right to violence is consigned to the king in return for the protection this gives to all (Hobbes 1651/2017).

In the following, we discuss the aspect of coherence theory that concerns the amount of data a theory seeks to integrate, i.e., comprehensiveness. To recapitulate, the aspect of comprehensiveness refers to the amount of data that a theory integrates. The more relevant data that are integrated, the better. A thought experiment can be created to demonstrate that theory A lacks specific relevant data or that theory B integrates important data, but most often it is demonstrated that one theory is superior to another because it manages to integrate a larger amount of relevant data. In the following, we run through some examples.

Jonathan Schaffer has an interesting examination of different views on the concept of causality and the connection between cause and effect (Schaffer 2007). The philosophical discussions on causality are full of thought experiments that are used to test different views (Schaffer 2007). There are two main views on what constitutes causality. The first is causation as probability-raising and the other is causation as process linkage. If Pam throws a stone at a window, for example, so that it breaks, the probability-raising view will say that Pam's stone-throwing was the cause since it increased the probability of a broken window, while the process-linkage view will say that Pam's stone-throwing was the cause because a process linked her arm, the stone, and the window. Thought experiments can be used as arguments against both views by describing events that none of the theories manage to integrate.

A thought experiment against causality as probability-raising, on the one hand, is the following: Pam is standing with a stone in front of the window, while at the same time the more reliable vandal, Bob, holds his throw waiting to see if Pam throws instead. When Pam throws, the probability that the window will break decreases, since there would be a higher probability of a broken window if Bob were the thrower, and he would have thrown if Pam had not.

A thought experiment against causation as process-linkage, on the other hand, is the following: Pam uses a catapult to throw a stone at the window. Pam pulls a lever to release a spring, and then the catapult throws a stone



on the window, and it breaks. Pam is process-linked to the lever and the catapult is process-linked the window, but there is no energy, force, momentum or other link between Pam and the window. Yet we want to say that Pam was the cause of the broken window.

Here we can see that thought experiments can be used to point to data that a theory does not manage to integrate. Defenders of the different theories could use these examples against each other to argue for the superiority of their own theory.

#### 4. Conclusion

In the previous section, we described how theories can be strengthened, weakened or compared by use of coherence and provided examples from existing thought experiments. Strengthening a theory can be understood as giving an argument for a theory. Weakening a theory can be understood as giving an argument against a theory. Comparing two theories to show that A is better than B, can be understood as giving an argument for A being better than B.

A deductive argument clarifies what is entailed in the premises. A deductive argument can clarify connections in a theory, and thus make it more coherent and better justified as true. It can also demonstrate the presence of an inconsistency or lack of coherence, thus weakening a theory. An inductive argument is an argument where the conclusion is not necessarily true even if the premises are true. How good the argument is depending on how relevant (“relevant” in the sense of logical strength) the premises are, if true. It is contested what makes inductive arguments relevant. We argue that the relevance of an inductive argument is the degree to which it makes one theory more coherent than the alternatives (or less coherent if it is a counterargument).

Given this understanding, thought experiments can obviously be both deductive and inductive arguments, used to strengthen, weaken, or compare theories. But scientific theories are not only strengthened and weakened by arguments, they are also strengthened and weakened by new data that we discover. Thought experiments can also be data, when we use a broad understanding of data—as we have good reasons to do.

We have already given examples of how thought experiments can be understood as new data. An area where they can obviously be new data is when the mind is the topic of scientific exploration. Thought experiments can teach us about things that are impossible to think or things where the negation cannot be thought.: For example, you cannot imagine an event separate from time and space (cf. Kant). You cannot consistently think that thoughts themselves are illusions (cf. Descartes).<sup>6</sup> Thought experiments can give us data about modal facts of possibility, impossibility, necessity, or transcendental conditions.

In many cases, thought experiments employ knowledge we have by empirical means. But empirical knowledge is also interpreted by thoughts. Thought experiments and empirical experiments are interwoven and have very similar and overlapping functions in science. One might think that thought experiments are mainly about deducing inconsistencies. But in this article, we have shown that pointing out inconsistencies very often has the inductive function of showing one theory to be more cohesive and comprehensive than another, while the theories can also be reconfigured and further nuanced to deal with the thought experiments. In other cases, the function of thought experiments is not about deducing inconsistency, but instead demonstrating consistency, cohesiveness or comprehensiveness. The goal of this article was to show this rich use of thought experiments and their close argumentative link to normal experiments owing to the fact that thought experiments are also data for theories to integrate.

The coherence theory we have here presented uses a broad understanding of data and of theory. We do not have access to the world in itself outside of our mind. All data are like small theories: interpretations of the world that can be wrong. Very often we have good reason to believe that what we observe is true, especially if many people observe it, and there is no coherent alternative explanation but to believe that what we observed was true. But many observations are also uncertain, contested and open to many interpretations. Both observations and thought experiments are truth candidates and thus data that theories should consider when trying to make

---

<sup>6</sup> Some have contested these claims, which we think is unfeasible given proper definitions of the terms—but there is not room for that discussion here.

the most coherent theory of the world. Some observations and thought experiments will be included and some will be discarded even in the most coherent theory.

When we learn something new about the world, it is not the case that the world in itself is revealed to us. What in fact happens is that observations, thought experiments, reflections on language and definitions, understandings of connections etc. help us understand that one theory of the world is more coherent than another. We then replace our earlier understanding with a more coherent understanding – often by integrating new data, but sometimes also by rejecting old data as false. This is how thought experiments teach us something new about the world, namely by strengthening, weakening or comparing theories, thus making us reconsider which understanding of the world is most likely to be true.

Sartoris theory of moving from internal to external validity is not wrong, but very narrow, focusing on a subset of thought experiments and not explaining how the external validity is established. Given coherentism, external validity is established by showing that a theory is more coherent than alternative theories. In this article, we hope to have contributed with both a broader and deeper understanding of how thought experiments function and give us new knowledge about the empirical world.

### References

- Brown, James Robert. 1991. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. London: Routledge.
- Brown, J. R. 2004. “Why thought experiments transcend empiricism.” In C. Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science*, 23–43. Oxford: Blackwell.
- Campbell, D. T. 1957. “Factors Relevant to the Validity of Experiments in Social Settings.” *Psychological Bulletin*, 54, 297–312. <https://doi.org/10.1037/h0040950>
- Cappelen, Herman. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.
- Cohnitz, Daniel. 2006. *Gedankenexperimente in der Philosophie*. Leiden: Brill.
- Dennett, D. 1996. “Intuition pumps.” In J. Brockman (Ed.), *Third Culture: Beyond the Scientific Revolution*, 181–197. New York: Simon and Schuster.
- Descartes, René. 1641/1984. *Meditations on First Philosophy*. Translated by John Cottingham. Cambridge: Cambridge University Press

- El Skaf, R. 2018. “The Function and Limit of Galileo’s Falling Bodies Thought Experiment: Absolute Weight, Specific Weight and the Medium’s Resistance.” *Croatian Journal of Philosophy*, 18(52), 37–58.
- Foot, P. 1967. “The Problem of Abortion and the Doctrine of the Double Effect.” *Oxford Review*(5), 5-15.
- Gendler, Tamar Szabo. 2004. “Thought Experiments Rethought—and Reperceived.” *Philosophy of Science*, no. 71, 1152-1163. <https://doi.org/10.1086/425239>
- Hacking, I. 1993. “Do thought experiments have a life of their own? Comments on James Brown, Nancy Nersessian and David Gooding.” In Hull, D., M. Forbes, & K. Okruhlik (Eds.), *Proceedings of the Philosophy of Science Association Conference 1992*, Volume 2, 291–301. Chicago: University of Chicago Press.
- Hobbes, Thomas. 1651/2017. *Leviathan*. New York: Penguin.
- Kuhn, Thomas S. (1977). “A Function for Thought Experiments.” *The Essential Tension: Selected Studies in Scientific Tradition and Change*, 240–265. Chicago: University of Chicago Press.
- Miscevic, N. (1992). “Mental Models and Thought Experiments.” *International Studies in the Philosophy of Science*, 6(3), 215–226. <http://dx.doi.org/10.1080/02698599208573432>
- Norton, J. D. (2004). “Why Thought Experiments Do Not Transcend Empiricism.” In C. Hitchcock (Ed.), *Contemporary Debates in the Philosophy of Science*, 44–66. Oxford: Blackwell.
- Palmieri, Paolo. 2018. “Galileo’s Thought Experiments: Projective Participation and the Integration Of Paradoxes.” In Michael Stuart, Yiftach Fehige, and James Robert Brown (eds.). *The Routledge Companion to Thought Experiments*, 92–110. London: Routledge.
- Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
- Puntel, Lorenz B. 2008. *Structure and Being. A Theoretical Framework for a Systematic Philosophy*. University Park: The Pennsylvania State University Press.
- Putnam, Hilary. 1975. “The Meaning of ‘Meaning.’” In Hilary Putnam. *Mind, Language and Reality. Philosophical Papers*, Vol. 2. <https://hdl.handle.net/11299/185225>
- Rawls, John. 1999. *A Theory of Justice*. Oxford: Oxford University Press.
- Rescher, Nicholas. 1973. *The Coherence Theory of Truth*. Oxford: Clarendon.
- Rescher, Nicholas. 2010. *Reality and its Appearance*. London: Continuum.
- Sartori, L. (2023). “Putting the ‘Experiment’ back into the ‘Thought Experiment.’” *Synthese* 201:34. <https://doi.org/10.1007/s11229-022-04011-3>
- Schaffer, Jonathan. 2016. “The Metaphysics of Causation.” *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/causation-metaphysics/>. Visited November 15, 2023.

- Shoemaker, Sydney. 1969. "Time without Change." *The Journal of Philosophy*, vol. 66, no. 12, 363–381. <https://doi.org/10.2307/2023892>
- Van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.
- Walton, Kendall L. 1990. *Mimesis as Make-Believe. On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.

## Free Will: A *consensus gentium* Argument

William Hunt\*

Received: 20 March 2023 / Revised: 1 November 2023 / Accepted: 7 January 2024


*Abstract:* This argument for free will is a probabilistic one based upon two conjectures: first, that of consensus; namely, that a large majority of people believe that they and others have free will and second, that *a priori* proofs against the existence of free will either fail or remain questionable. If these two conjectures hold, an inductive argument follows on the basis of beliefs founded upon justified auxiliary assumptions, assumptions that ensure a well-defined probabilistic relationship between the evidence of consensus and the proposition *free will exists* in an elaborated form. I will then demonstrate, through subjective Bayesian confirmation theory, that such evidence probabilistically confirms this proposition. Moreover, if one's prior degree of belief in the existence of free will is not very low - prior that is to consideration of the evidence - then, provided this evidence is factual, it is likely that one's resultant degree of belief in the veracity of the proposition is not only rational, but also compelling.

*Keywords:* Bayesianism; consensus; free will; libertarianism; probability.

---

\* MF Norwegian School of Theology, Religion and Society

 <https://orcid.org/0000-0002-8206-3413>

 MF Norwegian School of Theology, Religion and Society, P.O.Box 5144 Majorstuen, 0302 Oslo, Norway

 [willhunt@gmail.com](mailto:willhunt@gmail.com)



## 1. Introduction

The free will debate is both ancient and voluminous and is fundamentally divided into three questions: what is free will, does it exist and, if it does, how can it be coherently explained? The objective of this paper is to address the second question – does free will exist? This question has been intensely debated with several conceptual refutations. In contrast, evidence for its existence has been proffered; however, there is an omission in this evidential deliberation. There has been no assessment of the probabilistic potency of this evidence on the likely truth of the proposition *free will exists*. Evidential arguments that circumvent this probability perspective can lead to an exaggerated view of the force of the evidence - I intend to address this lacuna. Thus, my primary objective, as the title suggests, is to provide a *consensus gentium* argument or *agreement of the people*: the majority. Herein, I extend its application beyond a simple majority to a probabilistic conception in terms of degrees of belief. Then, through Bayesian confirmation theory, I apply the evidence of consensus and other relevant facts to the free will proposition. This will demonstrate probabilistic confirmation of the proposition that establishes free will is more likely to exist than not.

Although it is likely that there would be a general consensus in support of free will, the application of Bayes' theorem provides objectivity to any probabilistic connection.

### 1.1. The free will proposition

To attempt to ascertain the probability that the proposition *free will exists* is true, a more encompassing proposition is required to give substance to its meaning. There are a number of interpretations of the meaning of free will and the following proposition is one that encompasses both the compatibilist and libertarian perspectives including my own:

*h: Agents possess the capacity to make uncompelled reasoned choices between alternative possible actions so as to fulfil or resist a desire, whereby any resultant action or abstention to instantiate that choice is both intended and uncoerced.*

This free will proposition is commensurable with both the compatibilist and libertarian. For the compatibilist, the origin of the power that determines the agent's choice and any resultant action thereto is the causal nexus of a deterministic universe. By contrast, for the libertarian, the origin of the power that determines the agent's choice and any resultant action is the agent herself. For both parties, uncompelled choice and uncoerced action are necessary for predicating free will to the agent.

## 2. Probability

I now turn to the central theme of my argument – probability and the *consensus gentium* argument.

I contend that any argument for the existence of free will is primarily an evidential one, with the veracity of the free will proposition being subject to that evidence through a probabilistic analysis.

To ensure the proposition/evidence relationship is sound, any probabilistic analysis should be commensurable with the axioms of probability; herein, I apply the Kolmogorov axioms.

Probability theory includes a range of theories beyond the scope of this paper, but herein, I employ subjective probability.<sup>1</sup>

### 2.1 Subjective probability

Subjective probability is a form of epistemic probability which comprises two theories:<sup>2</sup> the logical theory and the subjective theory; (Gillies 2003, 37ff).<sup>3</sup> However, the logical theory is problematic as, although it complies with the axioms of probability, it relies upon the Principle of Indifference, a principle that leads to several paradoxes; (Gillies 2003, 33-49).

The subjective theory is based upon the personal credence someone gives to the chance of, in this case, a proposition being true. Warranted credence

---

<sup>1</sup> See Gillies (2003) for an analysis of the different theories of probability.

<sup>2</sup> Epistemic probability contrasts with objective probability of which there are two theories: frequency theory-Von Mises (1919) and propensity theory-Karl Popper (1959)

<sup>3</sup> Gillies also demonstrates that subjective probability is both necessary and sufficient for the axioms of probability; Gillies (2003, 59-64)



is expressed as a coherent degree of belief in a proposition given the evidence. Coherence is derived from the act of placing a bet, and defined in terms of avoiding a Dutch-book bet.<sup>4</sup> Like logical probability, credence is expressed as a numerical value between 0 and 1 on the probability continuum.

## 2.2 Conditional probability

Given a proposition  $h$  and evidence  $e$ , conditional probability is the probability of the truth of  $h$  given  $e$  expressed as  $\Pr(h|e)$  and defined as:

$$\Pr(h|e) = \frac{\Pr(h \& e)}{\Pr(e)}$$

provided,  $\Pr(e) \neq 0$

Herein, I employ conditional probability in the form of likelihoods, and for such likelihoods to be well defined, there are auxiliary assumptions that must be accounted for.<sup>5</sup> Auxiliary assumptions are crucial to this probabilistic analysis of free will and I consider them below.

## 2.3 Bayes' Theorem

A useful probabilistic tool in assessing whether some evidence provides justifiable credence in accepting a proposition to a certain degree is that of Bayes' theorem,<sup>6</sup> and the theorem is commensurable with subjective probability. The theorem is expressed as follows:

$$\Pr(h|e \& k) = \frac{\Pr(e|h \& k) \Pr(h|k)}{\Pr(e|k)}$$

---

<sup>4</sup> A Dutch-book bet is where odds are set by the bookmaker to win more money than the better, even if the better wins the bet. A Dutch-book is avoided by coherence with the axioms of probability. Coherence, as so defined, was proposed almost simultaneously by Frank Ramsey, 1926 and Bruno De Finetti, 1930

<sup>5</sup> A likelihood is a conditional probability function of the form  $\Pr(e|h \& k)$ ; where, in this case, the probability of the evidence  $e$  is conditional on the assumed truth of the proposition  $h$  and background knowledge  $k$ , which include auxiliary assumptions that create a well-defined probabilistic relationship between  $h$  and  $e$ .

<sup>6</sup> For challenges to the theorem and their defence see Earman (1996, Ch.4)

and can be reformulated, in its comparative form; i.e.,  $h$  compared to  $\neg h$  (mutually exclusive and exhaustive propositions), using a likelihood ratio as follows:

$$\Pr(h | e \ \& \ k) = \frac{\lambda \times \Pr(h|k)}{(\lambda \times \Pr(h|k) + (1 - \Pr(h|k)))}$$

Where the likelihood ratio  $\lambda$  is:

$$\lambda = \frac{\Pr(e|h \ \& \ k)}{\Pr(e|\neg h \ \& \ k)}$$

The function  $k$  represents background knowledge that also includes the all-important auxiliary assumptions.

This reformulation of Bayes' theorem is useful when used with subjective probability as likelihood ratios are much easier to assess subjectively than individual likelihood values.

$\Pr(h|e \ \& \ k)$  is the posterior degree of belief in the proposition  $h$ ; that is, the new degree of belief that would be formed if the person *conditionalised* on the evidence  $e$  with respect to  $h$  (see Bayesian conditionalisation below).  $k$  is background knowledge, which includes the auxiliary assumptions.

It can be seen that if  $\lambda$  is greater than 1, then probabilistic confirmation follows; i.e.  $\Pr(h|e \ \& \ k) > \Pr(h|k)$ . If it is less than 1 then probabilistic disconfirmation follows; i.e.  $\Pr(h|e \ \& \ k) < \Pr(h|k)$ . If it is 1 then there is no confirmation or disconfirmation; i.e.  $\Pr(h|e \ \& \ k) = \Pr(h|k)$ .

### 3. A *consensus gentium* argument

Having outlined my probabilistic methodology, I now turn to the probabilistic analysis for the existence of free will – a *consensus gentium* argument. I will argue for, and present values for the functions in the likelihood ratio form of Bayes' theorem above, with particular attention to the evidential function  $e$  – this represents the consensus element of the *consensus gentium* argument. In addition to this evidential element, I will focus on the auxiliary assumptions essential for any conditional probability analysis of this sort; that is, ensuring there is a well-defined probabilistic relationship between  $h$  and  $e$ .

### 3.1. Propositions $h$ , $\neg h$ and $e$

The free will proposition  $h$  is more than just *free will exists*, as it can imply different interpretations with different existential possibilities and only needs one case in an infinite universe to confirm its truth, which does not capture its intended meaning. Thus, for this Bayesian argument I expose  $h$ , the aforementioned free will proposition, and its negation  $\neg h$ , to the evidence of consensus:

$h$ : Agents possess the capacity to make uncompelled reasoned choices between alternative possible actions so as to fulfil or resist a desire, whereby any resultant action or abstention to instantiate that choice is both intended and uncoerced.

$\neg h$ : Agents DO NOT possess the capacity to make uncompelled reasoned choices between alternative possible actions so as to fulfil or resist a desire, whereby any resultant action or abstention to instantiate that choice is both intended and uncoerced.

A *consensus gentium* argument would normally be considered fallacious; evidence of consensus is, *prima facie*, subject to prejudice and can be unreliable. However, herein I justify its application by employing probability theory with robust auxiliary assumptions.

The vast majority of participants of a general survey on free will would be unaware of the nuances of the free will debate to make an informed decision. In fact, a loss of precise conceptual correspondence between individuals is likely to lead to confusion and imprecision, making any data untrustworthy. Thus, the population of this *consensus gentium* argument should comprise a body of participants likely acquainted with the ebb and flow of the free will debate. Given the above,  $e$  is defined as:

$e =_{df}$  *The mean degree of belief in free will expressed as a probability quotient within a given population acquainted with the free will debate is greater than 0.5 – more likely than not.*<sup>7</sup>

---

<sup>7</sup> I should add that  $e$  represents a consensus as independent agreement, not consensus by cooperative agreement as with intersubjective probability.

### 3.2. Auxiliary assumptions

Subjective Bayesianism is based upon subjective probability which, in this analysis, is a degree of belief in a proposition based upon evidence – a truth-conducive interconnection enabled by the auxiliary assumptions.<sup>8</sup>

Adopted auxiliary assumptions have themselves to be justifiable as there is a danger of bulking-up the likelihood with unjustifiable auxiliary assumptions to leverage the probability in one’s favour. “What is needed is not the invention of auxiliary propositions...but the identification of auxiliary information that is independently supported.” (Sober 2008, 168) I provide five auxiliary assumptions as follows:

#### 3.2.1. Naturalistic realism

Naturalistic realism assumes that, given all the possible explanations of reality, the best explanation is that provided by current scientific theory, which can change over time – it is defeasible; (Kuhn 2012). It projects its conception of reality beyond that which is observable, but is still commensurable with observable evidence. The science of cognitive psychology accepts the reality of the mind, in particular intentional agency, and is a form of naturalistic realism. Moreover, it is a widely accepted evidence-based science consistent with that of other human and social sciences.

The tenets of cognitive psychology (particularly intentional agency) under the lens of naturalistic realism would be a substantial auxiliary assumption in the likelihood and prior functions  $\Pr(e|h \ \& \ k)$  and  $\Pr(h|k)$ .

#### 3.2.2. Evolution

The emergence of consciousness, instincts and cognition in early life on Earth provided such life with improved survival chances. As the evolutionary course progressed, instincts and beliefs emerged that interfaced with

---

<sup>8</sup> The auxiliary assumption requirement is associated with the Duhem–Quine thesis. The thesis is a combination of Pierre Duhem’s 1904/5 thesis and Willard Van Orman Quine’s 1951 article Two Dogmas of Empiricism. In short, it is impossible to test a scientific hypothesis in isolation. It requires several background assumptions termed auxiliary assumptions or bundles of hypotheses; see Gillies (1993, 98–116)

environmental dangers, food supply, predators and many other advantages and threats, providing further survival chances and subsequent adaptation. Moreover, if those beliefs were true rather than false, survival chances would improve even further. Thus, the cognitive ability and motivation to harbour true beliefs delivers selective pressure on both animals and early humans which would have manifested itself in finding shelter, socialization, communication and so on. Indeed, Anabela Pinto argues that the complex beliefs that modern humans hold share a relationship with animal beliefs indicating biological roots of belief formation by adaptation. This, she argues, points the way to an evolutionary explanation for our complex linguistic belief concepts; (Pinto, 2022, 22).

Clearly, all our beliefs are not true, but we are motivated to harbour true rather than false beliefs and change them if shown to be false – an echo from our distant biological roots. Moreover, in modern life there is personal developmental pressure from a whole range of sources to form and harbour true beliefs over false ones. These two doxastic factors help define a probabilistic relationship between  $e$  and  $h$  in the likelihood  $\Pr(e|h \ \& \ k)$ . This is because a belief in free will assumes the truth of  $h$  rather than  $\neg h$ , and true beliefs are more likely with the evolutionary and developmental influences than without.

### 3.2.3. *Phenomenology*

The phenomenal experience of free will in terms of the first-person sense of being in control, losing control and regaining control (control-phenomena) are very common experiences for us all, albeit sensed differently. Similar experiences provide a sense of temporal and spatial awareness, self-awareness, social awareness and a host of other essential senses.

Control-phenomena suggests an ontological as well as qualia-logical content – it provides a lens on the power of self-determination possessed by us. This is central to the social science of cognitive psychology as outlined above. However, do control-phenomena ensure the existence of self-determination in the same way that self-awareness ensures our existence?

Self-determination implies that we can control our decision making. Loss of self-control means we are under the spell of our emotions, and decision-making becomes less rational – we're not fully in control of our decision making. If we're not fully in control of our decision-making then we cannot

be said to be self-determining our decisions – our emotions have taken over for example. However, if we begin to regain self-control, self-determination progressively returns. Even though self-control does not entail self-determination, the two concepts are related and control phenomena indicative of the reality of self-determination.

We have a distinct conscious awareness of losing control when we are at the mercy of causal-power such as the emotions of anger and fear, and we sense its reinstatement with the shift back to self-control - all frequently manifested in behaviour with concomitant third-party interpretation. Indeed, first-person experiences of pain, control, irritation, affection etc are frequently manifested in behaviour, and introspective predicates develop from that correlation; (Wittgenstein 1976, sec.244).

Token control-phenomena can be subject to confabulation in terms of scaling, particularly with retrospective rather than concurrent recall. Moreover, cognitive masking during loss of control such as shouting, or being spoken to whilst regaining control can weaken recall of the control-phenomena. Notwithstanding, this type of introspective experience is similar to pain in terms of recalling intensity - both types of experience are incorrigible, even if detailed recall is confabulated; (Shanahan 2010, 67-89).

The auxiliary assumption of control-phenomena helps create a well-defined probabilistic relationship with the function  $\Pr(e|h \ \& \ k)$  – also with the assessment of the prior  $\Pr(h|k)$ .

#### *3.2.4. Blame and liability*

When our choices are instantiated by us they can lead to blame and legal liability if there is a breach of moral rules or law respectively. The rule of law is a global form of social control even though laws can vary from country to country, but all are predicated on the understanding that if members of a social group breach such laws they will be held responsible. Moreover, many societies hold to the maxim that being responsible for breaking the law is a necessary criterion for blame and punishment, and being responsible means that actions are down-to-them, they are the source of the breach that could have been avoided by an alternative choice of action.

Although it could be argued that law is solely a means of social control, and punishment merely deterrence, the reality is that, in addition to any

social control, blame and punishment is fundamentally retributive. Indeed, Daniel McDermott argues that, analogous to financial debt, criminals incur a non-consequentialist “backward looking” moral debt to their victims (including society in some cases) and proportionate punishment, as retribution, represents a settlement of that debt; (McDermott 2002, 439–464).

Blame and punishment as retribution presupposes liability that, in turn, presupposes responsibility which then implies personal control in our choices of action unless proved otherwise; (Pereboom 2014, 153–160). It is the global proliferation of moral rules and the rule of law, together with assumed self-determination in breaches of them that, as an auxiliary assumption, helps create a well-defined probabilistic relationship between  $h$  and  $e$  in the likelihood  $\Pr(e|h \ \& \ k)$ . Moreover, this auxiliary assumption has application to the prior  $\Pr(h|k)$  as background knowledge  $k$  would include knowledge of moral rules and the rule of law and the related assumption of intentional agency.

### 3.2.5. *Scholarly error*

There can be widely held false beliefs among scientists; for example, the Newtonian concept of gravity or the hypotheses of phlogiston, vitalism and luminiferous ether. The cause of these false beliefs was primarily an absence of relevant facts, faulty or limited equipment and a lack of insight rather than self-delusion or mental weakness. Nevertheless, historical precedent and peer/societal pressure could have contributed, and the same could apply to free will; (Kuhn 2012, 66–76). Neuroscience or possibly physics may eventually fill any evidential and explanatory lacunae and refute the existence of free will – it is a defeasible notion.<sup>9</sup> In fact, defeasibility is an assumption of naturalistic realism and, given the history of scholarly error, an auxiliary assumption arises providing probabilistic definition to the function  $\Pr(e|\neg h \ \& \ k)$ . Residing in background knowledge  $k$ , this assumption may also affect the assessment of the prior  $\Pr(h|k)$ .

---

<sup>9</sup> The time-lapse in the Benjamin Libet et al (1983) experiment came close to such falsification with reflexive reactions; however, not with reflective reactions; also see List (2019, 141-147) and Rolls (2012)

## 4. Valuation

I shall now apply the presuppositions of subjective probability to ascribing values to the conditional probability functions of the likelihood ratio form of Bayes' theorem.

### 4.1. $Pr(h|k)$ : Prior belief in $h$

Because there are differences of opinion concerning the truth of the free will proposition,  $Pr(h|k)$  could be allocated 0.5 on the basis of the Principle of Indifference of logical probability. However, because of the paradoxes associated with the Principle, in this case the Book Paradox,<sup>10</sup> I resort to a subjective evens betting position and still set  $Pr(h|k) = 0.5$ . I do this on an assumption of no prior knowledge of the auxiliary assumptions.<sup>11</sup>

### 4.2. $Pr(e|h \ \& \ k)/Pr(e|\neg h \ \& \ k)$ : the likelihood ratio $\lambda$

Given the substantiality of the auxiliary assumptions that provide probabilistic definition between  $h$  and  $e$ , I would argue that  $Pr(e|h \ \& \ k) \gg Pr(e|\neg h \ \& \ k)$  and as such  $\lambda$  would be greater than 1 leading to confirmation; i.e.  $Pr(h|e \ \& \ k) > Pr(h|k)$ . Others may prefer  $Pr(e|h \ \& \ k) > Pr(e|\neg h \ \& \ k)$  or  $Pr(e|h \ \& \ k) \gg \gg Pr(e|\neg h \ \& \ k)$ .

Despite the validity of subjective probability, providing a precise value of  $\lambda$  would be somewhat arbitrary. Notwithstanding, given the subjective assessment  $Pr(e|h \ \& \ k) \gg Pr(e|\neg h \ \& \ k)$ , my evaluation of ratio  $\lambda$  is not less than 1.5 but certainly no greater than 3; i.e. a range of possible values.

Given  $Pr(h|k) = 0.5$  (on the basis of a subjective *evens* bet referred to above) and  $\lambda = f(y)$  which lies in the range  $[1.5, 3]$  then, from the comparative form of Bayes' theorem above:

---

<sup>10</sup> The falsity of any one of the criteria of  $h$  would ensure  $\neg h$ . Thus, there could be range of  $\neg h$  definitions, each with a probability of 0.5 given the Principle of Indifference which would breach axiom 1; see Gillies (2003, 37f)

<sup>11</sup> The problem of old-knowledge would challenge the use of Bayes' theorem in this application; see Glymour (1980, 86). However, see Howson & Urbach (1991, 270f) for a defence.



$$\frac{y \times 0.5}{(y \times 0.5) + (1 - 0.5)} = \Pr(h | e \ \& \ k) \text{ which lies in the range } [0.60, 0.75]$$

Thus, based upon an assumption of the evidence of consensus, the proposition *h free will exists* as so expressed is more likely true than false with a subjective probability value of between 0.60 and 0.75 – probabilistic confirmation. However, for Bayesian-conditionalisation to occur (coming to believe the above result) the evidence must be actual rather than assumed.<sup>12</sup>

### 4.3. Bayesian Conditionalisation

It is the reality of *e* that gives the Bayesian reason to accept the posterior probability value and conditionalise on that value, and it is to surveys I turn to provide that evidence.

There are two surveys that plainly fulfil the *acquainted with the free will debate* criterion within *e*: The *2009* and the *2020 Philpapers surveys*, both conducted by David Bourget and David Chalmers who posed questions to target audiences comprising philosophers on a range of philosophical issues, including one on free will (Bourdet and Chalmers 2013; 2020; 2023). A brief outline of the surveys is as follows:

- (i) *Target population (2009)*: The survey was taken by 3226 respondents with philosophical backgrounds from Australasia, Canada, Europe, UK and US.  
*Target population (2020)*: The survey was taken by 7685 respondents with philosophical backgrounds from New Zealand, Canada, Ireland, UK and US.
- (ii) *The free will question (2009)*: Accept or lean towards: compatibilism, libertarianism, or no free will?  
*The free will question (2020)*: Accept or lean towards: compatibilism, no free will, or libertarianism?
- (iii) *Numbers answering the free will question (2009)*: 931  
*Numbers answering the free will question (2020)*: 1758

---

<sup>12</sup> For a discussion on Bayesian conditionalisation see Howson and Urbach (1991, 67f).

- (iv) % *Results (2009)*: compatibilism - lean towards: 226 & accept: 324; libertarianism - lean towards: 56 & accept: 72; no free will - lean towards: 62 & accept: 53; and 'other' 138 of which 38 are only relevant to the question - being the agnostic response.
- (v) % *Results (2020)*: compatibilism - lean towards: 490 & accept: 550; libertarianism - lean towards: 138 & accept: 193; no free will - lean towards: 102 & accept: 95; and 'other' 190. The 'other' responses are more varied than the 2009 survey, but the only clear applicable result being the agnostic one of 80 responders.

What is interesting with both these surveys, is the number of participants that did not answer the free will question (2009: 2295; 2020: 5927) indicating agnosticism - a degree of believe of 0.5 on the probability continuum. In terms of degrees of belief, these agnostics cannot be ignored and neither can the disbelievers nor the 'other' group. Indeed, 38 responders of the 'other' group in the 2009 survey were agnostic; there is no indication that the remaining responses were relevant to the question, and therefore should not be included in the analysis.

What is also interesting from these surveys with regard to degrees of belief, is the division between 'lean towards' and 'accept'. Thus, modelling the probability continuum into equal subintervals to represent 'lean towards' and 'accept' for disbelief and belief, with agnosticism at the midpoint of the continuum we have: 0...0.17...0.34...0.50...0.67...0.84...1.<sup>13</sup>

Applying the above model to the 2009 survey, a mean degree of belief of  $\approx 0.53$  results and applying the above model to the 2020 survey, a mean degree of belief of  $\approx 0.53$  results. These two results are virtually identical adding further credence to warranted Bayesian conditionalisation of  $\Pr(h|e \& k)$  whose value lies in the range [0.60,0.75].

The individual degrees of belief among the respondents will vary despite being within the accept, lean towards or agnostic groups. Consequently, the above probability continuum model may not be precise; notwithstanding,

---

<sup>13</sup> The probability calculus assumes that probabilities are real numbers and each probability on the continuum should, theoretically, be represented by an infinite decimal (e.g. 0.1 is given by 0.999... $\infty$ ) because the probability space between any one point on the continuum and another is infinitely divisible. However, with subjective probability the calculus is an approximation as values are vaguer.

the values selected in the model are an even distribution of the continuum that reflect the differing degree of belief modes. Even skewing the model towards disbelief, say 0..0.1..0.17...0.50..0.60..0.67...1 yields a mean degree of belief greater than 0.5 for both the 2009 and 2020 surveys.

The above mean degree of beliefs results for the 2009 and 2020 surveys summate the results of the compatibilists and libertarians. However, this may not be justified given the differences between the two camps. Supporting facts in Bayesian conditionalizing are not intended to be entailments, but are persuasive facts to a greater or lesser extent. Thus, it is a question of whether  $h$  is congruent with both compatibilist and libertarian views of free will. I believe it is, and therefore contend that it is sound to combine both the compatibilist and libertarian results in the surveys, and that the two surveys provide justifiable evidence to Bayesian conditionalise on  $\Pr(h|e \ \& \ k)$  whose value lies in the range [0.60,0.75].

There are other free will surveys, but none as convincing and targeted as the two above.<sup>14</sup>

## 5. Preliminary conclusion

My preliminary conclusion is that, based upon the evidence of consensus as so defined and the justification of the auxiliary assumptions, the posterior value of the proposition *free will exists* as so expressed is greater than 0.5 – free will is more likely to exist than not. This does not ensure the truth of the proposition, and there is room for evidential refutations that neuroscience may provide. However, there are *a priori* refutations that threaten this preliminary conclusion.

---

<sup>14</sup> See for example Wisniewski et al (2019), where their survey found an 82.33% belief in free will in the US and 85.44% in Singapore. Also, in January 2015 Gary Stix carried out a survey for Scientific American with 4672 responders from the US including some from France, Australia, New Zealand, Kuwait, Israel, the Philippines and India. 59% believed in free will and 41% disbelieved.

## 6. A libertarian interpretation

There are different interpretations of free will that are commensurable with the free will proposition *h*, mine is best described as non-causal libertarianism. This is the perspective I employ in addressing the *a priori* refutations of free will below; as such, a more detailed explanation is required.

By non-causal libertarianism I assert that persons possess a distinct power of self-determination compatible with causal determinism should it exist – a compatibilist interpretation of free will.

This interpretation implies two freedoms, *the freedom to choose otherwise* and *the freedom to do otherwise*; that is, free will and free action – interrelated and goal directed concepts. Indeed, this power to have *chosen otherwise* is central to a libertarian perspective – a *multi-way* power; (Pink 2019a, 268). Indeed, we sense this multi-wayness when reflecting upon alternative possibilities at the point of choice. In addition to the freedom to choose otherwise, the possibility for voluntary action to fulfil a desire, including abstention is also fundamental to libertarianism.

Intentional agency is another key criterion to the libertarian perspective despite its superfluous presence for the epiphenomenalist, incompatibilist or even the classical compatibilist.<sup>15</sup> Intentional agency is a goal directed choice followed by a goal directed voluntary act; (Pink 2019a, 259-266).

The most overt threat to the libertarian is the hypothetical problem of causal determinism. However, there is an alternative to the power of causal determinism; that is, the power of self-determination – an intrinsic power that I term *will-power* as contrasted to *causal-power*.<sup>16</sup>

Will-power is a ‘difference making’ intrinsic power that is,<sup>17</sup> like causal-power, difficult to define in an ontic sense, but unlike causal-power it is not

---

<sup>15</sup> See Hobbes (1841, XX) – although Hobbes eschewed the will as the cause of voluntary action, he viewed free will as simply the unimpeded satisfaction of desires.

<sup>16</sup> I use the term *will-power* only as a contrast to *causal-power* not in the usual sense of fortitude.

<sup>17</sup> By ‘difference making’ or “matters” as Helen Steward terms it, I refer to facts that make a difference to an outcome as contrasted with effects from dynamic causal forces; Steward (2014, 212ff). Christian List also adopts this notion; List (2019, 131–140). Both Steward and List apply the notion to causation, herein I apply it to will-power.

realized by observing regularities in nature as Hume would have it; (Hume [1739] 1985, III, 117-123). In fact, our habits can be regular, but our choices are frequently not. Indeed, as well as lacking such regularities the difference between will-power and causal-power is stark - causal-power excludes goal-directed intention, choice and multi-wayness, blame and moral responsibility. These differences between will-power and causal-power remain even if the concept of causation is expanded from its dynamic character such as *the wind blew the chimney off*, to include making a difference such as *inflation soared because wages increased*. Compare this to *I refused a drink because I'm driving* – the above differences still apply.

From observing causal-power in the natural world, there is a temptation to predicate libertarian free will as causal - *agent-causal libertarianism*.<sup>18</sup> However, as I argue above, will-power is so different to causal-power that it warrants its own designation rather than being a sub-category of causal-power.

This power-difference perspective assumes that both free will and causation exist, with each having powers to bring about change in different ways. However, although related in this sense, the two powers cannot be conflated. Thus, the tag *non-causal libertarianism* has application to this power-difference perspective.

### 6.1. *The mechanics of libertarianism*

Libertarianism is intuitively compelling given our everyday phenomenological experiences, and some evidence does suggest a relationship between specific conscious decision making and concomitant action.

### 6.2. *Correlation*

The essence of the libertarian perspective is that this conscious decision making has a power over and above the causal nexus in which the neural networks are seated. The threat to libertarianism is that such mental states are superfluous to the train of the causal nexus and that there is only a correlation between causal neural activity and conscious decision making.

---

<sup>18</sup> See Pink (2019a; Ch.14) for a critique of agent-causal libertarianism.

This correlation perspective has gained credence in neuroscience from the notion of the neural correlate of consciousness (NCC) pioneered with the use of fMRI scanning together with reported conscious experiences. (Charmers, 2000, 17-39) However, libertarians need more than a mere correlation, they need an instrumental power that emanates from the agent.

### 6.3. *Integrated information theory*

An alternative explanation to correlation is that given by integrated information theory (IIT); (Tononi, 2004), derived and explained by a set of five axioms and resultant postulates;<sup>19</sup> (Tononi et al, 2023, 3-5). With IIT, the information element relates to neural systems functioning to reduce experiential uncertainty by ruling out experiences from a range of possible ones – *differentiation* or a not this or that scenario; (Seth, 2021, 52f).

The integration process is a function of the neural system as a self-causal unified whole rather than isolated individual systems; i.e. parts of the system affect other parts and, in turn, are affected by them – a cause/effect interdependence. There is synergy with such integration; i.e. extra information –  $\Phi$  being the measure of this holistic extra; at least in principle. When such integration reaches a high level (maximally irreducible conceptual structure (MICS))<sup>20</sup> the system is conscious – a self-generated emergent property of integrated information.

With IIT, there is an identity assertion - consciousness is MICS,<sup>21</sup> and has variable  $\Phi$  depending of the level of integration; dreaming having low  $\Phi$ . MICS is self-generated, intrinsic to the neural system giving it potential for a libertarian explanation for free will; in short, it is the power to make a difference to itself. Indeed, MICS supervenes on the substrate of consciousness; i.e. the integrating neural system. In this way, MICS should be seen as a holistic state comprising the phenomenal and neurological. Indeed, IIT claims just that, as consciousness is deemed identical to MICS; i.e. a maximal integrated system. Thus, consciousness is inextricably integrated

---

<sup>19</sup> For Tononi, the move from axiom to postulate is not deductive, but abductive; (Tononi et al, 2023, 3)

<sup>20</sup> Irreducible in the sense of greater than the sum of its parts.

<sup>21</sup> In the sense of numerical identity, despite a difference in meaning.

with the neural system. As such, MICS possesses the power to change the system where the phenomenal alone could not.

IIT provides a compelling explanation of how an agent possesses intrinsic self-generated power to make and fulfil her choices – libertarian free will.

IIT is a hypothesis, and as such, requires empirical evidence in support, and given that MICS is essentially intrinsic to the person, observation, other than measuring the firing and integrating of neural systems, is private. However, comparative brain functions provide persuasive evidence for IIT. Despite the cerebellum possessing c.70 billion neurons it has been shown to have no contribution to the conscious state. In contrast, the thalamocortical system possesses c.16 billion neurons yet is the core of consciousness. The fundamental difference between the two brain systems is that the thalamocortical system has a complex nexus of interconnecting neural fibres compared to the cerebellum, enabling a very high level of integration – supporting evidence of IIT; (Tononi, 2004, 10ff).

Having espoused my non-causal libertarian perspective of free will, I turn my attention to the *a priori* refutations of free will and my replies thereto.

## 7. Refutations and Replies

There are a number of objections, with concomitant defences, to the existence of free will and the veracity of subjective Bayesianism, and I have referred to some in this paper. However, there are potential *a priori* refutations of the free will proposition and Bayes' theorem has no application where the probability of the proposition is 0 (certainly false), and these *a priori* arguments threaten just that. Similarly, an ontology that views only evidence as real and relegates self-determinism and intentional agency to only useful but unreal constructs ensures the failure of a Bayesian analysis of free will.

Currently, there are no *a priori* arguments that establish the certain truth of the free will proposition. However, there are arguments that suggest the free will proposition is certainly false, or unreal in the case of scientific instrumentalism. I examine such arguments as follows:

### 7.1. Strawson's infinite regress argument

Galen Strawson's argument assumes a premise that free will entails ultimate moral responsibility for acts freely instantiated; (Strawson 1986, 292ff).<sup>22</sup> This implies that the agent's intentional choice is, itself, freely preferred for certain reasons. In this way moral responsibility for the ensuing act can be predicated to the agent. However, these reasons must be persuasive enough for the agent to prefer her choice, i.e. "principles of choice...preferences, values, pro-attitudes, ideals..." (Strawson, 1986, 25) Nevertheless, for moral responsibility to hold, such persuasiveness cannot just emerge; it too must have principles of choice that provide reasons persuasive enough for the agent to be persuaded that her choice is best. Thus, to ensure moral responsibility and hence the agent's free will, an infinite regress of reasoned persuasion arises along with recursive moral responsibility which is impossible. Consequently, moral responsibility and therefore free will is forfeit; (Strawson 1986, 24f).

Our desires, wants and passions are not reasoned into existence, but just come over us as a result of our instincts, observations and the ever developing autobiographical-self; (Søvik 2018,106-126). The agent becomes motivated to satisfy those desires, wants and passions, and intentional agency aims to do just that. There is no prior persuasion that requires justification for those desiderata; they just come over her. Her goal directed voluntary action to satisfy the desiderata follows. It is at that point that moral responsibility has application, not for any self-reflective mental processes that rationalize the desiderata into mental existence - the locus of responsibility is the act or abstention.<sup>23</sup>

### 7.2. Incompatibilist refutations

Although, there are arguments that deny determinism on the grounds that there are no universal laws of nature; (Cartwright 1999), there are incompatibilist arguments against the existence of free will based upon the truth of determinism.

---

<sup>22</sup> Ultimate responsibility implies the agent is the sole originator of the act.

<sup>23</sup> See Mele (1995, 223-225) for a counterfactual example.



Established incompatibilist arguments are the Consequence Argument, the Origination Argument and the Mind Argument. The Consequence Argument, championed by Peter van Inwagen, claims that the invariable state of past events (fixity of the past) together with the laws of nature, determine all future events; (Van Inwagen 1983, 16). Consequently, volitional and voluntary acts are illusionary. He structured this argument formally, and there are challenges to the validity of his argument; (Van Inwagen 2002, 158–177).<sup>24</sup>

The Origination Argument is also based on the truth of determinism and states that if determinism is true, an agent's volitions do not originate with her but are extrinsically caused – the agent is not the ultimate source of her volitions. Given this, and assuming the necessity of origination for free will, determinism is a serious defeater; (Kane 1996, 79f).

The mind argument is so called because it challenges the existence of free will by the effects of determinism (and equally indeterminism) on the mental-state of freely choosing. The argument against free will has three strands; the first strand echoes the Consequence Argument, claiming that our choices are outside of our control and pre-caused by the progress of the causal nexus through time. The alternative is indeterminism, and if this means a random setting, then free will is still deniable – apparent actions are really just chance events. The second strand claims that volitional acts and voluntary acts are not acts at all unless they have a prior agential cause. The third strand concerns the action of choosing between alternative possibilities. Again, this echoes the Consequence argument as, given determinism, alternative possibilities are not possible; (Van Inwagen 1983, 126–152).

If true, self-determination as a *sui generis* power defeats all three arguments from incompatibilism. Determinism is more precisely termed *causal determinism*, and will-power, with its distinctive properties, is not causal and therefore, is not subject to the power of causal determinism. Thus, if determinism is true as a universal causal system, the agent can still intervene in its component causal sub-systems by her voluntary acts; (Ismael (2016, Chs. 4 & 5).

---

<sup>24</sup> Also see the Agglomeration Argument; McKay and Johnson (1996, 113–122)

Motivated by passive desires, wants and passions and influenced by the autobiographical self, the agent's will-power originates from her. Hence, the causal-nexus does not determine the agent's choices – she does. Further, voluntary acts are acts, not because of agential causation, but because of agential self-determination and, as I argue above, multi-wayness is a characteristic of will-power unlike causal-power. Thus, alternative possibilities are feasible irrespective of the truth of determinism. Also, indeterminism as viewed as chance events, is not a challenge to free will, as the power of self-determination eclipses any associated random possibilities.

There are additional replies to the three arguments from incompatibilism other than the power of self-determination. Indeed, incompatibilism and replies to it are a central item on the free will agenda but beyond the scope of this paper.

### 7.3. Instrumentalism

A scientific perspective in contrast to naturalistic realism is instrumentalism which is also evidence based, but it does not assert a true unobserved reality beyond the specific evidence. Any inference to the best explanation upon such evidence is based upon utility - the more useful the inference the more worthy its adoption. Thus, the instrumentalist would likely argue that cognitive psychology explains behavioural evidence, but such evidence is all that is real. Extrapolation from that evidence to intentional agency is not an assertion of reality, but an assertion of instrumentality in that it is practically useful; (List 2019, 74–77).

*Prima facie* it appears that instrumentalism, if sound, is an auxiliary assumption supporting the likelihood  $\Pr(e|\neg h \ \& \ k)$  as free will is a hypothesis based upon evidence. The hypothesis can be useful but cannot, itself, be considered true and in this likelihood the hypothesis is not considered true. However, instrumentalism does not consider  $h$  and  $\neg h$  in ontological terms, only in terms of their usefulness as constructs - the only component in the above Bayesian analysis that is real is  $e$ . So if instrumentalism is considered a sound scientific perspective of reality, then existential predicates do not apply to  $h$  and  $\neg h$  and the posterior function  $\Pr(h|e \ \& \ k)$  is otiose. However with instrumentalism, there is an overweighted emphasis

on distinct observation over the clear theoretical implications of evidence, bringing a vagueness to defining reality.

In summary, if instrumentalism is adopted, then explanatory power and correct predictions based upon non-existent scientific constructs seems to be more miraculous than rational; (Putnam 1975, 72f). Indeed, “Experimental physics provides the strongest evidence for scientific realism. Entities that in principle cannot be observed are regularly manipulated to produce new phenomena and to investigate other aspects of nature.” (Hacking, 1982, 71)

The notion of self-determination as will-power casts doubt on the soundness of the potential *a priori* arguments against free will, and instrumentalism is a questionable scientific principle. Based on the above replies, I do not believe that these refutations ensure that the probability of the free will proposition is zero, or in the case of instrumentalism, that Bayes’ theorem has no application.

## 8. Conclusion

The objective of this paper was to substantiate my contention that the debate concerning the existence of free will is essentially an evidential one rather than conceptual one; thus, lending itself to a probabilistic analysis. From an introduction to subjective probability, I moved my analysis to subjective Bayesianism applied to the free will proposition based upon the evidence of consensus – a *consensus gentium* argument.

I argue that evidence from consensus is only sound if the population selected is cognizant of the issues relevant to the question posed to it, in this case, *does free will exist?*, and the free will debate is both extensive and complex. Moreover, consensus itself is a belief orientated notion, and belief comes in degrees. Thus, I adjusted the evidential proposition in Bayes’ theorem to account for these two nuances. I also employed a likelihood ratio in the comparative form of Bayes’ theorem.

To ensure a well-defined probabilistic relationship between the free will proposition and its negation and the evidence of consensus, I introduced a set of auxiliary assumptions. These assumptions also had application to the prior belief in the free will proposition given background knowledge. Thus,

I examined five auxiliary assumptions: naturalistic realism, evolution, phenomenology, blame/liability and scholarly error. From Bayes' theorem, together with these auxiliary assumptions, I concluded that the posterior probability value of the free will proposition on the evidence would fall within the range [0.60,0.75]. However, for this result to be credible, such a theoretical analysis needs evidential support itself – Bayesian conditionalisation. Thus, I introduced the results of two surveys that supported this posterior value and concluded that the posterior probability value of the free will proposition on the evidence of consensus would fall within the range [0.60,0.75], so it is more likely than not that free will exists.

There are evidential challenges to the free will proposition that I referred to throughout the paper. However, a more serious challenge to the proposition loomed – *a priori* refutations. To reply to these refutations, I first provided my interpretation of free will that is congruent with the free will proposition h – *non-causal libertarianism*. This interpretation indicates a personal power possessed by free agents to make choices between alternative possibilities and to instantiate those choices. A power manifesting properties different to that of causation – I term this power *will-power* in contrast to *causal-power* – interrelated concepts that cannot be conflated. From this differentiation, it was clear that my argument would take on a compatibilist perspective, at odds with incompatibilism and event/agent-causal libertarianism. Given the threat of incompatibilism, I provided a brief insight into the likely neural mechanics of libertarianism – integrated information theory (IIT).

The application of Bayes' theorem presupposes that the probability of the proposition is not zero, and there are several *a priori* refutations of the free will proposition. I considered the primary ones and applied the concept of libertarian self-determination in defence of them, arguing that there is doubt as to their *a priori* status. I also cast doubt on the veracity of an instrumentalist perspective of reality that would have excluded a Bayesian analysis of free will.

Although there is strong evidence in support of the free will proposition, it does not ensure its truth. In fact, although my analysis demonstrates that the probability of the truth of the free will proposition is fairly high, this also implies there is a probability that it is false. The proposition is

defeasible, and new evidence, say from neuroscience and/or quantum science, may change this true/false-balance either way.

### References

- Bourdet, David. and David J. Chalmers. 2013. "What do Philosophers Believe?" *Philpapers*: <https://philpapers.org/archive/BOUWDP>
- Bourdet, David and David J. Chalmers. 2020. "Philosophers on Philosophy." *The 2020 PhilPapers Survey*: <https://survey2020.philpeople.org/survey/results/all>
- Bourdet, David. and David J. Chalmers. 2023. "Philosophers on Philosophy." *The 2020 PhilPapers Survey* Philosophers' Imprint 0: 0: 1-53 [BOUPOP-3 \(philarchive.org\)](https://philarchive.org)
- Cartwright, Nancy. 1999. *The Dappled World*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139167093>
- David, Chalmers. 2000. *What is a Neural Correlate of Consciousness?* In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, edited by Thomas Metzinger, 17-39. Cambridge, Massachusetts: The MIT Press.
- Earman, John. 1996. *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*. Cambridge, Massachusetts: MIT Press, first published 1992.
- Gillies, Donald. 1993. *Philosophy of Science in the Twentieth Century: Four Central Themes*. Oxford: Blackwell.
- Gillies, Donald. 2003. *Philosophical Theories of Probability*. London: Routledge.
- Glymour, Clark. 1980. *Theory and Evidence*. Princeton: Princeton University Press.
- Hacking, Ian. 1982. "Experimentation and Scientific Realism." *Philosophical Topics*, 13(1): 71-87. <http://dx.doi.org/10.5840/philtopics19821314>
- Hobbes, Thomas. 1841. *The Questions Concerning Liberty, Necessity and Chance*, In *The English Works of Thomas Hobbes*, vol. 5, edited by Sir William Molesworth, Bart. London: John Bohn, first published 1654-1656.
- Howson, Colin and Peter Urbach. 1991. *Scientific Reasoning: The Bayesian Approach*. Illinois: Open Court Publishing Company, first published 1989.
- Hume, David. 1985. *A Treatise of Human Nature*. London: Penguin Classics, first published 1739.
- Inwagen, Peter van. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.
- Inwagen, Peter van. 2002. *Free Will Remains a Mystery*, In *The Oxford Handbook of Free Will*, edited by Robert Kane, 158-177. New York: Oxford University Press.

- Ismael, Jenann. 2020. *How Physics Makes Us Free*. Oxford: Oxford University Press, first published 2016.
- Kuhn, Thomas S. 2012. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, first published 1962.
- Libet, Benjamin et al. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness Potential). The Unconscious Initiation of a Freely Voluntary Act." *Brain*, 106: 623–642.  
<http://dx.doi.org/10.1093/brain/106.3.623>
- List, Christian. 2019. *Why Free Will is Real*. Cambridge, Massachusetts: Harvard University Press.
- McDermott, Daniel. 2002. "Debts to Society." *The Journal of Political Philosophy* 10(4): 439-464. <http://dx.doi.org/10.1111/1467-9760.00160>
- McKay, Thomas and David Johnson. 1996. "A Reconsideration of an Argument against Compatibilism." *Philosophical Topics* 24: 113-122.  
<http://dx.doi.org/10.5840/philtopics199624219>
- Mele, Alfred R. 1995. *Autonomous Agents*. New York: Oxford University Press.
- Pereboom, Derek. 2014. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Pink, Thomas. 2019a. *Self-determination: The Ethics of Action*. Oxford: Oxford University Press, first published 2016.
- Pink, Thomas. 2019b. "Freedom, Power and Causation." *Organon F*, 26(1): 141–168. <https://doi.org/10.31577/orgf.2019.26109>
- Pinto, Anabela. 2022. *An Evolutionary Approach to the Adaptive Value of Belief in Evolutionary Psychology Meets Social Neuroscience*, edited by Rosalba Morese et al, Ch.2. London: IntechOpen.
- Putnam, Hilary. 1985. *Mathematics, Matter and Method*. Cambridge: Cambridge University Press, first published 1975.
- Shanahan, Murray. 2010. *Embodiment and the Inner Life: cognition and consciousness in the space of possible minds*. Oxford: Oxford University Press.
- Sober, Elliott. 2008. *Evidence and Evolution: The Logic behind the Science*. Cambridge: Cambridge University Press.
- Søvik, Atle O. 2018. *Free will, Causality and the Self*. Berlin: De Gruyter.
- Anil, Seth. 2021. *Being You*. London: Faber & Faber.
- Steward, Helen. 2014. *A Metaphysics for Freedom*. Oxford: Oxford University Press, first published 2012.
- Strawson, Galan. 1986. *Freedom and Belief*. Oxford: Clarendon Press.
- Tononi, Giulio et al. 2004. "An Information Integration Theory of Consciousness." *BMC Neuroscience*, 5:42. <https://doi.org/10.1186/1471-2202-5-42>
- Tononi, Giulio et al. 2023. "Only What Exists Can Cause: An Intrinsic Powers View of Free Will." arXiv. <https://doi.org/10.48550/arXiv.2206.02069>

Wisniewski, David et al. 2019. “Free Will Beliefs Are Better Predicted by Dualism than Determinism Beliefs across Different Cultures.” *PLoS ONE*, 14(9).

<https://doi.org/10.1371/journal.pone.0221617>

Wittgenstein, Ludwig. 1976. *Philosophical Investigations*. Oxford: Basil Blackwell, first published 1953.

# Chrysippus' Conditional Captured from a Non-Axiomatic Computer Program

Miguel López-Astorga\*

Received: 11 September 2023 / Accepted: 11 February 2024

*Abstract:* It is usually accepted that Chrysippus of Soli proposed the “connexivist view” of the conditional. I assume here that Chrysippus also supported the “inclusion view” and that differentiated between two kinds of conditionals: strong and weak conditionals. The latter assumptions allow me to link Chrysippus' interpretation of the conditional to a computer program such as Non-Axiomatic Reasoning System (NARS). The inclusion view enables to deem Stoic conditionals as inheritance relations in NARS. The distinction between strong and weak conditionals helps assign values of frequency and confidence such as those NARS inheritance relations have to Stoic conditionals.

*Keywords:* connexivist view; inclusion view; inheritance relations; non-axiomatic reasoning system; Stoic conditional.

## 1. Introduction

In the logic Chrysippus of Soli proposed, the conditional is not classical logic conditional. Chrysippus' view is often linked to the “connexivist” tra-

---

\* University of Talca, Chile

 <https://orcid.org/0000-0002-6004-0587>

 Institute of Humanistic Studies, Research Center on Cognitive Sciences, University of Talca, Talca Campus, Chile

 [milopez@utalca.cl](mailto:milopez@utalca.cl)





dition (e.g., O'Toole and Jennings 2004). It has been addressed from different approaches, even from the modern framework of modal logic (e.g., Lenzen 2019). What I try to do here is to consider Chrysippus' interpretation from a current non-axiomatic system coming from Artificial Intelligence (AI). That system is NAL (Non-Axiomatic Logic) (Wang 2013).

The advantage of that consideration is twofold. On the one hand, because NAL is a term logic, it means to relate Stoic logic to a term logic, and, therefore, to a logic akin to the Aristotelian. On the other hand, given that NAL enables to build NARS (Non-Axiomatic Reasoning System), that is, a computer program (see also, e.g., Wang 2006), it also makes it possible to relate Stoic logic to AI. The latter relation is very interesting as NARS is based on an assumption (see also, e.g., Wang 2011): The Assumption of Insufficient Knowledge and Resources (AIKR). This assumption allows NARS to work in a cognitive context akin to that habitual for the human mind. Hence, to link the interpretation of the conditional Chrysippus of Soli gave to NAL can lead to think about Stoic logic as a logic near the manner human beings draw conclusions.

The paper will have three parts. First, I will discuss the way Chrysippus of Soli understood the conditional. To clarify this point is important to determine to what extent the relation to NAL is justified. Then, I will describe the characteristics and components of NAL necessary to link both approaches. The last part will indicate how the relation between Chrysippus' view of the conditional and NAL can be provided. It will also show how, by virtue of the relation, the Stoic conditional can be captured from a computer program such as NARS.

## 2. The conditional in Stoic logic

It can be thought that the way Chrysippus of Soli interpreted the conditional is the way the Old Stoa understood it (O'Toole and Jennings 2004). This way is different from that of classical logic. As indicated by Cicero (*Academica*) and Sextus Empiricus (*Adversus Mathematicos, Pyrrhoniae Hypotyposes*) there was an intense debate about the characteristics of the sound or true conditional in the fourth century B.C. It is not clear whether the words 'sound' and 'true' were synonymous in Stoic logic (for a discussion

in the context of the debate, see, in addition to O'Toole and Jennings (2004), e.g., Mates (1953). However, because that does not have an influence on the develop of the present paper, I will assume that they were synonymous. Sextus Empiricus spoke about four different opinions in this way. One of them is that of Philo of Megara (Sextus Empiricus, *Adversus Mathematicos*, 8.113; *Pyrrhoniae Hypotyposes*, 2.110), which has been deemed as the account corresponding to modern classical logic and, accordingly, to the material view of the conditional (see also, e.g., Bocheński 1963). Nevertheless, that is none of the two accounts relevant here.

One of the accounts important for this paper is that claiming the need for a connection or bond between the two clauses, that is, the antecedent and the consequent. That connection is expressed as a fight between the opposite of the second clause, that is, the consequent, and the first clause, that is, the antecedent (Sextus Empiricus, *Pyrrhoniae Hypotyposes*, 2.111). Sextus Empiricus did not point out who the proponents of this account were. Nonetheless, following other ancient sources too (Cicero, *De Fato*, 12), most of the authors think that there is no doubt that this is Chrysippus' account (see also, e.g., Gould 1970). This view is often called the 'connexivist view' of the conditional (e.g., O'Toole and Jennings 2004).

While the connexivist view is hard to assume from propositional calculus, it is not from modern logic in general. For example, the relations to the strict implication (Lewis 1918) seem to be evident (for a discussion, see, e.g., Gould 1970); see also (Lenzen 2019), where links between the strict implication, the connexivist view, and Leibniz's work are provided). Nonetheless, to address one more interpretation of the conditional in the debate in the fourth century B.C. can be more interesting now. Other account Sextus describes is that of the philosophers considering the consequent to be included in the antecedent (Sextus Empiricus, *Pyrrhoniae Hypotyposes*, 2.112). This interpretation is often named the 'inclusion view' (e.g., (O'Toole and Jennings 2004)).

The inclusion view is relevant because it has been thought that it is not very different from the connexivist view (e.g., Kneale and Kneale 1962; Long and Sedley 1987). It has been said even that it is possible that Chrysippus offered the connexivist criterion in order to develop and clarify the inclusion criterion (O'Toole and Jennings 2004). Besides, the algebra

Leibniz proposed, which also presented 'containment relations' between two clauses, has been related to the connexivist view as well (Lenzen 2019).

All of this leads to think that there are reasons for assuming that the criterion Chrysippus of Soli supported includes both the connexivist view and the inclusion view. Thus, I will accept that assumption in this paper. To my aims here, only one more point needs to be reviewed. It seems that Chrysippus also distinguished between strong and weak conditionals (e.g., Sedley 1984). The thesis is based on sources such as Cicero (*Academica*, II.99-100; *De Fato*, 15-16). It appears to be that, when a conditional relation is coherent with Chrysippus' account, it should be worded in natural language as a conditional. But if we are not absolutely sure that the denial of the second clause is not compatible with what the first one states, it should be worded as a conjunction that is negated, and in which one of the conjuncts is negated too. The distinction Sedley supports is also addressed in works such as López-Astorga (2015a).

Let A be the antecedent in a conditional relation. Let B be its consequent. The idea in the previous paragraph means that, in a natural language such as English, if the conditional relation is clear (according to Chrysippus' view), the sentence indicating that relation should be "If A then B." If there is uncertainty on the conditional relation, the correct sentence should be "It is not the case that A and not B."

This thesis shows the distance existing between Stoic logic and modern propositional logic. In the latter, conditionals and negated conjunctions are interchangeable. Nevertheless, beyond the fact that the distance is obvious and has already been highlighted in different senses (e.g., Bobzien 1996), we have reasons for accepting the difference between strong and weak conditionals. Works such as that of Sedley (1984) illustrate the reasons. Therefore, I will assume the difference here as well. The difference will be essential for the relation I will establish between Stoic logic and NAL below.

### 3. Inheritance relations in NAL

NAL is a term logic (Wang 2013). It has several rules and a grammar so broad that it cannot be described in entirety in this paper. I will only

refer to the components of NAL necessary to provide a relation to Chrysippus' logic.

In NAL, there subjects and predicates linked by means of inheritance relations. An inheritance relation between a subject 'S' and a predicate 'P' is expressed in NAL as (INH1).

(INH1)  $S \rightarrow P$

(INH1) literally appears and is explained in (Wang 2013, Definition 2.2).

The concepts of extension and intension define what inheritance relations are. All the subjects had by a predicate are the extension of the predicate. On the other hand, all the predicates had by a subject are the intension of the subject (Wang 2013, Definition 2.8). For example, let us suppose the following inheritance relations:

ARISTOTLE  $\rightarrow$  PHILOSOPHER

CHRYSIPPUS  $\rightarrow$  PHILOSOPHER

CICERO  $\rightarrow$  PHILOSOPHER

ARISTOTLE  $\rightarrow$  GREEK

ARISTOTLE  $\rightarrow$  HUMAN BEING

If only these inheritance relations are considered (as a subset of all of the inheritance relations), it can be said that,

- "Aristotle," "Chrysippus," and "Cicero" are the extension of "philosopher."
- "Philosopher," "Greek," and "human being" are the intension of "Aristotle."
- "Philosopher" is the intension of "Chrysippus."
- "Philosopher" is the intension of "Cicero."
- "Aristotle" is the extension of "Greek."
- "Aristotle" is the extension of "human being."

But AIKR is essential in NAL. This is because we cannot be sure about the inheritance relations. For this reason, NAL inheritance relations have both a value of frequency "f" and a value of confidence "c." In this way, inheritance relations are not represented as (INH1), but as (INH2).

$$(INH2) \quad S \rightarrow P (f, c)$$

(INH2) appears and is explained in, for example, (Wang 2013, Definition 3.8).

The formulae to calculate  $f$  and  $c$  (as they are literally in (Wang 2013, Definition 3.3)) are:

$$\text{Frequency: } f = W^+/W$$

$$\text{Confidence: } c = W/(W + K)$$

Where “ $W^+$ ” expresses “positive evidence,” “ $W$ ” is “total evidence,” and “ $K$ ” is a constant whose value is generally 1.

Although the definition of “ $W^+$ ” is more complex (see Wang 2013, Definition 3.2), for the goals of the present paper, it is enough to assume that, given this inheritance relation,

$$\text{PHILOSOPHER} \rightarrow \text{GREEK}$$

“Aristotle” is positive evidence, but “Cicero” is not. “Cicero” is part of the total evidence, but not positive evidence.

NAL includes different rules. However, to deal with the deduction rule as an example can suffice here. Following the schema of the latter rule in (Wang 2013, Chapter 4), this inference would be a case of deduction:

$$\text{GREEK} \rightarrow \text{PHILOSOPHER} (0.8, 0.83)$$

$$\text{LOGICIAN} \rightarrow \text{GREEK} (0.6, 0.91)$$

---


$$\text{Therefore, LOGICIAN} \rightarrow \text{PHILOSOPHER} (0.48, 0.36)$$

The numbers in the three statements are explained as follows:

- We have checked five Greek people, and four of them are philosophers. Hence,  $f = 4/5 = 0.8$ ;  $c = 5/6 = 0.83$ .
- We have checked ten logicians, and six of them are Greek. Hence,  $f = 6/10 = 0.6$ ;  $c = 10/11 = 0.91$ .
- The frequency and confidence of the conclusion are calculated following the formulae indicated in (Wang 2013, Table 4.7). For frequency, the frequencies of the two premises are multiplied. For confidence, four numbers are multiplied: the frequencies and the confidences of the two premises. Hence;  $f = 0.8 \times 0.6 = 0.48$ ;  $c = 0.8 \times 0.83 \times 0.6 \times 0.91 = 0.36$ .

Much more inference rules are in NAL. Nevertheless, this example of deduction is enough to present how the system works.

#### 4. Stoic conditional and NAL

This is not the first paper trying to relate Stoic logic to the way human beings think. For example, there are works linking the latter logic to contemporary psychology theories such as the theory of mental models (e.g., Johnson-Laird 2023; Khemlani and Johnson-Laird 2022); an example of work linking Stoic logic to the theory of mental models is López-Astorga (2015b) and the mental logic theory (e.g., O'Brien 2014 and O'Brien 2021); an example of work linking Stoic logic to the mental logic theory is (López-Astorga 2015b). On the other hand, links between the connexivist view and a term logic are also to be found in the literature (e.g., Lenzen 2019). Nonetheless, as far as I know, there are not any works providing relations between Stoic logic and systems of AI such as NARS. I will establish a relation between Chrysippus' view of the conditional and NAL, which is the foundation of NARS, in this section.

To relate a Stoic conditional to a NAL inheritance relation is not difficult if it is assumed, as I did above, that Chrysippus' criterion is both the connexivist and the inclusion criteria. As indicated, there are reasons for the assumption. One of the reasons is particularly interesting. As also mentioned, Leibniz's algebra has been related to the connexivist view (Lenzen 2019). That algebra includes containment relations. In it, it is possible to state that "a term A contains a term B." In addition, the algebra also resorts to the concepts of extension and intension. So, one might think that the theoretical arguments to support that there are not many differences between the connexivist and the inclusion criteria are strong.

From this point of view, if the antecedent of a conditional includes the other clause, that means that the latter is part of the intension of the antecedent, and the antecedent is part of the extension of the second clause. Thus, if the antecedent of a conditional consistent with the inclusion account is deemed as "S" and its consequent is denominated "P," those clauses satisfy (INH1).

NAL inheritance relations are actually represented as (INH2). But this is not a problem if the distinction between strong and weak conditionals (e.g., Sedley 1984) is taken into account. If the frequency of an inheritance relation is not 1, that inheritance relation cannot correspond to a strong conditional. It can only capture a weak conditional, that is, a conditional about which we are not sure whether it is compatible with Chrysippus' interpretation or not. This is because an inheritance relation has value of frequency 1 only when all the cases reviewed are positive evidence, that is, are part of  $W^+$ . As far as confidence is concerned, it is obvious that it should be high as well. However, a value of confidence 0.9 is already achieved only checking ten cases ( $c = 9/10 = 0.9$ ). This is important, since, given the formulae above, the value of confidence (and frequency) of any inheritance relation can only vary between 0 and 1.

Therefore, in an inheritance relation such as (INH2), if  $f = 1$ , then it should be formulated as "If S then P." If  $f < 1$ , then it should be couched as "It is not the case that S and not P."

As it can be noted, the computational processing of my proposal is really easy. Although the habitual languages for NARS at present are Prolog and Java (Wang 2013), I can offer a trivial example of code in Common Lisp (LispWork Personal Edition). The code is very simple, but it can differentiate between strong and weak conditionals by virtue of their values of frequency. It is the following:

```
(defun Chrysippus (L1 L2 N)
  (IF (= N 1) (append '(if) L1 '(then) L2)
      (append '(it is not the case that) L1 '(and not) L2)))
```

In Common Lisp, "defun" allows defining functions. Here, the function created is named "Chrysippus." "L1," "L2," and "N" are the three variables of function "Chrysippus." "L1" corresponds to the first clause or antecedent. "L2" refers to the second clause or consequent. Variable "N" is the number for frequency. I resort to function "IF," which establishes a condition:  $N = 1$  "(= N 1)." If that condition is correct, what is indicated between the next brackets, that is, "(append '(if) L1 '(then) L2)," is run. Function "appends" joins "if" + L1 + "then" + L2. If the initial condition is not correct, that is,  $n \neq 1$ , what is in the third line, that is, "(append '(it is not the case that)

L1 ‘(and not) L2)’ is run. In the latter situation, “append” binds “it is not the case” + L1 + “and not” + L2.

If, for example, we write,

> Chrysippus ‘(dog) ‘(MAMMAL) 1

The system will return:

IF DOG THEN MAMMAL

If the information I give is,

> Chrysippus ‘(dog) ‘(dalmatian) 0.9

The answer will be,

IT IS NOT THE CASE THAT DOG AND NOT DALMATIAN

The first example corresponds to a conditional coherent with Chrysippus’ account. If an animal is not a mammal, that animal cannot be a dog. For this reason, the sentence is written as strong conditionals in Stoic logic. In the second example, there is no consistency with Chrysippus’ view. An animal may not be a dalmatian and keep being a dog. So, in the latter example, the formulation corresponding to weak conditionals in Stoic logic is that to be used.

The drafting of both function ‘Chrysippus’ and the information given in the examples can be improved. For instance, in the case of the information written, the data could be those:

> Chrysippus ‘(this is a dog) ‘(this is a mammal) 1

What the system would return would be,

IF THIS IS A DOG THEN THIS IS A MAMMAL

But the examples indicated make the point the present paper is intended to do.

## Conclusions

It is possible to relate the logic Chrysippus of Soli proposed to a computer program making inferences in a similar manner as human beings.



Probably, there are many ways to do that. I have tried to do it resorting to NAL here.

We only need to assume two points. First, the connexivist and the inclusion criteria Sextus Empiricus distinguished are the same (or, at least, very alike) and, accordingly, both of them typify Chrysippus' account of the conditional. Second, Stoic logic differentiates between strong and weak conditionals. The first ones should be phrased as conditionals are habitually expressed. The second ones should be built by means of conjunctions, which in turn should be negated and have one of their clauses negated too.

This allows relating Stoic conditionals to NAL inheritance relations. The antecedent of a Stoic conditional can be deemed as the subject in a NAL inheritance relation. The consequent of that very conditional can be deemed as the predicate in that very inheritance relation. This is because, both in Stoic conditionals and in NAL inheritance relations, the first clause encloses the second one. The concepts of extension and intension enable to see this.

NAL inheritance relations have values of frequency and confidence associated. Hence, if an inheritance relation does not have the highest value of frequency (i.e.,  $f < 1$  in it), the Stoic conditional corresponding to it cannot be strong. Only the conditionals with  $f = 1$ , that is, with the highest value of frequency, can be considered strong conditionals in Stoic logic.

NAL is the logical skeleton of NARS. So, the arguments above allow linking Stoic logic to an AI program deriving conclusions in a similar manner as people do. Both Stoic logic and NARS include much more components than those addressed here. Nevertheless, to find a connection between the view of the conditional Chrysippus of Soli supported and NAL inheritance relations already shows that there is, at a minimum, one link between the two systems. Further research should try to relate more components from both frameworks.

## Funding

PIA Ciencias Cognitivas, Centro de Investigación en Ciencias Cognitivas, Instituto de Estudios Humanísticos, Universidad de Talca.

Proyecto “Dialéctica virtuosa de la educación: De la fundamentación filosófica y antropológica a implementaciones concretas”, código 15/I122 SAL127/19, UNMdP.

## References

- Bobzien, Susanne. 1996. "Stoic Syllogistic." In *Oxford Studies in Ancient Philosophy*, 133-192. Oxford: Clarendon Press.
- Bocheński, Józef Maria. 1963. *Ancient Formal Logic*. Amsterdam: North-Holland.
- Gould, Josiah B. 1970. *The Philosophy of Chrysippus*. Albany, New York: State University of New York Press.
- Johnson-Laird, Philip N. 2023. "Possibilities and Human Reasoning." *Possibilities Studies & Society*, 1(1-2): 105-112. <https://doi.org/10.1177/27538699231152731>
- Khemlani, Sangeet, and Philip N. Johnson-Laird. 2022. "Reasoning about proper-  
ties: A computational theory." *Psychological Review*, 129(2): 289-312.  
<https://doi.org/10.1037/rev0000240>
- Kneale, William and Kneale, Martha. 1962. *The Development of Logic*. Oxford: Oxford University Press.
- Lenzen, Wolfgang. 2019. "Leibniz's Laws of Consistency and the Philosophical Foundations of Connexive Logic." *Logic and Logical Philosophy*, 28: 537-551.  
<http://dx.doi.org/10.12775/LLP.2019.004>
- Lewis, Clarence Irving. 1918. *A Survey of Symbolic logic*. Berkeley: University of California Press.
- Long, Anthony A., and Sedley, David N. 1987. *The Hellenistic Philosophers*. Cambridge: Cambridge University Press.
- López-Astorga, Miguel. 2015a. "The Relationship between Conditionals and Denied Conjunctions in Stoic Logic." *Analele Universitatii din Craiova, Seria Filosofie* 36(2): 82-94.
- López-Astorga, Miguel. 2015b. "Chrysippus' *indemonstrables* and mental logic." *Croatian Journal of Philosophy*, 15(43): 1-15. <https://doi.org/10.5840/croat-jphil20151511>
- Mates, Benson. 1953. *Stoic Logic*. Berkeley and Los Angeles: University of California Press. <https://doi.org/10.1525/9780520349070>
- O'Brien, David P. 2014. "Conditionals and Disjunctions in Mental-Logic Theory: A Response to Liu and Chou (2012) and to López-Astorga (2013)." *Universum*, 29(2): 221-235. <http://dx.doi.org/10.4067/S0718-23762014000200015>
- O'Brien, David P. 2021. "Natural Logic." In *The Handbook of Rationality*, 215-224. Cambridge, Massachusetts: The MIT Press. <https://doi.org/10.7551/mit-press/11252.003.0021>
- O'Toole, Robert R., and Jennings, Raymond E. 2004. "The Megarians and the Stoics." In *Handbook of the History of Logic, Volume 1. Greek, Indian and Arabic Logic*, 397-522. Amsterdam: Elsevier. [https://doi.org/10.1016/S1874-5857\(04\)80008-6](https://doi.org/10.1016/S1874-5857(04)80008-6)

- 
- Sedley, David. 1984. "The Negated Conjunction in Stoicism." *Elenchos*, 5: 311-317.
- Wang, Pei. 2006. *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.  
<https://doi.org/10.1007/1-4020-5045-3>
- Wang, Pei. 2011. "The Assumptions on Knowledge and Resources of Rationality." *International Journal of Machine Consciousness*, 3(1): 193-218.  
<https://doi.org/10.1142/S1793843011000686>
- Wang, Pei. 2013. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. Singapore: World Scientific. <https://doi.org/10.1142/8665>

## The Alethic Status of Contradictions in Fictional Discourse

Vladimir Vujošević\*


Received: 21 September 2023 / Revised: 19 February 2024 / Accepted: 22 February 2024

*Abstract:* Whether contradictions could be “true in fiction” has become an unavoidable topic in the debates on the bounds of fictionality. This paper claims that genuine contradictions in fiction are far more infrequent phenomena than is usually claimed. The majority of cases that have been put forward as examples of contradictory fictions can be convincingly understood either as instances of rhetorical pseudo-contradictions or (in the case of the so-called “forking-path“ narratives) as disjunctions of possible outcomes rather than contradictory conjunctions of simultaneously enacted exclusive scenarios. The only philosophically interesting category of contradictory fiction would be the one in which a single “root” contradiction is explicitly affirmed as the central element of the story (in the third-person, authoritative narrative voice). The paradigmatic example would be the revised version of Graham Priest’s “Sylvan’s Box” this paper presents. However, it could be argued that the problem with such narratives is that they unsuccessfully attempt to perceptually code what remains exclusively propositional content. They are, thus, fatally under-described, and the truth of the contradictory proposition fails to be adequately established in fiction. The idea that one can posit contradictions as fictional facts is

---

\* University of Donja Gorica

 <https://orcid.org/0000-0003-0078-5103>

 Faculty of Philology, Oktoih 1, ME-81000 Podgorica, Montenegro

 [vladimir.vujošević@udg.edu.me](mailto:vladimir.vujošević@udg.edu.me)



based on oversimplified notions of narrative conventions and truth in fiction.

*Keywords:* truth in fiction; contradiction; the law of non-contradiction; “Sylvan’s Box.”

## 1. The Principle of Poetic Licence (PPL) vs. the Law of Non-Contradiction (LNC)

In the debates on the nature of fiction, two claims seem to be commonly accepted:

- (i) We take “various propositions to be true according to a particular [fiction]” (Nolan 2021, 14).
- (ii) What is true in any given fiction is not necessarily bound by the standard of actuality.

The first claim has become familiar under the phrase “truth in fiction.” It’s a way of distinguishing fictional facts from, e.g., (possibly misguided) beliefs of various fictional characters. In *Don Quixote*, to offer Doležel’s example, it’s not true that the protagonist fights monstrous giants, but it’s true that he charges at windmills (delusionally convinced that these are the monstrous giants). Since Don Quixote tilting at windmills is an event that “really” took place in the “fictional world of [Cervantes’ novel]” (Doležel 1998, 149), we can describe the proposition that asserts it as being “true in fiction.”

The second claim simply means that what is true in fiction “may [...] deviate enormously from the actual world” (Routley 1979, 6). Although we are aware that things like sapient teapots (*The Beauty and the Beast*) and sloth bears endowed with the command of human language (*The Jungle Book*) are physically or biologically impossible, we, nevertheless, concede that they are fictional possibilities.

Does this mean that authors can make “anything whatsoever true in their fictions” (Xhignesse 2016, 149)? Some people have argued precisely so. The idea is encapsulated in what Harry Deutsch has dubbed *the principle of poetic license* (PPL): for any proposition *p*, one can produce a fiction

in which  $p$  is true.<sup>1</sup> According to this approach, the authorial “sayso” (Nolan 2021, 16) is enough to make a proposition true in fiction.

However, there is one obvious problem with the PPL: the price for upholding it seems to be “simply too high” (Xhignesse 2016, 161). If the PPL holds, then even logical contradictions could be true in fiction. We can accept that “impossibility with respect to reality is significantly different from impossibility with respect to [fiction]” (Ashline 1995, 231). For me to transform into a talking ashtray would be physically impossible in the world as it is. However, we can easily imagine logically possible worlds in which such transformations routinely take place. Such bizarre events are still logically possible. The laws of logic, however, are not just dispensable elements of possible worlds but are the prerequisite of their very possibility.<sup>2</sup>

---

<sup>1</sup> For various formulations of Deutsch’s principle, see (Xhignesse 2016, 149) and (Hanley 2004, 121).

<sup>2</sup> Here, one terminological elucidation concerning “worlds” is called for. For instance, imagine a student taking an English literature exam and being asked how many people did Frankenstein’s creature murder. She’s well aware that “Frankenstein’s creature” has never accurately referred to anything in the actual world, and she may reasonably believe that the existence of such a being would even be physically impossible. However, the question doesn’t seem meaningless since there appears to be a correct answer to it. This is because we understand all such questions as being discretely prefixed by an operator: “In work of fiction  $f$ , ...” (Lewis 1978, 38). According to Lewis, our engagements with fictions require us to agree to an act of “pretense” (1978, 40). The storyteller pretends “to be telling the truth about matters whereof [she] has knowledge” (Lewis 1978, 40). When Mary Shelley produced her novel, she wanted her readers to make-believe that it was a factual report “rather than fiction” (Lewis 1978, 40). Since many of the claims in the novel are obviously “misdescriptions” (Kroon, Voltoni 2019) of the actual world (it’s safe to say, e.g., that there was never an 18th-century Swiss natural philosopher who successfully reanimated a creature comprised of discarded body parts with the use of a voltaic pile), we should understand them as descriptions of some possible worlds in which “the act of storytelling” is “what here it falsely purports to be: truth-telling” (Lewis 1978, 40). Simply put, to say that a claim is “true in fiction” means that it is true in some possible worlds described by the fiction in question. According to the Lewisian approach, the act of pretense involves possible worlds semantics. This model has been influential in both the analytic philosophy of fiction and contemporary narratology and it has been employed to a large extent in discussions of logically impossible fictions by authors such as (Alber 2016, Badura and Berto 2019, Doležel 1998, Fořt 2016, Priest 1997, Ryan 2019,

*The law of non-contradiction* (LNC) is, according to Aristotle's famous definition, "firmest of all" logical laws and "non-hypothetical": it asserts that for all propositions of the type  $p$  and  $\sim p$  to be true "simultaneously and in the same respect is [absolutely] impossible" (*Metaphysics*  $\Gamma$  1005b15-20). There is no possible world in which  $\sim(p \wedge \sim p)$  does not obtain: the LNC is "true at all members of any set of worlds, and so is true in every fiction" (Hanley 2004, 117) since fictions are descriptions of possible, unactualized states of affairs. There is no fiction (since there's no possible scenario) in which the LNC could be violated. We cannot have the PPL and the LNC both. It seems that we have reached the frontier of truth in fiction. Fictional possibilities are associated "with logical laws" (Alber 2016, 30). Try as he might, one cannot create a fiction in which propositions of the type " $p \wedge \sim p$ " are true. It appears that we should reject the PPL. So why not stop here and call the case settled? There are two main reasons.

First, various authors have claimed that there is a "special range of [fictional] possibility" that is wider than "the range of [logical possibility] philosophers have tended to consider" (Nolan 2015, 62). To prove this point, Graham Priest has constructed a much-debated fiction called "Sylvan's Box." Two philosophers are going over the archive of their deceased friend and mentor when suddenly they uncover a cardboard box with the inscription "Impossible object" on its lid. After opening it, they are stupified by what seems to be an explicit, observable contradiction: "The box was absolutely empty, but also had something in it. Fixed to its base was a small figurine carved of wood" (Priest 1997, 576). After initially questioning their sanity, the philosophers carefully reexamine the box, trying to come to terms with the far-reaching consequences of their finding on logic: "This was no illusion. The box was really empty and occupied at the same time. The sense of touch confirmed this" (Priest 1997, 576).

One could instantly pose a simple question: "What's true in this fiction?" (Berto, Jago 2019, 246). It seems that "the most straightforward reading" (Berto, Jago 2019: 246) would be to say that it is *true* in the story that there exists a box that is simultaneously empty and non-empty, or to

---

etc.). Therefore, I will use the Lewisian possible worlds framework throughout much of this paper to address some of these arguments.

put it more formally:  $(\exists x) (Fx \wedge \sim Fx)$ .<sup>3</sup> There's no way around it: the whole point of the story is the discovery of an object that violates the LNC. The assumption that the LNC doesn't obtain in this fiction is "essential for understanding [it]" (Berto, Jago 2019, 246). To comprehend the fiction in any other way would be to misread it.<sup>4</sup>

Second, one could argue that "Sylvan's Box" is not just an odd "philosophical corner-case" but, in a way, a "fictional commonplace" (Xhignesse 2016, 152). In various fictional works, the violation of the LNC seems to be "a central poetical device" (Ronen, qtd. in Ryan 2019, 67). The fact that "the logically impossible" is "a salient feature" (Ashline 1995, 215) of many fictional narratives could be seen as further evidence for the PPL. After all, we "have these stories, we read them, we understand them" (Xhignesse 2021, 3170) and engage with them without substantial problems. It appears that, in our ordinary practice, we accept contradictions as fictional facts that are somehow *true* in these particular storyworlds.

Fowles' *French Lieutenant's Woman* is an often-mentioned example. The story, set in 19th-century England, portrays the intense relationship between Charles Smithson, a young gentleman soon to be married to a rich heiress, and Sarah Woodruff, an ostracized Victorian "fallen woman" whose social standing was destroyed by a short-lived romance with an ill-reputed French military man. However, the novel has three "logically incompatible endings" (Alber 2016, 173) that appear to be simultaneously enacted. In one of these endings, Charles, after a brief dalliance, breaks up with Sarah, never to meet her again, and marries his fiancée. The affair with Sarah becomes an unpleasant minor episode in his otherwise respectable life. In the second ending, Charles calls off the engagement, not without a public

---

<sup>3</sup> For similar notations see, (Horn 2018).

<sup>4</sup> Such an opinion was upheld not just by Priest himself (who is a prominent advocate for dialetheism, the idea that contradictions can sometimes be true) but also by Fořt, who seems to argue that "the notion of an impossible possible world within fictional discourse" could be "profitable" (2016, 51). A similar position is maintained by Barto and Jago (2019) and narratologists such as Jan Alber, who claims that it's possible to successfully embed "logically impossible elements [...] in the world of fiction" (2009, 80), and Ruth Ronen who distinguishes possible worlds from fictional worlds, since the latter "can be [logically] impossible" (qtd. in Bell, Ryan 2019, 13).



scandal, and eventually ends up with Sarah, fathering a child with her. In the third ending, Charles breaks up the engagement but is ultimately turned down by Sarah, who appears to be no longer interested in him. It seems that the answer to a simple question: “Did Charles and Sarah get married in Fowles’ novel?” would imply a contradiction: “Yes and no.”

Robert Coover’s short story “The Babysitter” is another notable fiction that combines “multiple, mutually incompatible plotlines” (Alber 2016, 175). The narrative begins with the Tuckers hiring a babysitter to watch over their children while they attend a party. However, after the babysitter arrives, the story breaks down into a sequence of incompatible episodes (that are seemingly taking place all at once). In one plotline, Mr. Tucker returns alone from the party to make advances to the babysitter, while in the other (parallel one), he remains at the party with his wife until the end. At one point, things go tragically awry: the babysitter is distracted by a film on TV, and one of the children chokes on a diaper pin. However, the subsequent paragraph suggests that, thankfully, nothing memorable has happened that night: the babysitter prepares the children for bed, watches TV, and then dozes off, waiting for the Tuckers to return. In one storyline, the Tuckers return from the party to find the house in perfect order, and the babysitter leaves. But in another, they discover that their house has become a crime scene: the babysitter has been murdered (by her boyfriend and his friend). According to Alber, the story is “logically impossible” (2016, 25) and violates the LNC since contradictory propositions (like “The babysitter is murdered” and “She is not murdered”) are simultaneously true in it.

Caryl Churchill’s play *Traps* is also frequently invoked as yet another example of fiction that “does not conform to the [LNC]” (Alber 2009, 83). The play features various characters whose lives and mutual relationships substantially differ from scene to scene. In the first act, Albert and Syl are a couple, and they have a baby, but in the subsequent one, they never had children and they even converse about the prospects of raising a child together someday. In another entrance, Albert is completely absent from the picture (as if he never existed as a substantial presence in Syl’s life), and Syl and Jack are lovers. A while later, however, it’s suddenly suggested that Albert and Jack were a couple all along, and Syl was only their friend. Eventually, Albert commits suicide but is, nevertheless, alive and well

afterward. It seems that the characters are going through various contradictory experiences at the same time. Things are simultaneously happening and not happening in the play: Syl (e.g.) is married to Albert but is also not married to him. She has a child but also doesn't have one. She is in a relationship with Jack but also not in a relationship with him, etc. Churchill herself, in "Performance Notes," compares the play to "an impossible object" that can be actualized only "on stage, but [in] no other reality" (1985, 71). It seems that she maintains that contradictions could be realized in the realm of fiction. Fictionality, for Churchill, seems unbound by the LNC.

Other much-debated examples of contradictory fictions include Danielewski's *House of Leaves* and Calvino's *The Nonexistent Knight*. *House of Leaves* is a frame tale that employs the gothic convention of the found manuscript: Johnny Truant, the narrator of the novel, discovers an alleged documentary record, compiled by a man called Zampano, that describes a series of uncanny events surrounding a house inhabited by a famous photographer Will Navidson and his family. All sorts of anomalies are taking place inside the house (e.g., its interior appears larger than its exterior, new chambers suddenly materialize, etc.). However, according to Alber (2016, 188), at least one of these eerie disturbances seems to be logically impossible. A haunted dark hallway mysteriously emerges on the north wall of the living room of the Navidson house, only to be subsequently asserted that the same uncanny hallway has always been located only on the west wall. Thus, according to the Zampano record, the hallway appears to be located on the north wall and not on the north wall (at the same time). This incongruity seems to produce a "logically impossible spatial parameter," violating thus "the [LNC]" (Alber 2016, 188).

In Calvino's short novel, we are introduced to Agilulf, Charlemagne's paradoxical paladin. While inspecting his troops, Charlemagne's attention is drawn to a strange knight whose body and face are thoroughly concealed by military gear. When questioned about the reasons for the insolency of hiding his face in front of the emperor, the knight calmly responds that he doesn't exist and, raising the visor of his helmet, reveals the gaping emptiness inside. Everyone (including the emperor himself) comes to accept the contradictory fact that Agilulf "doesn't exist" (Calvino 2012, 6) while simultaneously possessing various properties of existence (like chivalrously

riding to the rescue of distressed maidens). It appears that the knight doesn't exist, yet exists (since existence seems to be the necessary prerequisite of, e.g., wielding a sword and fighting in a battle).

Following Xhignesse (2016, 152) and Berčić (2021, 167), we can call the authors who claim that truth in fiction is not bound by the laws of logic (such as the LNC) *impossibilists*. According to this position, logical impossibilities are “the very possibility of fiction” (Alber 2016, 3). The most plausible explanation of (e.g.) “Sylvan's Box,” according to the impossibilist position, is the one in which “Priest has the right belief, and there actually is a fempty [both full and empty] box, without trivialization” (Badura, Berto 2019, 188).

Now, the author who wants to preserve the LNC “as an important background principle” (Xhignesse 2016, 161)—let's call him, following Xhignesse and Berčić, *the possibilist*—would need to provide a model for dealing with contradictory fictions according to which whenever a contradictory claim is put forward within the fictional discourse it will fail to automatically convert into a fictional fact. There are two promising strategies one could employ in such a venture.

## 2. The LNC preservation strategies

According to the possibilist position, when confronted with any contradictory claim in fiction, the reader can simply argue it off by claiming that either (a) it's not really a contradiction (i.e., there is a plausible and convincing non-contradictory explanation of what is happening) or (b) that the contradiction is claimed but not really achieved (e.g., a case can be made that the conveyor of information is unreliable). Matravers (2014, 131) names these (a) the *reconciliation* and (b) the *rejection* strategies. In the same way, Johnny Truant, the narrator of *House of Leaves*, after he stumbles upon contradictory information in a manuscript he reads (concerning the location of the uncanny hallway that mysteriously appeared in the Navidson house), engages in the following interpretative process: “Maybe there's some underlying logic to the shift. Maybe it's a mistake. [Heck] if I know” (Danielewski 2000, 970). What is signaled here is a natural way of dealing with contradictions in fiction. Either such claims are simply

erroneous (they are *mistakes* that indicate the narrator's unreliability), or there's some possible and satisfying *underlying* explanation for the seeming contradiction. If someone wants to create a fictionally true violation of the LNC, he must eliminate the plausibility of both reading strategies. This seems to be a serious challenge for the impossibilist author.

(a) "Reconciliation" Strategies

*Rhetorical (Pseudo-)Contradictions.* Not every phrasal form of contradiction is genuinely contradictory. We should not take all such sentences at face value since they can be part of the metaphorical use of language one can often encounter in fictional discourse. When, for example, Agilulf in *The Nonexistent Knight* is described in a contradictory fashion—as “one who exists without existing” (Calvino 2012, 14)—we should be wary of understanding such a claim literally, especially if the text offers valuable clues for a non-literal reading. After carefully inspecting the novel, one would notice that the term “nonexistent” is contrasted with “possessing a body”: being a “nonexistent” entity, Agilulf feels alone in “the realm of bodies” (Calvino 2012, 9). Devoid of physical substance, he doesn't know what it feels like to “shut one's eyes,” so he is envious of “the faculty of sleep possessed by people who existed” (Calvino 2012, 8-9). The description of Agilulf as “non-existent” seems to belong to a figurative, hyperbolic use of language that doesn't commit us to genuine contradictions. It's an imprecise, poetically provocative way of saying that Agilulf is “disembodied,” which then “scratches a different itch altogether, with different epistemic standards” (Xhignesse 2021, 3179).

“[W]hen you need to say something vividly,” you should “say it with a contradiction” (Sorensen 2002, 353). If I were to describe Sherry Levine's 1981 piece of appropriation art *After Walker Evans* as *originally unoriginal*, such a claim would semantically mimic the basic logical form of contradiction ( $p \wedge \sim p$ ), but it would not be one since there is a consistent meaning to it.<sup>5</sup> These rhetorical pseudo-contradictions are catchy phrases, like oxymorons, and one only needs to “sharpen” and “precisify” a “vague

---

<sup>5</sup> To claim that *After Walker Evans* is *unoriginal* suggests that Levine basically re-photographed and displayed Evans' 1935 work *Alabama Tenant Farmer Wife*. To describe it as *original* would mean that, unlike Evans, who was portraying “the

predication” (Horn 2018) for the sense of contradiction to dissipate. Thus, it seems that Alber is mistaken in claiming that (for example) the well-known phenomenon of the post-death narration (when the narrator is dead but nevertheless capable of telling a story) violates the LNC because such narrators “are alive [...] and not alive at the same time” (2019, 162). If we recall the influential Aristotle’s understanding of the LNC in *Metaphysics*, we can see that he mentions an important “qualification” (1005b 19-20) governing the principle: contradictory predicates cannot hold for the same subject “at the same time, and in the same respect” (as translated in Horn 2018). Hence, there’s no violation of the LNC if we assume that someone is alive and not alive in different respects (e.g., one can be physically *not alive* and “spiritually” *alive*). Thus, the first step in the possibilist argumentation would be to detect whether we are actually dealing with rhetorical (pseudo-)contradictions. In such cases, no violation of the LNC is achieved since contradictory claims are not affirmed in the same regard.

“*Slip-up*” *Contradictions*. An oft-repeated example of fictional contradiction concerns the “location of Watson’s old war wound” (Lewis 1978, 46) in Sherlock Holmes stories. Watson had only one wound, but some fictional accounts in the Holmes canon situate it on his shoulder and others on his leg. However, such contradictions are uninteresting since they are only incidentally part of fiction. They are authorial blunders that are merely unintentional interruptions in the fictional going-on. When Robinson Crusoe strips naked to swim to a shipwreck and then fills his pockets with the provisions he finds there, there’s no good reason to assume that Defoe’s novel describes a logically contradictory world in which one can be simultaneously naked and not naked. Such incidents are best understood as “slip-up[s] on the author’s [or the narrator’s] part” (Hanley 2004, 113). Thus, it would seem that Alber is mistaken when he assumes that *House of Leaves* has a “logically impossible [plotline]” (2016, 188). The contradiction regarding the position of the hallway could simply be attributed to the narrator’s misstep, which can suggest his unreliability. No contradictory state of affairs is thus generated but merely an ambiguity is created concerning the precise location of the hallway. Such

---

suffering of ordinary people during the depression in America’s Deep South” (Hudson Hick 2017, 128), Levine was dealing with a completely different issue of originality and authorship.

inconsistencies should be treated as “special case[s] of indeterminacy” (Hanley 2004, 117), not contradiction.

*Lewis’ Method of Union.* What about fictions like *The French Lieutenant’s Woman*, “The Babysitter,” and *The Traps*? It seems that, in these texts, contradictory storylines are simultaneously enacted, not by chance, but quite deliberately. Lewis’ possibilist way of dealing with such fictions was to divide them into consistent segments: “[W]here we have an inconsistent fiction, there also we have several consistent fictions that may be extracted from it” (Lewis 1983, 277). Instead of reading, e.g., *The French Lieutenant’s Woman* as one fiction with a contradictory pair of statements, we should take it as separate *fictions* in which contradictory statements are independently actualized. This is what Lewis calls *the method of union*: “ $\varphi$  is true in the original fiction iff  $\varphi$  is true in some fragment” (1983, 277). It’s true that, in one segment, Charles and Sarah are married, and it’s true in another that they are not, but there’s no segment in which the conjunction of these claims is true. Even the narrator of Fowles’ novel explicitly says that he cannot make two separate endings simultaneously true in fiction: “I cannot give both versions at once, yet whichever is the second will seem, so strong is the tyranny of the last chapter, the final, the ‘real’ version” (Fowles 2010, 347). Instead of claiming (like Alber has done) that the proceedings of the novel violate the LNC, one could merely say that we’re dealing here with a “forking-path” narrative that “develop[s] several possible storylines out of a common situation” (Ryan, Bell 2019, 23). The same seems to be the case with “The Babysitter.” The story is not a contradictory conjunction of exclusive options but a disjunction of diverse possible outcomes.<sup>6</sup>

---

<sup>6</sup> The impression that “forking-path” narratives are contradictory is based on the erroneous idea that one fiction depicts one possible world. The very language we use to talk about fictional works, as Lewis (1978, 42) suggests, leads us to this slippery terrain: we are prone to sayings like “in the world of *The Great Gatsby*” or “in the world of *The Magic Mountain*,” etc. But there’s never one single world that exclusively belongs to any particular fiction. Every fiction is a draft of countless possibilities. In one world compatible with *The Great Gatsby*, Gatsby has blue eyes, but in the other his eyes are green, etc. Fitzgerald’s novel (as far as I recall) remains silent on this issue and “is [essentially] partial in what it explicitly represents” (Berto, Jago 2019, 225). There are numerous possible worlds (with greater or lesser differences

In the case of *Traps*, the situation seems complicated by the fact that the author herself insists that the play behaves like “an impossible object” (Churchill 1985, 71). Even if by this she understands logical impossibility, the possibilist should not be troubled. For one, her remark is not part of the fiction itself (being in “Performance Notes,” it represents an extra-fictional commentary). One cannot establish fictional facts by subsequent extra-fictional calibrations. We can put this idea to a simple test: we are in the audience, watching *Traps*. In one scene, Alber and Syl are married, and in the other—they are not. There would be no single moment in which they are both married and not married. We would see merely a kaleidoscope of possibilities that follow one another. A rule of thumb here should be that if we can reject the contradictory reading of a fiction “without affecting the plot structure, then [the contradiction] does not belong to its real content” (Berčić 2021, 166).

(b) The “Rejection” Strategy

*Root vs. Branch Contradictions.* However, there are some contradictory fictional scenarios for which none of the abovementioned accommodation strategies works well. Let’s take “Sylvan’s Box” as an example. The proposition “The box was both full and empty” purports to be true in this particular fiction. It cannot be easily dismissed as a rhetorical (pseudo-)contradiction. Also, it cannot be conveniently brushed away as an authorial “slip-up” since the whole point of the story is to portray a violation of the LNC. The contradiction is intentional. Furthermore, Lewis’ method of union “does not work well in cases [like ‘Sylvan’s Box’] where the original fiction contains a single ‘root’ responsible for each branch of a contradiction”

---

between them) that correspond to Fitzgerald’s fiction. Thus, there’s no “unique world we can call the world of [*The Great Gatsby*]” since “[f]ictions are incomplete” (Berto, Jago 2019, 242), and necessarily so for it is impossible to produce a perfectly exhaustive fiction in which all particular details would be settled. See, Doležel 1998, 22. No storyworld corresponds to only one possible world since no storyworld is a singular world but a script for “a class of worlds” (Berto, Jago 2019, 242) that are compatible with it. In the case of *The French Lieutenant’s Woman*, the way out of contradiction appears to be a straightforward one: in some possible worlds, Charles and Sarah get married, while in others, they don’t. The fiction encompasses and portrays these possibilities with no contradiction incurred.

(Proudfoot 2018, 99). We cannot divide the story into consistent segments without the whole plot structure being destroyed.

One can thus distinguish between two types of contradictory fictions. Let's describe them (following Proudfoot 2018) as fictions with *branch* contradictions and fictions with *root* contradictions.

( $\alpha$ ) The term “branch” contradiction describes “a pair of statements of the form  $\varphi$  and not- $\varphi$ ” (Varzi 2004, 95) in a single fiction. These are the fictions with contradictory segments. This is the case with all “forking-path” narratives where we have “the representation of logically incompatible situations (such as the various scenarios or plotlines)” (Alber 2019, 159). Such contradictions are “divisible” by Lewis’ method of union, and, as we have shown, “there is the opportunity to interpret [them] in an uninteresting way” (Sorensen 2002, 347). These “contradictions” are “venial” (Kroon, Voltoni 2019) since we can preserve the plot structure without committing ourselves to truthful violations of the LNC in fiction. There is no hermeneutical need to interpret the story in a contradictory fashion.

( $\beta$ ) The phrase “root contradiction,” on the other hand, does not describe a pair of exclusionary propositions but “a [single] statement of the form  $\varphi$  and not- $\varphi$ ” (Varzi 2004, 95). When a genuine root contradiction appears, we are dealing with “fatally inconsistent fictions” (Hanley 2004, 113) since the contradiction is indivisible. There are no separate contradictory situations, but only one integral contradictory situation. We cannot divide it into consistent segments or read it convincingly as authorial mistake. This means that the price for eliminating such a contradiction from the story would be the “wholesale destruction of the [plot structure]” (García-Carpintero 2022, 319).

In the cases of such blatant contradictions, the possibilist cannot convincingly resort to reconciliation strategies. He would have to straightforwardly reject the possibility of such claims being true in fiction.

*The Case for Unreliability.* Let's start with something that seems like a truism: “It's true in every story that the story is told” (Hanley 2004, 118). That is, there typically is some narratorial agency that conveys fictional events. Such agency is called *the narrator*. Thus, for a fictional claim  $\varphi$ , we can introduce an operator that points out “the intensional context” of the claim: “According to S:  $\varphi$ ” (Varzi 2004, 97), where S is some narratorial



agency. “According to S:  $\varphi$ ” describes a “propositional attitude” (Fořt 2016, 34) like “Peter told me that  $\varphi$ ” or “According to Paul,  $\varphi$ .”

Let’s now state another well-known fact: “[T]he mere [...] utterance of [ $\varphi$ ] does not suffice to make it true [in fiction]” (Xhignesse 2016, 153). This phenomenon is known as *narratorial unreliability*. Narrators can sometimes be mistaken, highly biased, or deceptive. For instance, Johnny Truant, the narrator of *House of Leaves*, is a drug addict with a distorted grasp of reality. One should be wary of taking anything he says at face value. Sometimes the best explanation of a fictional fact is not necessarily the one provided by the narrator.

Thus, the obvious possibilist strategy of dealing with the root contradiction in “Sylvan’s Box” would be to reject it on the grounds of narratorial unreliability and so preserve the LNC as a basic principle. If we rephrase the problematic proposition so that its reference is included, we get the following sentence: “According to the narrator of ‘Sylvan’s Box,’ there is a box that is both full and empty.” We are not dealing anymore with a contradictory fact, but merely with the narrator’s contradictory belief. The sentence is suddenly much less philosophically interesting since people claim and “believe all sorts of things” that “needn’t be true, or even reasonable” (Berto, Jago 2019, 235). The possibilist can thus argue that “Sylvan’s Box” is a possible fiction “where we are presented with unreliable [narrator] with inconsistent beliefs” (Nolan 2021, 18).

How do we know that we are dealing with unreliable narration? Some tell-tale signs are to be assessed “case by case” (Varzi 2004, 97). Regarding “Sylvan’s Box,” there are a few clues that can help us make a solid case for narratorial unreliability.

First of all, even if we ignore the fact that the narrator believes that there is a box that is both full and empty, there are other unaccounted inconsistencies in Priest’s fiction. We are told that there is nothing special about the box itself. “It was of brown cardboard of poor quality” (Priest 1997, 575). However, towards the end of the narration, the box (not just its content) starts to behave in a contradictory fashion without anyone being particularly puzzled by it: “I carried the box outside; Nick carried the box outside. I opened the car door; Nick picked up a spade and dug a hole. I put the box in the car; Nick put the box in the hole. I closed the door on the

box and locked it; Nick covered the box with dirt and stamped it down” (Priest 1997, 579). As readers, we can be puzzled by the main event. Something strange is happening inside the box that baffles the narrator and his companion into thinking that they have discovered a logically impossible object. But with the proliferation of contradictions in the last paragraph “without [any] plausible explanation supplied,” we “have gotten off train” (Hanley 2004, 125). The story appears to be not the narrator’s mimetic account of his experiences but a joke.

Furthermore, the narrator seems to be biased. Even before discovering the box, he was already a staunch believer in true contradictions. That was the subject of debate between the late Richard Sylvan and him: “When I first met Richard, we had disagreed over whether the actual world could contain contradictions. I thought that maybe it could” (Priest 1997, 577). Whatever was happening inside the box, the narrator was already predisposed to interpret it in one way rather than another. And the narrator’s account itself leaves some room for doubt about what has happened. He uses cautious language: “What I had discovered *seemed* [my emphasis] so unlikely, impossible even—just as the box said” (Priest 1997, 576).<sup>7</sup>

Whenever a first-person narrator says things that sound improbable (or even contradictory) or problematic in any way, we tend to shift our focus from what has been said to who is saying it. Since unreliability is always a latent possibility in the first-person accounts, such narrators often lack the means to establish a definite report that cannot be doubted.

By employing the so-called reconciliation strategies, the possibilist can demonstrate that the majority of fictions that have been usually regarded as contradictory do not actually violate the LNC. It appears that if we can

---

<sup>7</sup> One could add that the narrator also employs language that signals fabulation, not the recounting of authentic facts that are taking place within the storyworld. His account is characterized by poetical literary exacerbations and is overtly adorned with figures of speech: “[I]t was the magic time of day, that time when the sun mercifully elects to hide for a few hours, and the roasted earth heaves a sigh of relief” (Priest 1997, 573). This seems to be an idiom one would use when deliberately producing a fictitious account. After all, “[made-up] stories are hardly ever told in the same way as factual ones, and fictionality can usually be detected in a blind test” (Ryan, Bell 2019, 16).

convincingly read (e.g.) *The French Lieutenant's Woman* in a consistent manner, then we have a good reason to believe that it's a consistent fiction. By preserving the consistency of plot structure, the possibilist approach has the additional "virtue of doxastic conservatism"<sup>8</sup> since it allows us to keep the LNC as the basic law of logic. When it comes to fictions with branch contradictions, the possibilist interpretation simply appears to be more convincing than the impossibilist one. Thus, the only real candidates for contradictory fictions are those with root contradictions: when the narrator asserts a single proposition that explicitly violates the LNC. In such cases, the possibilist can point out the fact that even if fiction explicitly contains the assertion of the type  $p \wedge \sim p$ , this is still "a very long way from establishing that it is true in the story that  $[p \wedge \sim p]$ " (Hanley 2004, 120). Due to the general prospect of narratorial unreliability, the most we can say about fictions like "Sylvan's Box" is that something puzzling has happened. But we are not obliged by the fictional content to take any further steps. It seems more plausible to suppose that the narrator is (for example) in a state of cognitive disarray than that the LNC has really been violated.

### 3. Impossibilist objections and ways forward

*The Convention of Reliability.* However, for impossibilists, the matter is far from settled. Their reply could take the following course. Let's suppose that the narrator of "Sylvan's Box" is indeed unreliable. After all, first-person narrators can more often than not be convincingly challenged by the unreliability charges. But what if we rewrite the story so that the narrator is no longer some (first-person) Australian philosopher but the third-person narrative agency, the one that cannot be identified with any of the characters? The impossibilist's argument runs as follows:

- (i) We can rewrite "Sylvan's Box" in such a manner that the claim "The box is both full and empty" is uttered in the third-person.
- (ii) All claims uttered in the third person are true (by convention).

---

<sup>8</sup> I thank the anonymous *Organon F* reviewer for this formulation.

- (iii) Thus, it would be true in the fiction that “the box is both full and empty.”

Concerning (ii), it seems that “[t]he basic convention of [fiction] is that narrative sentences not produced by the characters are true” (Culler 2004, 27). The third-person utterances are generally understood not as someone’s opinions but as fictional facts. Third-person agency is not a person but a “narrating function (*Erzählfunktion*),” devoid of “gender, [...] name, [...] age or I” (Behrendt, Hansen 2011, 227). It seems to be a completely objective generator of fictional facts. If the possibilist somehow argues against (ii), by introducing the concept of third-person unreliability, “something that has been regarded as unimaginable” (Behrendt, Hansen 2011, 219-20), he would be attacking a major convention of fictional discourse, and his argumentation would, as the impossibilist contends, “smack of the *ad hoc*” (Xhignesse 2016, 152).<sup>9</sup> In order to save fictional events from contradictions, the possibilist would be sacrificing our standard way of dealing with fiction.

*The Possibilist Reply: Coding Error.* What could be the possibilist answer here? One could argue that (ii) represents an oversimplification of the conventions governing narration. It’s generally true that we take pieces of information promulgated in the third person as fictional facts. However, this is not always the case. We need a more nuanced approach.

Imagine that you are engaged with a fiction narrated in the third person. The narratorial agency reveals certain facts (let’s say it’s a story about a ruthless politician who slanders his opponents and destroys their lives through fabricated scandals), and you accept them as true in that fiction. Everything is running smoothly until the third-person voice of the narrative describes the politician as “a sensitive man, overtly obsessed with ethical issues.” This seems obviously false. Maybe it’s meant ironically. But, could a non-personal narrative agency be capable of irony and sarcasm, or does this “flash [it] out as a character[-narrator]” (Behrendt, Hansen 2011, 222) after all? Suddenly, you want to rethink what’s going on in the fiction and who is narrating it, despite the apparent third-person voice. Maybe we are

---

<sup>9</sup> The possibilist’s recourse to unreliability in the case of the third-person narration would ultimately lead to some sort of narrative solipsism (there would be no fictional facts we could assert beyond the existence of the narrator).

dealing here with the *ersatz* third-person narration.<sup>10</sup> Or, it could be a case of “psycho-narration” (Behrendt, Hansen 2011, 236), the situation in which the third-person narrator merely conveys and mirrors the (potentially misguided) thoughts of the characters, not the objective facts (by omitting the attribution: “..., he thought”). We accept the factuality of the third-person narration only until we encounter something problematic and ambiguous that compels us to reassess the information that has been provided. The mere presence of the third-person voice doesn’t automatically mean that the story is narrated by some non-personal narrative agency that conveys information with utter authority.<sup>11</sup>

Truth in fiction is not so much an alethic as it is a *pragmatic* concept. We take the narrated information as true in fiction until we have some pragmatic reason to doubt it.<sup>12</sup> It is an issue of “smoothness” of narration, not of the infallible narrative convention that automatically establishes random facts *ex opere operato*. As Hanley puts it, “judgments of truth in fiction [are] probable rather than certain” (2004, 116).<sup>13</sup>

---

<sup>10</sup> This seems to be the case in Dinesen’s “The Sailor-Boy’s Tale,” a story that appears to be narrated in the third person until the last sentence where, as Behrendt and Hansen (2011, 236) note, we find clues that this could actually be the case of a covert first-person narration. Perhaps in order to feign objectivity or to distance himself from traumatic events, the first-person narrator may mimic the third-person narrative attitude.

<sup>11</sup> A similar line of reasoning can be employed in solving the so-called puzzle of imaginative resistance. See, Vujošević 2023.

<sup>12</sup> In third-person narratives, we may initially assume that the story is told by some non-personal, “objective” agency until we encounter something that prompts us to “personalize” the narrator (e.g., there is an obvious error, or an overtly personal or neurotic tone suddenly emerges in narration, etc.).

<sup>13</sup> Generally, truth in fiction seems to be an aesthetically trivial concept, for it has little bearing on our standard engagements with fiction. For example, is Myrtle’s death in *The Great Gatsby* an accident or a premeditated act on Daisy’s part? There is no way to know for sure, and this indeterminacy even contributes to the aesthetic value of the novel. We cannot furnish conclusive evidence for much of the fictional happenings. The majority of great fictional works are games of interpretation where what is (really) true (in fiction) may remain radically elusive.

So, what is not running “smoothly” in “Sylvan’s Box” that make us hesitate to accept that “the box is both full and empty” as true in fiction, despite the fact that the story is now narrated in the third person?

Every fiction “presents areas of radical indeterminacy” (Ryan, Bell 2019, 10-11). Fictions do not provide all the information about their respective storyworlds: we never get to know, e.g., what Heathcliff was doing for three years after running away from Wuthering Heights. However, such “under-descriptions” are contingent. The authors could have filled these blanks if they wanted to. The problem with “Sylvan’s Box” is that the crucial event of the story must remain underdescribed.

Let’s imagine that Priest was making an indie film called “Sylvan’s Box,” and he wanted to present the discovery of a receptacle that is both full and empty. He could try to achieve this only in two ways: by some second-hand announcement (the audience never sees the actual content of the box, but merely observes the characters exclaiming in utter surprise: “The box is both full and empty!”) or by misrepresenting the content: for example, the audience sees a figurine in the box, flickering in and out of existence, “like a malfunctioning cloaking device” (Xhignesse 2021, 3180). In the first case, we are only provided with an indirect clue (the character’s testimony) that the box is “fempty,” but this “does not make the contradiction true [in fiction]” (Xhignesse 2021, 3181). What is going on remains hopelessly fuzzy. In the second case, we can claim that what characters describe as a contradiction is no contradiction at all. Priest simply cannot forge a fiction in which it would be plainly true that the box is both full and empty.

Now, someone could say that a proposition can be true in fiction even if it cannot be visually represented (or clearly perceptually imagined). There is no direct link between truth in fiction and visual representation. Priest offers the example of “a chiliagon (a regular 1000-sided figure)” (2016, 2659). One could even produce a story in which an immortal being creates a megagonic structure (an edifice with one million angles). We cannot clearly and distinctly imagine such a thing, but nevertheless, our general intuition is that such an event could still be true in fiction. What can be replied here is that it’s true that we cannot clearly imagine a megagon, but the imaginative impossibility is not structural here. The restriction is within

our mental capacities, not in the object as such. That is, I cannot clearly imagine a figure with one million sides, but I can still imagine it in a modular way. Instead of imagining a million angles, I imagine a figure with many, many angles. Such a figure serves as an imaginative model for the megagon (since it's not structurally dissimilar from it). It merely differs in the degree of completeness. However, there is no such a model for a box that is both full and empty. If I imagine the content of the box as flashing in and out of existence (like the digits on an alarm clock), I'm imagining something structurally different from simultaneous existence and non-existence. There's no imaginative model for contradictions.

However, the impossibilist could further argue that some fictional facts that cannot be imagined perceptually (even in a modular fashion) could still be true in fiction. One can distinguish between “two [kinds of] coding” (Berto, Jago 2019, 34) in fiction: perceptual and propositional. Perceptual representations are “characterized by reference to sensory perception” (Berto, Jago 2019, 34). They are “pictorial” (Berto, Jago 2019, 34), and they provide instructions on how to imagine something. An example of perceptual representation could be the crime scene in the Luriston Gardens in Doyle's *A Study in Scarlet*. We have a detailed description of the room with the corpse lying on the floor. The event is perceptually coded. What is to be imagined is explicitly stipulated. One could make a film or a play out of these sensory pieces of information. But not all fictional facts are perceptual. There are propositional representations that deal with “abstract scenarios.” They are “amodal” since “they are disconnected from sensory modalities” (Berto, Jago 2019, 35). They lack any relevant perceptual stipulation. It seems that contradictory claims in fiction could belong only to the domain of propositional representation. That is, if we can imagine contradictions at all, we do so only on the propositional level since they come without any stipulation on how to perceptually imagine them. They are exclusively propositionally coded.<sup>14</sup> But this is exactly the problem with “Sylvan's Box.”

---

<sup>14</sup> Not everyone agrees with this. Priest seems to argue that we imagine and even perceive contradictions on a routine basis. For example, consider this simple scenario: While walking out of my apartment, for a split second I'm perfectly “symmetrically poised” (Priest 2004, 28) in the doorway so that my left foot is still inside the apartment while my right foot is in the hallway. Let's “freeze” this particular moment and

What's taking place in this fiction is a specific type of *coding error*. That contradictions could be imagined propositionally is beside the point in the case of "Sylvan's Box," since, in this fiction, we are invited to imagine the focal contradiction perceptually as a realized state of affairs in the story-world (because the contradiction is something that the characters feel, see, and touch). However, no instruction is offered (nor it could be adequately offered) on how to do so. The very invitation to imagine such a proposition is void, yet persistent (since we are dealing here with a sensory content). This creates a zone "of radical indeterminacy" (Ryan, Bell 2019, 10-11) concerning what is really going on in the fiction. All conceivable stipulations on how to imagine the contradictory proposition in "Sylvan's Box" would not be about the contradictory proposition but about some other (non-contradictory) state of affairs. One can produce a fiction about some

---

ask a simple question: "Am I in or not in the [apartment]?" (Priest 2004, 28). It seems that "I am both in and not in" (Priest 2004, 28). Seemingly, we are dealing with a perfectly imaginable yet contradictory situation. However, the possibilist need not be particularly troubled by the examples like this one. These scenarios are fundamentally different from the one outlined in "Sylvan's Box." The ambition of "Sylvan's Box" is to portray an "ineliminable contradictory [fictional fact]" (Mares 2004, 271). No reconciliation strategy works here. The contradiction cannot be eliminated from the story without rejecting the narrator's authority. The narrator offers a metaphysical claim: he came to believe that there are certain "aspects of the world [he inhabits] for which any accurate description will contain a true contradiction" (Mares 2004, 270). The box is empty and not empty "at the same time, and in the same respect" (Aristotle, as translated in Horn 2018). This is not the case with the doorway scenario since one can "re-describe [it] [...] consistently without sacrificing accuracy" (Mares 2004, 270), by (e.g.) offering some correct protocol-like report of the situation that avoids the contradictory formulation (my left foot is in the room and my right foot is in the hallway, etc.). In any case, it is not exactly true that while I am in the room, I am also not in the room "in the same respect" (which seems a necessary prerequisite for genuine contradictions). Ultimately, even if one accepts the (moderate) dialethic stance that some "semantic contradictions [...] are [...] 'true,' [...] or not necessarily false" (Grim 2004, 55), such as the Liar sentence or the examples offered by Mares (whose article was suggested to me by anonymous referee), and that "inconsistencies [...] may arise because of the relationship between language and the world" (Mares 2004, 265), this still does not entail "any painful metaphysical commitment" to such "untoward entities" (Mares 2004, 274) as empty boxes.



contradictory situation, but he will repeatedly fail to make it the undoubted fact in the story. There is a fatal incompleteness in such fictions, and the question of fictional truth cannot be properly settled.

#### 4. “Radical indeterminacy” of impossible fictions

Should we reject the LNC in favor of the PPL? We’ve seen that the impossibilist argues so while drawing our attention to a plethora of works of fiction that seemingly contain contradictions as their essential elements. Readers engage with contradictory fictions all the time and seem to understand them.<sup>15</sup> According to the impossibilist, we should take a dim view of the possibilist’s claim that the right way to read these stories has to be the one not intended by the authors and not pursued by the majority of ordinary readers. As Nolan points out, “[t]he main drawback of [the possibilist strategy] is that [it] often seems undermotivated by the texts and audience reception” (2021, 10).

However, there is one thing that may prevent someone from automatically accepting the impossibilist stance that there may be genuine violations of the LNC in fiction. According to the influential Lewisian understanding of “truth in fiction,” a proposition is true in fiction if it “[obtains] in a possible world [or a set of worlds]” (Currie 1990, 54). There are possible worlds where, for example, a reanimated outlandish corpse reads *The Sorrows of Young Werther* and recites, in eloquent remorse, passages from *Paradise Lost* over Victor Frankenstein’s dead body. However, it appears that if we concede that even logically contradictory propositions can be true in fiction, this would force us to “admit appropriately selected impossible worlds to the set of worlds that realize what is told in [...] a story” (Kroon, Voltolini 2019) and such a commitment to “impossible possible worlds” (Lewis 1983, 275) seems to be “for many a difficult pill to swallow” (Kroon, Voltolini 2019). One tempting alternative, for impossibilists, would be to simply abandon Lewis’ model of “truth in fiction” and consider some rival

---

<sup>15</sup> Priest (2004, 35) even suggests that contradictory fictions (like “Sylvan’s Box”) would ring more consistent than some logically possible fiction in which the protagonist (for instance) randomly turns into a fried egg.

theories that seemingly retain the idea that a proposition can be “true in fiction” (which appears to be in accordance with how we generally talk about fiction) without subscribing to “the machinery of possible worlds” (Currie 1990, 147). Perhaps, there is a way to think about the contradictory claim in “Sylvan’s Box” as being fictionally true without having to grapple with what it would mean for some possible worlds to contain impossible objects like fempty boxes.

*The Waltonian Model.* Kendall Walton has proposed a pragmatic account of truth in fiction that is not dependent on possible world semantics. For Walton, a work of fiction “is a prop in a game of make-believe, where the function of the prop is to prescribe imaginings” (Kroon, Voltolini 2019). A proposition is true in fiction “if there is a prescription to the effect that it is to be imagined” (Walton 1990, 61).<sup>16</sup> For instance, we can say that, in Charlotte Brontë’s novel, it is fictionally true that Jane Eyre is a strong-willed 19th-century English governess who knows French and has green eyes (since there’s a prescription to imagine these things). But it is not fictionally true that she is a coarse 18th-century coachman with hazel eyes. To imagine her as such would be an “unauthorized [move]” (Walton 1990, 60) in the game of make-believe.

While rejecting possible world semantics, Walton still employs the concept of fictional worlds, but in a pragmatic fashion. “Fictional worlds” are “associated with [...] cluster[s] of propositions” (Walton 1990, 64) that are true in a certain fiction (which simply means that such propositions carry the invitation to be imagined). Fictional worlds are not full-blown worlds in which all sorts of bizarre phenomena are taking place. To say that Graham<sup>17</sup> and Nick discover a fempty box (in the fictional world of Priest’s story) does not commit us to “impossible possible worlds” where such an event would take place. Unlike possible worlds, fictional worlds “are sometimes impossible” (Walton 1990, 64) and they are not really worlds.

---

<sup>16</sup> Walton is cautious not to equate imagining with mental imagery since “imagining can occur without [mental images]” (1990, 13).

<sup>17</sup> I refer here to the fictional character-narrator of Priest’s story as Graham to distinguish him from the actual author, Graham Priest (who, presumably, doesn’t actually believe that he discovered a box that is both full and empty, in R. Sylvan’s house).

Considering the problem “impossible fictions” could pose for his account of truth in fiction, Walton suggests two solutions. First, he allows for the possibility that “contradictions can be imagined in [some] relevant sense” (Walton 1990, 64). He seems to suggest that, when we are dealing with a contradictory conjunction in fiction, we can understand it as involving “separate prescriptions to imagine  $p$  and to imagine not- $p$ ” (Walton 1990, 64) in the game of make-believe. However, this won’t do since such an imagining would be a direct violation of what the contradictory fiction prescribes us to imagine. For example, “Sylvan’s Box” does not invite us to imagine that Graham and Nick first discover a box with a figurine in it (in  $t_1$ ), and then (in  $t_2$ ), they find out that the box is empty. In doing so, we would be imagining a completely different scenario (that doesn’t violate the LNC), and such a move seems “unauthorized” by the fiction in question.

Walton then briefly considers the possibility that contradictory propositions cannot be imagined. However, he adds that this would not affect his “understanding of fictionality” (Walton 1990, 64). He claims that there “can be prescriptions to imagine a contradiction even if doing so is not possible” (1990, 64). Kroon and Voltolini understand this passage as suggesting that contradictions can be imagined propositionally. Therefore, “what is at stake here is propositional imagining,” not imagination that “relies purely on mental imagery” (Kroon, Voltolini 2019). As readers, we are invited to propositionally imagine that Graham has discovered an object that violates the LNC (without perceptually imagining anything specific). This seems like a very convenient strategy: I imagine a non-contradictory (perceptual) situation described in the story (i.e., after opening a certain box, two people are claiming that it is both empty and non-empty), and then I add: “I perceptually imagine that what they are claiming is true in fiction.” However, here one encounters a similar problem as with the Lewisian model.

Let’s suppose that we are dealing with two fictional variants of “Sylvan’s Box” called  $SB_1$  and  $SB_2$ . Let’s further say that, in  $SB_1$ , it is fictionally true that a contradictory object exists and is discovered by Graham and Nick. However, in  $SB_2$ , they only *believe* that they’ve discovered such an object. These scenarios have to be radically different since  $SB_1$  is logically inconsistent and  $SB_2$  is perfectly consistent (after all, people do believe all sorts

of things). However, the basic plots of  $SB_1$  and  $SB_2$  seem indistinguishable. That is, what is stipulated to be imagined perceptually is identical in both stories. In both  $SB_1$  and  $SB_2$ , it is fictionally true (in Waltonian terms: prescribed to be imagined) that two friends discover a box, and that, upon opening it, they come to believe that they've found an object that violates the LNC. Their belief is not merely propositional. They are not simply considering some abstract semantic scenario, but are dealing with something that they can touch, see, and even move across the room. However, in both stories, they are radically unable to describe the object in any relevant detail. The narrator of "Sylvan's Box" admits this: "[I]t is impossible to explain what the perception of a contradiction, naked and brazen, is like" (Priest 1997, 576). If we try to imagine such an object perceptually, we'll end up imagining something non-contradictory: a flickering figurine, a translucent holographic image, etc. There can be no close-up of the object in either of the stories.  $SB_1$  and  $SB_2$  are identical in this regard.

However, the same perceptual coding leads to different prescriptions for propositional imagining in  $SB_1$  and  $SB_2$ . There is an additional prescription in  $SB_1$  to (propositionally) imagine the characters' belief as being true in the story, while the most generous thing we can say for  $SB_2$  is that it invites us only to imagine that the fiction remains undecidable in its crucial aspect. But what would warrant such a prescription in  $SB_1$ ? The prescription to (propositionally) imagine a contradictory situation must be produced exclusively by some narratorial assertion since nothing else would do (there can be no further description of the contradictory situation in any relevant detail). However, as we have seen, narratorial assertions alone (even when they are uttered in the third-person) are not strong enough to produce an unavoidable imaginative prescription in the game of make-believe, so that if we are not acting by such a prescription, we are "misusing the work" (Walton 1990, 60).

The truth of contradictory fictions always depends solely upon the narrator's claims that cannot be backed up by any nuanced further elaboration. There is nothing in  $SB_1$  that prescribes the acceptance of the impossible situation as true in fiction (in both Lewisian and Waltonian framework), except the narrator's assertion. But we've shown that one need not rely blindly on such authority. Valid imaginative engagement with fictions may

ignore such “blank” narratorial prescriptions. Such an assertion is the only difference between  $SB_1$  and  $SB_2$ . It is an additional, *phantom* quality of  $SB_1$  that adds nothing relevant to the fictional content. It vacuously repeats the narrator’s claim from  $SB_2$  (but, perhaps, in the third-person voice).

For example, imagine Charlotte Brontë writing *Jane Eyre* in the third person with a goal to establish the fictional fact that her character is a paragon of virtue and late-Georgian morality. To use the Waltonian idiom: she wants to create a “prescription to imagine” (Walton 1990, 139) such a thing (as fictionally true). She cannot adequately achieve this by merely stating this fact (even in the third-person voice) since, later on, she may exaggerate in descriptions of her heroine’s upright behavior so that the final impression is that Jane Eyre is not so much a nice, virtuous person, but an obnoxious and tedious character. Or she may subsequently hint, in portraying Jane, that she could be someone who merely uses the mask of virtue to achieve her own selfish goals, etc. What makes Jane Eyre a virtuous character is not a mere stipulation on the narrator’s part (even in the cases of third-person narration), but the general outlay of the story that requires some narrative skill to produce. But in the case of “Sylvan’s Box,” no relevant further elaboration is possible since such fictions are created around essentially indeterminate situations (something strange has happened, but we cannot be sure what). There is simply no narrative way to expand or resolve the case in the manner that Priest would want to. Rather than modifying theories of truth in fiction so that they can accommodate real-world violations of the LNC, it may be more propitious to simply abandon the venerable old notion of *divin’ artista*, the omnipotent Author who can make anything whatsoever true in his story “simply by fiat”<sup>18</sup> (the idea that is currently known as the PPL).

## 5. Conclusion

We can say that  $SB_2$  stands for Priest’s original fiction. The character-narrator obviously believes that he has discovered some contradictory object. This seems to be the most one can say about “Sylvan’s Box.” The

---

<sup>18</sup> See (Liao 2016, 475).

author cannot upgrade SB<sub>2</sub> fiction to SB<sub>1</sub> by providing some additional pieces of information about the nature of the characters' impossible discovery. Priest (1997, 576) accepts this. One simply cannot expand the story in any relevant detail. The only promising strategy is to distance the assertion from the character-narrator and to convey it in a different narrative voice (as an objective fact in the story). But this maneuver rests on a misconception about what third-person narration really means. It's not a magical device that establishes fictional facts (or generates authentic imaginative prescriptions *ipso facto*).

The ongoing debate about logically impossible fictions is (to some extent) due to the assumption that it is easy to make something true in a story by merely stating it (through some narratorial agency). However, this is not always the case. "Sylvan's Box" is an example of "radically indeterminate" fiction. These are fictions that must remain fatally underdescribed. That is, no narrative agency can establish the contradictory fictional fact since such an event simply "cannot be described in adequate [and relevant] detail" (Ashline 1995, 222). "Sylvan's Box" thus remains radically inconclusive about what the characters have discovered inside the box. The story is centered not around a physically realized contradiction but rather a "blind spot" or an enigma (at best) of what has happened.

Fictions like "Sylvan's Box" are "interesting thought-experiment[s]" (Xhignesse 2021, 3183), but their existence is simply insufficient to prove that contradictions can be true in fiction. By constructing such fictions, one cannot prove that fictional possibility is broader than logical possibility (as it is classically understood). Such endeavors are based on an oversimplified view of truth in fiction and narrative conventions. As Xhignesse puts it, "Priest wants his readers to reflect on the possibility that the logic of fiction is paraconsistent. To get us to do so, he must first tell us to do so, but he needn't [and, in fact, he doesn't] succeed in actually making it so in the story we read" (2021, 3183).

## References

- Alber, Jan. 2009. "Impossible Storyworlds—and What to Do with Them." *Story-Worlds: A Journal of Narrative Studies*, 1(1): 79-96.  
<https://doi.org/10.1353/stw.0.0008>

- Alber, Jan. 2016. *Unnatural Narrative: Impossible Worlds in Fiction and Drama*. Lincoln: University of Nebraska Press. <https://doi.org/10.2307/j.ctt1d4v147>
- Alber, Jan. 2019. "Logical Contradictions, Possible Worlds Theory, and the Embodied Mind." In *Possible Worlds Theory and Contemporary Narratology*, edited by Alice Bell and Marie-Laure Ryan, 157-176. Lincoln: University of Nebraska Press. <https://doi.org/10.2307/j.ctv8xng0c.11>
- Ashline, William L. 1995. "The Problem of Impossible Fictions." *Style*, 29(2): 215-34.
- Badura, Christopher, and Francesco Berto. 2019. "Truth in Fiction, Impossible Worlds, and Belief Revision." *Australasian Journal of Philosophy*, 97(1):178-193. <https://doi.org/10.1080/00048402.2018.1435698>
- Behrendt, Poul and Per Krogh Hansen. 2011. "The Fifth Mode of Representation: Ambiguous Voices in Unreliable Third-Person Narration." In *Strange Voices in Narrative Fiction*, edited by Per Krogh Hansen, Stefan Iversen, Henrik Skov Nielsen, and Rolf Reitan, 219-252 Berlin and Boston: De Gruyter. <https://doi.org/10.1515/9783110268645.219>
- Berčić, Boran. 2021. "Art and the Impossible." *Croatian Journal of Philosophy*, 21(61): 155-177. <https://doi.org/10.52685/cjp.21.1.9>
- Berto, Francesco and Mark Jago. 2019. *Impossible Worlds*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198812791.001.0001>
- Calvino, Italo. 2012. *The Nonexistent Knight*, translated by Archibald Colquhoun. Boston and New York: Houghton Mifflin Harcourt.
- Churchill, Caryl. 1985. *Plays: One*. London and New York: Routledge.
- Culler, Jonathan D. 2004. "Omniscience." *Narrative*, 12(1): 22-34. <https://doi.org/10.1353/nar.2003.0020>
- Currie, Gregory. 1990. *The Nature of Fiction*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511897498>
- Danielewski, Mark Z. 2000. *House of Leaves*. New York: Pantheon Books.
- Doležel, Lubomir. 1998. *Heterocosmica: Fiction and Possible Worlds*. Baltimore: The Johns Hopkins University Press. <http://dx.doi.org/10.56021/9780801857492>
- Forť, Bohumil. 2016. *An Introduction to Fictional Worlds Theory*. Frankfurt am Main: Peter Lang. <https://doi.org/10.3726/978-3-653-06323-3>
- Fowles, John. 2010. *The French Lieutenant's Woman*. London: Vintage.
- García-Carpintero, Manuel. 2022. "Truth in Fiction Reprised." *The British Journal of Aesthetics*, 62(2): 307-324. <https://doi.org/10.1093/aesthj/ayab066>
- Grim, Patrick. 2004. "What is a Contradiction?" In *The Law of Non-Contradiction: New Philosophical Essays*, edited by Graham Priest, JC Beall, and Bradley Armour-Garb, 49-72. Oxford: Clarendon Press. <https://doi.org/10.1093/acprof:oso/9780199265176.003.0004>

- Hanley, Richard. 2004. "As Good As It Gets: Lewis on Truth in Fiction." *Australasian Journal of Philosophy*, 82(1): 112-128. <https://doi.org/10.1080/713659790>
- Horn, Laurence R. 2018. "Contradiction." *The Stanford Encyclopedia of Philosophy*. (Winter 2018 Edition), ed. by Edward Zalta. <https://plato.stanford.edu/archives/win2018/entries/contradiction>
- Hudson Hick, Darren. 2017. *Introducing Aesthetics and the Philosophy of Art*. London: Bloomsbury Publishing. <https://doi.org/10.5040/9781350006935>
- Kroon, Fred and Alberto Voltolini. 2019. "Fiction." *The Stanford Encyclopedia of Philosophy*. (Winter 2019), edited by Edward Zalta. <https://plato.stanford.edu/archives/win2019/entries/fiction>
- Lewis, David. 1978. "Truth in Fiction." *American Philosophical Quarterly*, 15(1): 37-46.
- Lewis, David. 1983. "B. Postscripts to 'Truth in Fiction.'" In *Philosophical Papers Volume I*. New York: Oxford University Press. <https://doi.org/10.1093/0195032047.003.0015>
- Liao, Shen-yi. 2016. "Imaginative Resistance, Narrative Engagement, Genre." *Res Philosophica*, 93(2): 461-482. <https://doi.org/10.11612/resphil.2016.2.93.3>
- Mares, Edwin D. 2004. "Semantic Dialetheism." In *The Law of Non-Contradiction: New Philosophical Essays*, edited by Graham Priest, JC Beall, and Bradley Armour-Garb, 264-275. Oxford: Clarendon Press. <https://doi.org/10.1093/acprof:oso/9780199265176.003.0017>
- Matravers, Derek. 2014. *Fiction and Narrative*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199647019.001.0001>
- Nolan, Daniel. 2015. "Personification and Impossible Fictions." *The British Journal of Aesthetics*, 55(1): 57-69. <https://doi.org/10.1093/aesthj/ayt053>
- Nolan, Daniel. 2021. "Impossible fictions part I: Lessons for fiction." *Philosophy Compass*, 16(2), e12723, <https://doi.org/10.1111/phc3.12723>
- Priest, Graham. 1997. "Sylvan's Box: A Short Story and Ten Morals." *Notre Dame Journal of Formal Logic*, 38(4): 573-582. <https://doi.org/10.1305/ndjfl/1039540770>
- Priest, Graham. 2004. "What's So Bad About Contradictions?" In *The Law of Non-Contradiction: New Philosophical Essays*, edited by Graham Priest, J.C. Beall, and Bradley Armour-Garb, 23-38. Oxford: Clarendon Press. <https://doi.org/10.1093/acprof:oso/9780199265176.003.0002>
- Priest, Graham. 2016. "Thinking the impossible." *Philosophical Studies*. 173(10): 2649-2662. <https://doi.org/10.1007/s11098-016-0668-5>
- Proudfoot, Diane. 2018. "Sylvan's Bottle and other problems." *Australasian Journal of Logic*, 15(2): 95-123. <https://doi.org/10.26686/ajl.v15i2.4858>
- Routley, Richard. 1979. "The Semantical Structure of Fictional Discourse." *Poetics*, 8(1-2): 3-30. [https://doi.org/10.1016/0304-422X\(79\)90013-5](https://doi.org/10.1016/0304-422X(79)90013-5)



- Ryan, Marie-Laure and Alice Bell. 2019. "Introduction: Possible Worlds Theory Revisited." In *Possible Worlds Theory and Contemporary Narratology*, edited by Alice Bell and Marie-Laure Ryan, 62-87. Lincoln: University of Nebraska Press. <https://doi.org/10.2307/j.ctv8xng0c.5>
- Ryan, Marie-Laure. 2019. "From Possible Worlds to Storyworlds: On the Worldness of Narrative Representation." In *Possible Worlds Theory and Contemporary Narratology*, edited by Alice Bell and Marie-Laure Ryan, 62-87. Lincoln: University of Nebraska Press. <https://doi.org/10.2307/j.ctv8xng0c.7>
- Sorensen, Roy. 2002. "The Art of the Impossible." In *Conceivability and Possibility*, edited by Tamar Szabo Gendler and John Hawthorne, 337-368. Oxford: Clarendon Press. <https://doi.org/10.1093/oso/9780198250890.003.0010>
- Varzi, Achille C. 2004. "Conjunction and Contradiction." In *The Law of Non-Contradiction: New Philosophical Essays*, edited by Graham Priest, JC Beall, and Bradley Armour-Garb, 93-110. Oxford: Clarendon Press. <https://doi.org/10.1093/acprof:oso/9780199265176.003.0007>
- Vujošević, Vladimir. 2023. "Narrative Omniscience and the Problem of the Fictional Truthfulness of Deviant Evaluations." ["Pripovjedačko sveznanje i problem fikcionalne istinitosti devijantnih evaluacija."] *Filozofska istraživanja*. 43(1): 181-198. <https://doi.org/10.21464/fi43110>
- Walton, Kendall L. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. London and Cambridge, Massachusetts: Harvard University Press.
- Xhignesse, Michael-Antoine. 2016. "The Trouble with Poetic Licence." *The British Journal of Aesthetics*, 56(2): 149-161. <https://doi.org/10.1093/aesthj/ayv053>
- Xhignesse, Michael-Antoine. 2021. "Imagining fictional contradictions." *Synthese*, 199: 3169-3188. <https://doi.org/10.1007/s11229-020-02929-0>

## Lampert on the Fixity of the Past

Brian Garrett\* – Jeremiah Joven Joaquin\*\*

Received: 24 February 2024 / Accepted: 27 February 2024

*Abstract:* In ‘A Puzzle about the Fixity of the Past’, Fabio Lampert argues that the principle of the fixity of the past is at odds with standard views about knowledge and the semantics for ‘actually’. In this paper, we show that Lampert’s argument fails because of its use of the material conditional.

*Keywords:* Material conditional; fixity of the past; free will; fatalism.

### I

Fabio Lampert (2022) endorses an argument intended to undermine the principle of the fixity of the past. This principle is formulated thus:

(FP) For any action  $\phi$ , agent  $S$ , times  $t$  and  $t'$  (where  $t \leq t'$ ) and possible world  $w$ ,  $S$  is able at  $t$  to  $\phi$  at  $t'$  in  $w$  only if there is

---

\* Australian National University

 <https://orcid.org/0000-0002-0604-0260>

 School of Philosophy, Australian National University, 146 Ellery Cres, Acton, ACT 2601, Australia

 [brian.garrett@anu.edu.au](mailto:brian.garrett@anu.edu.au)

\*\* De La Salle University

 <https://orcid.org/0000-0002-8621-6413>

 Department of Philosophy, De La Salle University, 2401 Taft Avenue, Malate, Manila 0922, Philippines

 [jeremiah.joaquin@dlsu.edu.ph](mailto:jeremiah.joaquin@dlsu.edu.ph)

a possible world  $w'$  with the same past as that of  $w$  up to  $t$  in which  $S \phi$ -s at  $t'$ . (Lampert 2022, 426)

This principle plays an important role in standard fatalist and incompatibilist arguments for the conclusion that we are not free, where freedom is understood as the ability to do or to have done otherwise.

## II

What then is Lampert's argument? He writes:

Let ' $\Box$ ', ' $A$ ' and ' $K$ ' stand for 'necessarily', 'actually' and 'it was, is or will be known that', respectively, while ' $\supset$ ' is the material conditional.

Then:

- (1)  $q \supset ((q \supset p) \supset p)$
- (2)  $Ap \supset \Box Ap$
- (3)  $\Box(Kp \supset p)$ ; so
- (4)  $Ap \supset \Box(K(Ap \supset p) \supset p)$ . (Lampert 2022, 426)

## III

That is his main argument, but it is then applied to a particular example in which the following three propositions are assumed to be true:

- (5) S actually  $\phi$ -s at  $t''$ .
- (6) S is able at  $t'$  to not  $\phi$  at  $t''$ .
- (7) It was known that (S actually  $\phi$ -s at  $t''$  only if S  $\phi$ -s at  $t''$ ) at  $t$ .

From (4) and (5) we can derive:

- (8) Necessarily, if it was, is or will be known that (S actually  $\phi$ -s at  $t''$  only if S  $\phi$ -s at  $t''$ ), then S  $\phi$ -s at  $t''$ .

Then, from (FP) and (6), (9) follows:

- (9) There is a possible world  $w'$  with the same past as that of the actual world up to  $t'$  in which  $S$  does not  $\phi$  at  $t''$ .

(9) implies:

- (10) It is not the case that  $S$   $\phi$ -s at  $t''$  in  $w'$ .

From (7) and (9), and since  $w'$  and the actual world share the same past up to  $t'$ , it follows that:

- (11) It was known that ( $S$  actually  $\phi$ -s at  $t''$  only if  $S$   $\phi$ -s at  $t''$ ) at  $t$  in  $w'$ .

(11) implies:

- (12) It was, is or will be known that ( $S$  actually  $\phi$ -s at  $t''$  only if  $S$   $\phi$ -s at  $t''$ ) in  $w'$ .

Finally:

- (13)  $S$   $\phi$ -s at  $t''$  in  $w'$ .

follows from (8) and (12), contradicting (10).

Lampert claims that “We have thus arrived at a contradiction given (FP), (4) and the trio (5), (6) and (7)” (2022, 428).

## IV

Obviously, since (5), (6) and (7) frame our example, (FP) is threatened if (8) was soundly derived. But plainly it was not since it is clearly false. As Lampert concedes, (contingent yet *a priori*) conditionals of the form  $Ap \supset p$  are *true whatever p's truth value* (Lampert 2022, 427; 432; 433n6). Even if  $p$  is false, it is true that  $Ap \supset p$ . But how then can (8) be true? How can my knowledge that  $Ap \supset p$  imply  $p$  if that knowledge is compatible with  $p$ 's falsity?

Something has gone wrong. But what? Note that it is always potentially worrying when a philosophical argument, as opposed to an argument in a logic textbook, makes use of the material conditional. Any such argument

will be of limited or zero interest if it exploits logical features of the material conditional, which intuitively diverge from those of the (natural language) indicative conditional. And so it is here.

(1) is a classical tautology but, intuitively, its natural language counterparts are not. Consider this example:

- (1a) IF Trump wins, THEN if Trump wins then Trump loses, then Trump loses.

It is reasonable to regard (1a), not just as something that no one in their right mind would assert, but as plain false (or anyway untrue).

## V

We are thus in a rather curious situation. There is nothing wrong with the (1) – (4) argument, given the rules governing ‘ $\supset$ ’. But if the indicative is not material, then Lampert’s argument is merely a technical exercise that fails to validate (8), which is formulated using the English indicative. The only conclusion to be drawn, if we are happy with the English variants of (2) and (3) – ‘if Ap then necessarily Ap’ and ‘Necessarily if Kp then p’ – is that Lampert’s argument is further proof that the indicative conditional is not material.

## Reference

Lampert, Fabio. 2022. “A Puzzle about the Fixity of the Past.” *Analysis*, 82(3): 426-434. <https://doi.org/10.1093/analys/anab092>