A Worlds Account of Thought Experiments

Santiago Rodriguez
University of Central Florida
27 June 2021

Abstract

Besides logical arguments, thought experiments are used extensively in philosophy and science to acquire knowledge about the natural world. Many times, these experiments led to new discoveries. However, there are also various instances of similar thought experiments producing wildly contradictory results. This raises the epistemological challenge of the reliability of thought experiments. Multiple theories have been proposed that aim to justify their usage. However, they fail to properly treat experiments about more intricate, nuanced subjects infamous for their vagueness and ambiguity. This paper attempts to address this difficulty by formulating thought experiments under possible worlds semantics in order to take advantage of the modal machinery that has been developed over the past decade.

Introduction

Throughout history, thought experiments held its place as an invaluable asset for seeking truth, on par with logical arguments and empirical evidence. However, as noted by Adriano Angelucci and Margherita Arcangeli in their paper "Introduction: New Perspectives on Philosophical Thought Experiments", influential criticisms since the late 20th Century casted doubt on the reliability of thought experiments as objective arbiters of truth. Most concerning was that this threatened to undermine a majority of philosophy that relied on thought experiments to refine intuition and beliefs. Thus, we are faced with the following epistemological challenge: can we gain knowledge through thought experimentation? And if so, what kind of knowledge would this be?

The main issue with the practice is "its alleged reliance on intuitions as a trustworthy source of evidence" since intuitions have repeatedly been shown to be biased by truth-irrelevant factors like culture, geography, and socioeconomic background [1]. As such, various ingenious solutions have been proposed that reformulate thought experiments in a more rigorous form. Prominent theories include John D. Norton's argument view (that all thought experiments are formal arguments in disguise) and Nancy Nersessian's mental-model account (that all thought experiments are simulations of idealized scenarios). The goal of this reconstitution program is to justify inferences about the natural world from thought experiments without losing their convincingness.

While the argument view is probably the best at justifying thought experiments, it's the least like how we use them. Similarly, while the mental-model account faithfully mimics how we use thought experiments, it's not sufficiently justified with how conclusions follow. Given that neither resolves the issue completely, an alternative approach or synthesis of these theses is needed. Fortunately, we don't have to look far. Given that thought experiments are essentially imagined scenarios, we can interpret them as experiments performed in possible worlds. As such, we have all the tools of modality at our disposal to discover facts about these worlds that may apply to reality.

Literature Review

A variety of accounts have been proposed that address the epistemological challenge faced by thought experiments. For the sake of clarity, we will look at the four most common as identified in the Stanford Encyclopedia of Philosophy's article titled "Thought Experiments", including: (A) the "skeptical objection", (B) the "intuition-based account", (C) the "argument view", and (D) the "mental-model account".

(A) The Skeptical Objection

> The view that thought experiments are too unreliable for use in serious inquiry. This is based on a family of challenges raised by the skeptic over the reliability of intuition, which they claim thought experiments are inextricably tied to. Two main challenges can be found in the literature.
>
> Daniel Dennet argues that thought experiments act as persuasive tools designed to lead the audience to a particular intuition, which he calls "intuition pumps". While Dennet accepts that intuition pumps are useful at illustrating concepts, he fears that a majority are malformed since they rest on faulty common sense. Take Lucretius' flying spear experiment:
>
> > If there is a purported boundary to the universe, we can toss a spear at it. If the spear flies through, it isn't a boundary after all; if the spear bounces

> back, then there must be something beyond the supposed edge of space, a cosmic wall [...] that is itself in space. Either way, there is no edge of the universe; thus, space is infinite. [2]

Although this thought experiment has an intuitive conclusion, it is nevertheless flawed. Imagine walking in any direction on the surface of the globe, a two-dimensional space. We would never find an edge of the world, yet the space is finite. The universe might be a three-dimensional version of this topology.

Ultimately, Dennet and others charge thought experiments with being too conservative, that they conform to current intuitions and lead to little progress, which is why they can be so misguided. This contrasts physical experiments in science which many times produce highly counterintuitive results. However, this may be an unfair charge. Take Galileo's seafaring experiment:

> If we are inside a ship and perform a number of experiments, such as walking about, tossing a ball, watching birds fly about, we could not tell whether we are at rest in port or sailing over a smooth sea. The upshot is that nature behaves the same either way; the laws of nature are the same in any inertial frame. [2]

This thought experiment leads to a profound result, the principle of relativity, that remains even today in Einstein's theory of relativity. While we cannot guarantee that the principle of relativity will always hold, we can clearly reach counter-intuitive results from naïve intuitions through thought experiments. However, there is a stronger skeptical challenge that rears its head frequently in ethics and philosophy of mind.

Kathleen Wilkes argues that any successful thought experiment must be fully expressed and abide by all known laws of nature, otherwise we risk confusing ourselves. Many philosophical thought experiments on morality and personal identity fail this criterion because they either lack sufficient detail or are too detached from reality. Wilkes goes one step further and suggests that the lack of description is unavoidable. Take Derek Parfit's people splitting like amoebas which aims at provoking questions about personal identity. Wilkes contends that the scenario is not only inadequately described but also theoretically absurd and should be dismissed entirely. Another thought experiments that has received major criticisms is Robert Nozick's utility monster.

Imagine a utilitarian world inhabited by a being capable of amassing enourmous pleasure from the sacrifices of others, greater than what others lose. Since the world is utilitarian, the only actions that are moral are those that maximize the

monster's pleasure even at the expense of everybody else, however philosophers like Derek Parfit have complained that the thought experiment is absurd (ironic coming from Parfit) because we do not know how to make sense of a being that gains more pleasure than the pain of everybody else.

While many would agree that these type of thought experiments are problematic, it is not clear whether they should be so easily dismissed. They may say more about the complexity of the subject than the usefulness of the thought experiment and can help guide further inquiry. Furthermore, the claim that thought experiments must satisfy our current scientific knowledge appears unbased. Reasoning about scenarios that violate the laws of nature can teach us a great deal about the world and our theories.

(B) The Intuition-Based Account

The view that thought experiments access a priori knowledge of nature from intuition. This is based on the idea that we have an immense amount of knowledge (some potentially false) stored in our intuition that we may not be aware of. For example, we can easily distinguish between cats and dogs without being consciously aware of what fundamentally defines a cat and a dog. Thought experiments, then, make apparent this hidden knowledge we hold. Take Galileo's falling body experiment:

> [T]he then reigning Aristotelian account [...] holds that heavy bodies fall faster than light ones (H > L). But consider [...] a heavy cannon ball (H) and light musket ball (L) [that] are attached together to form a compound object (H+L); the latter must fall faster than the cannon ball alone. Yet the compound object must also fall slower, since the light part will act as a drag on the heavy part. Now we have a contradiction: H+L > H and H > H+L. That's the end of Aristotle's theory. But there is a bonus, since the right account is now obvious: they all fall at the same speed (H = L = H+L). [2]

In this thought experiment, we never actually dropped a canon ball and musket ball, yet we seem to have learned something new about the world. No new empirical data was introduced, nor did we derive the conclusion from old empirical data, so how can this be? It seems as if we pulled the conclusion out of a magic hat. The intuition-based account claims that we always had access to this knowledge (a priori), just that we needed a bit of intuition to get us there. There are two main flavors regarding what these intuitions are.

James Robert Brown argues that our intuition is an account of the laws of nature. The laws of nature, in this case, are "relations among objectively existing abstract

entities" [2]. Thus, Brown proposes a form of Platonism similar to Platonic accounts of Mathematics where mathematical objects like numbers are treated as real abstract entities. Given that we do not know how Platonic intuition works, it appears to many as miraculous. It is for this reason that the position has few supporters. For many, Platonic intuition can be replaced with a far more familiar faculty: sense perception.

Elke Brendel argues that intuitions are "mental propositional attitudes accompanied with a strong feeling of certainty" [2]. These "mental propositional attitudes", or beliefs and prejudices, are developed through our sensory experiences and biological instincts. Thus, Brendel proposes a form of Naturalism since it fully describes intuition by natural phenomena. A strength of this position is that it recognizes that our intuitions change and adapt. For example, the intuitive claim that all infinities are the same turns out to be false with closer inspection. Brendel remarks that, although our intuitions are relative, we would be foolish to ignore them since they serve an important step in acquiring knowledge.

Regarding the epistemological challenge, thought experiments acquire knowledge through the intuitions developed. While appealing, the intuition-based account says little about distinguishing trustworthy from misleading intuitions. Given this shortcoming, thought experiments may require more than just intuition to be compelling.

(C) The Argument View

The view that thought experiments are arguments in disguise. This is based on an empiricist philosophy which deems experience as the only source of knowledge. While there are many nuanced positions based on this view, we will look at one of the more influential advocates.

John D. Norton argues that all thought experiments are arguments that follow a chain of reasoning from premises grounded in experience. The narrative and imagery accompanying thought experiments, then, are purely cosmetic which serve only for rhetorical purposes. Take Galileo's falling body experiment again.

Although no empirical data was analyzed, we were led to accept that objects fall at the same speed regardless of mass. While Brown would attribute this knowledge to intuition, Norton would find an associated argument and claim that our knowledge came from the implicit premises. In this case, the premise that composite objects preserve the property of their parts was crucial in showing a contradiction to Aristotle's theory. Fortunately, the premise is supported on empirical grounds, so Galileo's experimental result holds.

Interestingly, the argument view shares many similarities with the Naturalist flavor of the intuition-based account since both rely on experience as the basis of their theories. However, they disagree on where the conclusions arise. For Brendel, knowledge comes directly from the experiences we had which got translated to intuition. Norton, on the other hand, argues that knowledge is a product of inference on our interpretations of our experiences.

Norton takes a rather strong position compared to other empiricists. He argues that any other account of thought experiments implies a commitment to "asking the oracle". If Norton is right that every thought experiment has an associated argument, then we would do best to skip the oracle and head straight to the more objective argument. This gives the position its strength.

Of all the views presented, the argument view is likely the best at justifying thought experimental results. Regarding the epistemological challenge, thought experiments acquire knowledge through the premises introduced. However, where the premises come from remains unanswered since it is unclear why some experiences are used over others. On top of that, the reconstruction from thought experiment to argument is not as clear cut as hoped.

(D) The Mental-Model Account

The view that thought experiments are simulated mental models. This is based on the observation that in certain problems people reason by manipulating models of situations or events. Similar to physical models, mental models target the underlying phenomenon of the scenario, which is why they can be so informative. Take Judith Thomson's violinist experiment:

> [I]magine a famous violinist falling into a coma. The society of music lovers determines from medical records that you and you alone can save the violinist's life by being hooked up to him for nine months. The music lovers break into your home while you are asleep and hook the unconscious (and unknowing, hence innocent) violinist to you. You may want to unhook him, but you are then faced with this argument put forward by the music lovers: the violinist is an innocent person. All innocent persons have a right to life. Unhooking him will result in his death. Therefore, unhooking him is morally wrong. However, the argument [...] does not seem convincing in this case. You would be very generous to remain attached for nine months, but you are not morally obligated to do so. [2]

Clearly, this thought experiment parallels a popular anti-abortion argument of the same structure. The conclusion that abortion may sometimes be morally

permissible is disputed of course. However, the experiment also has an underlying conclusion which we cannot so easily dismiss. Specifically, the experiment effectively distinguishes two concepts previously ran together: the right to life and the right to what is needed to sustain life.

While it is tempting to claim that we used our moral intuition to reach the underlying conclusion, the intuition in this experiment contradicts our intuition for the actual anti-abortion argument; which one is correct? The ambiguity should be no surprise since the intuition-based account does not mention how to distinguish trustworthy from misleading intuitions. Similarly, in order to reconstruct Thomson's violinist experiment as an argument, we would need as premise the underlying conclusion. Although we may have always known this premise from experience, we were not aware of it until the end which begs the question where did the premise come from.

The mental-model account may be more accurate in this case. We were given a model of a scenario that, ubenknownst to us, is governed by the underlying conclusion. In manipulating the model, we exposed the phenomena and learned something new in the process. What exactly is this process?

Nancy Nersessian argues that the narrative of a thought experiment "functions as a kind of user-manual for building the model, but it isn't identical to a thought experiment" [2]. Once we have built a model in our minds, the mental model gains a life of its own as we simulate its features. This is a thought experiment as viewed by Nersessian.

Since the mental-model account matches with how we use thought experiments, it is no surprise that this position garners the most followers. Regarding the epistemological challenge, thought experiments acquire knowledge through the phenomena they interact with. Unfortunately, the mental-model account makes no attempt at explaining how we are justified in reaching the conclusions.

In reviewing the four most common responses to the epistemological challenge, a common point of concern arises: how can we be sure if a thought experiment accurately reflects reality? The skeptical objection outright denies that thought experiments could ever represent reality. The intuition-based account fails at identifying intuitions we can trust and so the reliability of experiments are left undetermined. The argument view shifts the goalpost to figuring out whether our initial premises are true. And the mental-model account makes no mention as to whether reality ever encounters the phenomena studied in thought experiments.

The argument view is the most straightforward in its resolution: we would need to also validate our premises. As mentioned before, this makes the position most likely

the best at justifying our thought-experimental claims about reality. Unfortunately, reconstructing a thought experiment into its argument form loses its typical force unlike in the intuition-based and mental-model accounts. Can we find an account of thought experiments that is as rigorous as the argument view and as epistemically appealing as the mental-model account?

Going back to the basics, we know that thought experiments involve the imagination. Lucky for us, there has been extensive investigation on the relationship between imagination and knowledge. Specifically, imagination appears essential in acquiring what is referred to as modal knowledge. Maybe this is the key we need to understand how thought experiments can lead us to knowledge. What follows is a brief review of possible worlds and modality as discussed in the Stanford Encyclopedia of Philosophy's articles titled "Possible Worlds" and "Modal Logic".

Although we are directly aware only of our immediate surroundings, many of us believe that our current situation is part of "a series of increasingly more inclusive, albeit less immediate, situations": the situation of you reading this paper is part of the one in the city you live in, the country, the continent, the Earth, the Solar System, the Milky Way galaxy, and so on [3]. It seems reasonable to assume that this series has a limit, "a maximally inclusive situation encompassing all others", the actual world [3].

Many of us also believe that history could have turned out very differently: galaxies might have never formed; dinosaurs might have never gone extinct; wars won might have been lost; we might not have been born. Either way, "no matter how things had gone they would still have been part of a single, maximally inclusive, all-encompassing situation, a single world" [3]. Thus, the actual world, the one we are a part of, may be one among many possible worlds.

In similar fashion, we may also believe that some things just absolutely cannot happen: the number 1 cannot not equal itself, we cannot both occupy the same point in space at the same time, something cannot be simultaneously true and false. Maximally inclusive situations that contain these absolute impossibilities are impossible worlds.

The concept of possible and impossible worlds is quite appealing for its creativity. Beyond being imaginatively interesting, possible worlds provide an elegant solution to the semantics of modal logic known as "Possible Worlds Semantics".

Modality refers to expressions involving modals (like "necessarily" or "possibly"). Modal logic, then, studies the deductions we can make from modal expressions [4]. To be clear with our usage of modals, we introduce two new operators, "$\Box$" for "it is necessary that" and "$\Diamond$" for "it is possible that", to the standard classical operatives, "$\wedge$" (for "and"), "$\vee$" (for "or"), "$\neg$" (for "not"), and "$\rightarrow$" (for "if...then"). Therefore, a sentence like $\Box p \rightarrow \Diamond p$ can be read as "if it is necessary that p, then it is possible that p" or more

succinctly "if p is necessary, then p is possible".

In order for us to determine if a modal argument is valid, we need to understand what it means for a modal proposition to be true or false. In other words, what do we mean when we say that a modal claim like $\Box p$ is true? This is where Possible Worlds Semantics comes into play.

1. $\Diamond p$ is true if and only if p is true in some possible world.
2. $\Box p$ is true if and only if p is true in all possible worlds.

With these semantics, we can now make sense of the validity of modal arguments and begin to form a rigorous logic. Many modal logics have been developed with varying degrees in expressive power. The specifics of these logics is outside the scope of this paper, however, you can read more about it in [4] where they discuss in depth logics like $S4$ and $S5$. Instead, we will look at the general mechanisms on which modal arguments are formed.

When we make modal arguments, we may desire to restrict the range of worlds. For example, I might say that it is necessary for me to sleep, even though I know there is a possible world where I neglect sleep. If we want to remain consistent with our usage of $\Box$, we need to say that "$\Box A$ is true in [world] w [if and only if] A is true in all worlds that are related to w in the right way" [4]. To accomplish this, we impose a binary relation R (referred to as the accessibility relation) on the set of worlds. If wRu holds, then we say u is accessible from w. Thus, we end up with the following updated semantics:

1. $\Diamond p$ is true in w if and only if p is true in some world accessible from w.
2. $\Box p$ is true in w if and only if p is true in all worlds accessible from w.

How we define the accessibility relation will determine the modal logic we can use [4]. Because of this dependence on accessibility, we can neatly accomodate multiple modal logics at once.

The function of imagination in all of this is to populate the set of worlds so we can use our modal logics to derive modal facts about the actual world. A point of contention when imagining worlds, however, is figuring out whether they are possible or not. Numerous tactics have been proposed to clarify the issue. For our purposes, we will only focus on the fact that, in any case, imagining is an essential step to accessing worlds from which we can infer modal knowledge.

Since thought experiments are in the business of imagination, we can reasonably say

that they have the capacity to lead us to modal knowledge. Now with this background knowledge under our belt, we continue with a discussion on a possible solution (pun intended) to the epistemological challenge faced by thought experiments.

Discussion

If we can account for thought experiments under possible and impossible worlds, we can take advantage of all the work done on modal logic to form an epistemology of thought experiments. Fortunately, thought experiments lend themselves rather naturally to the concept of possible and impossible worlds due to their imaginative core. With this opening, we will now explore what I term the "worlds account" of thought experiments.

When we perform a thought experiment, we imagine some scenario ϵ that we would like to reason about. We can go one step further and place this experiment inside an imagined world w. This motivates the following definition:

(INST) ϵ is a thought experiment provided it is in at least one imagined world w.

Importantly, we make no commitment as to whether w is possible in any mode. This allows us to account for impossible thought experiments as well. Next, we need a set of worlds over which we can meaningfully discuss modality.

(SET) For each thought experiment ϵ, we define W_ϵ as the set of worlds that evaluate all assertions inferable from ϵ as true.

Like before, we do not restrict W_ϵ to possible worlds. As explained in the Stanford Encylopedia of Philosophy's article titled "Impossible Worlds", modal logics over arbitrary worlds is severely limited because of the lack of structure. Thus, if we want to infer any useful modal facts, we need to impose much stricter accessibility relations over relevant worlds [5].

Fortunately, Wulf Rehder identified three accessibility relations that generate much more expressive modal logics. Namely, he associates the physical equivalence relation R_1 with realizable thought experimens, the homeomorphism relation R_2 with idealized thought experiments, and the semantic equivalence relation R_3 with abstract thought experiments. Each generates S5, S4, and B modal logics respectively [6].

A crucial condition for Rehder's model to work is that W_ϵ be restricted to only possible worlds. Unfortunately, this denies inferring directly from thought experiments in impossible worlds and undetermined worlds. But we need not be discouraged

since there are a couple of methods we can employ that will allow us to infer modal knowledge from (A) the impossible and (B) the undetermined via counterpossible reasoning. The following two approaches are inspired by Anand Jayprakash Vaidya and Michael Wallner's work on modal epistemology.

(A) Impossible Worlds

(Dispensability Thesis) *If a proposition p entails a contradiction by virtue of the essence of some object(s) x, then p is necessarily false. Symbolically,* $\Box_x(p \rightarrow \bot) \rightarrow \Box\neg p$.

In other words, if a proposition must always lead to a contradiction, then it must be false; *reductio ad absurdum* in action. Given that the actual world is a realized possibility, it must satisfy all necessities for possible worlds. Thus, if we can prove that an impossible thought experiment is essentially inconsistent, we would have proven a modal fact about the actual world.

An epistemology of essence is required to complete these proofs; a topic well beyond our scope. However, it is sufficient to show that all worlds in W_ϵ are untenable (that is, logically impossible). Take Galileo's falling body experiment ϵ once more.

By (SET), W_ϵ is the set of worlds that confirm the following proposition p: (1) Classical logic holds, (2) Aristotle's theory of falling bodies holds, and (3) Principle of Compositionality holds.

By (INST), we are guaranteed that there is a world w in W_ϵ that contains ϵ completely. If we carry out Galileo's experiment in w as described before, we end up with a contradiction which makes w an impossible world.

Since all other worlds in W_ϵ do not seem to be relevantly different to w, they must also hold contradictions derived from p. Therefore, all worlds in W_ϵ are untenable and so, by the dispensability thesis, p is necessarily false for possible worlds.

Note that in this example we were only able to prove that the claims in p cannot be simultaneously true, but in the original experiment, we specifically accused Aristotle's theory as the culprit. What gives? Turns out Galileo's experiment relied on more than just Aristotle's theory to reach the contradiction which is why p has multiple claims. However, since we can be reasonably assured that (1) and (3) are true, we are left with (2) as the only troublemaker.

(B) Undetermined Worlds

(Exportation Thesis) *If some world w is a situation in a possible world u, then each proposition p of w is possible. Symbolically,* $w \subseteq u \rightarrow \forall p(v_w(p) \rightarrow \Diamond p)$

In other words, if a possible world contains a model of some arbitrary world, everything we can say in this arbitrary world might actually be true. This is based on the observation that we can model situations that may not hold in general. Since these models are part of a possible world, anything we can say in them will indeed be possible.

Note that when w is a possible world, $u = w$. In this special case, all the propositions of w are possible by the exportation thesis, which matches with what we expect of any possible world.

For all other cases, the result that all propositions in w are possible is grounded on *modus ponens*. Namely: p is true if certain conditions are met; a possible world u meets these conditions by w; therefore, p is true in u. The power of this thesis is that it allows us to prove positive modal claims without demanding that the reference world be possible.

Consider the following active field in philosophy and mathematics: non-classical logics. As the name suggests, this field researches the properties and consequences of logics unlike classical logic. Prominent examples include intuitionistic logic (classical logic without law of the excluded middle), paraconsistent logics (any logic without principle of explosion), and fuzzy logic (logic dealing with partial truths). In a very broad sense, we reason about these logics through thought experiments that assume the truth of one of these.

Clearly, all of these logics are not equivalent to classical logic and will say very different things about the truth of various statements. As a consequence, worlds possible in one logic may not be possible in another. Critically, we do not know for sure that our world obeys classical logic. However, if we find a model of any of these logics in a possible world, we can pull all the modal facts from worlds of that logic via the exportation thesis without having to assume that our world truly obeys that logic.

Based on the above analysis, the worlds account describes the process of thought experimenting as follows:

1. When performing a thought experiment ϵ, we reason through the situation in some world w. Note, all conclusions we make from ϵ are only shown to be true in W_ϵ. That is, we have said nothing about the actual world.

2. To change that, we need to show either that (1) at least one world in W_ϵ is contained in a possible world or (2) ϵ is necessarily impossible. Neither is exactly easy to prove. However, once a condition is met, inference to our world becomes straightforward.

Regarding the epistemological challenge, thought experiments acquire knowledge through the essence of relevant object(s), provided we succeed in the second step.

Limitations

We have seen that under the worlds account any thought experiment ϵ can lead to modal knowledge provided we know the composition of the corresponding set W_ϵ. Does this fare better than the other accounts discussed previously? The two most promising accounts are the argument view and the mental-model account.

In the first step of the worlds account, we reason through the experiment much like the mental-model account. Thus, the worlds account has epistemic appeal. Where they differ, however, is determining which world the conclusions apply to. The mental-model account claims that the results of simulating a mental model automatically applies to the actual world. In contrast, the worlds account claims that the results so far have only been shown to apply to the world where the experiment took place. Since thought experiments are frequently misleading, it may be better to take the more conservative approach (that is, the worlds account).

In the second step of the worlds account, we attempt to establish an epistemic bridge between the set of worlds that the thought experiment holds in and the actual world. This is much like the argument view when we attempt to show that the premises are sound. The main difference is in the outcome if we are successful in establishing the connection. While the argument view leads you directly to the results of the thought experiment, the worlds account frustratingly requires additional steps to show that the results are actual possibilities, not just mere ones. The reason for this discrepancy is that the argument view uses empirical facts (committed to actual) while the worlds account uses essentialist facts (committed to possible) to bridge the gap.

As a consequence, the worlds account requires an epistemology of essence to complete proofs that establish some relationship between the actual world and the worlds in W_ϵ. Unfortunately, knowledge of essence is rather difficult to justify, barring some trivial cases, unlike knowledge of empirical facts. This point may be the strongest objection to the worlds account.

Lastly, the argument referenced in the argument view is a nontrivial reconstruction of the thought experiment which is relatively clear once we know how the setup interacts to lead to the conclusion. In contrast, the argument referenced in the worlds account is a proof that relates the thought experimental worlds to the actual world which does not have as obvious of a construction. Thus, the worlds account provides epistemic appeal at the expense of added complexity to the justification procedure.

Whether the additional complexity is unreasonable depends on our confidence in a given thought experiment. For relatively mundane thought experiments, an argument-view reconstruction is likely sufficient even if we may not know entirely where the premises came from. On the other hand, more intruiging thought experiments are usually nuanced and controversial. With these experiments, the stronger demands by the worlds account may be ideal for handling the intricacies of the subject.

Future Scope

Given the very broad definitions and procedure in the worlds account, it may be possible and worthwhile to generalize the account to a much more expansive definition of thought experiments that include not only scientific and philosophical thought experiments, but also fictions and other storytelling mediums that prime us to imagine alternative worlds. A reason we may want to include some literary works under the concept of thought experiments is because many narratives, while entertaining, also serve as commentary about our world. Take dystopian novels like George Orwell's "1984" and Joseph Heller's "Catch-22" that explore fictional worlds that share many characteristics with our own. Or mathematical fictions like Edwin Abbott Abbott's "Flatland: A Romance of Many Dimensions", Hamilton Luske's "Donald in Mathmagic Land", and John D. Barrow's "The Infinite Book" that explore fantastical worlds featuring mesmerizing mathematics and logic.

Of course, before going on applying the worlds account to the other facets of imagination, it is crucial that we resolve the concern over having an epistemology of essence. As such a study into the theories of essence and how they intersect with thought experiments would greatly contribute to the subject matter. On similar footing, finding other ways to derive modal facts from impossible worlds and undetermined worlds can enrich the worlds account with a much more approachable toolset that philosophers can use to ease justification of a given thought experiment.

Conclusion

Ultimately, we ended with an account of thought experiments that uses the concept of possible and impossible worlds to explain how we acquire knowledge by just imagining contrived scenarios and how we are justified in doing so. The actual act of thought experimenting remains the same. The hope of this paper is to aid in what comes after, deriving knowledge from our imagination. Particularly, to provide a framework on which we can rigorously and coherently justify claims about our world from thought experiments.

Since the worlds account directly concerns itself with what is possible, hopefully it can

also be used as a tool to judge the efficacy of thought experiments that go well beyond what is real. This is especially useful when we consider thought experiments in ethics and metaphysics where bizarre and fantastical scenarios abound like Robert Nozick's utility monster and Derek Parfit's amoeba people.

* * *

References (In order of appearance)

[1] Angelucci, Adriano, and Margherita Arcangeli. “Introduction: New Perspectives on Philosophical Thought Experiments.” *Topoi*, vol. 38, no. 4, 2018, pp. 763–768. *Crossref*, doi:10.1007/s11245-018-9580-2.

[2] Brown, James Robert, and Yiftach Fehige. "Thought Experiments." *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, Sep. 2019, plato.stanford.edu/archives/win2019/entries/thought-experiment/.

Annotated Bibliography

[3] Menzel, Christopher. "Possible Worlds." *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, Feb. 2016, plato.stanford.edu/archives/win2017/entries/possible-worlds/.

After reading about thought experiments and imagination, the reference to modality led me to research more on the subject. Turned out almost all the literature on the topic assumes that I already knew about possible worlds semantics so that meant I needed to learn that first.

The SEP article, like always, was fascinating as it discussed the various philosophical conceptions of what a possible world may be. What especially struck me as important was the discussion on how possible worlds provide a rigorous framework for understanding modal sentences like "it is necessary that p" and "it is possible that p". Specifically, it defined these modalities as quantifiers over some set of worlds.

Now with a suitable understanding of possible worlds and possible worlds semantics, I can dive into the technicalities of modal logic and modal epistemology. In addition, the "possible" in possible worlds hinted to me that there might also be a concept for impossible worlds which would be great for reasoning about worlds that are logically contradictory, something that happens quite a lot when thought experimenting.

[4] Garson, James. "Modal Logic." *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, Sep. 2018, plato.stanford.edu/archives/sum2021/entries/logic-modal/.

Almost every source I have read so far had at least one mention to modality. As a subject in and of itself, modality is very intricate since it involves the study of rather extroardinary claims like necessity and possibility. Something I was even less certain about was how we can make any formal argument usings modals without resorting to fanciful thinking. To clear some of my headaches, I dove into

a trustworthy article about the logic of modality.

The article, in its entirety, was overwhelming as it attempted to condense decades of research into a single comprehensive entry. Although I cannot say that I am an expert now in the field, the article did leave me well informed on the relevant issues and solutions when it comes to logically manipulating modalities via accessibility relations and conditioned frames

Almost the entire article assumed a set of possible worlds over when defining each modal logic. The next question is if it can also be applied to impossible worlds even though almost all of them inherit logical impossibilites.

[5] Berto, Francesco, and Mark Jago. "Impossible Worlds." *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, Aug. 2018, `plato.stanford.edu/archives/fall2018/entries/impossible-worlds/`.

While reading about modal epistemology, I attempted to develop modal claims that could be associated with particular thought experiments. Reasoning about positive assertions was straightforward. However, I got stumped with negative assertions since their respective worlds turned out not to be possible. This got me thinking about how one could reason about impossible scenarios. As with any unfamiliar philosophical concept, I went straight to the SEP for a crash course on impossible worlds.

I found illuminating the article's treatment of modal logics in (im)possible worlds. Notably, we can think of modal truth as world-dependent where modal sentences on statement A hold in a world w depending on if A holds in all or some of the accessible worlds of w. While that description was a bit technical, the gist is that it allows one to accommodate for different modal logics even in impossible worlds where contradictory facts abound. This is similar to how we can meaningfully debate between classical and non-classical logic by supposing other axioms while still using the same set of mathematical concepts.

With this conception of modal truth, it would be helpful to get a bit more familiar with quantifiers and the accessibility relation in modal logic. This is because thought experiments typically do not commit themselves to any particular world beyond having the essential ingredients for the experiment. Thus, to think sensibly about thought experiments as providing substantial modal knowledge, it would be beneficial to understand how to group worlds into classes from which arguments can be made.

[6] Rehder, Wulf. "Thought-Experiments and Modal Logics." *Logique Et Analyse*, vol. 23, no. 92, 1980, pp. 407–417. *JSTOR*, `www.jstor.org/stable/44083924`.

After reading all the previous sources (and then some) and contemplating their significance to thought experiments, I was able to narrow down my subject to justifying modal knowledge via thought experimentation. With this more developed lexis, I now could look into sources that directly talk about my research interest. By doing this, I may be able to find open problems I can tackle.

To my surprise and glee, the author in this paper proposed a categorization of thought experiments based on the modal logic they correspond to. This formal analysis aims to make explicit the valid inferences one can make about the actual world from thought experiments conceived in possible worlds. Particularly, the author identified three types of thought experiments feasible, idealized, and abstract that correspond to modal systems S5, S4, and B, respectively.

Since this is the first source I've read directly about my subject, I can't confidently say that I am aware of the open problems of the field. However, the technical subtleties crucial to the paper did open up a new rabbit hole. Namely, instead of linking thought experiments in possible worlds to the actual world (which is a difficult thesis to defend), what if we allow for arbitrary world thought experiments to which the actual need not apply?

[7] Vaidya, Anand Jayprakash, and Michael Wallner. "The Epistemology of Modality and the Problem of Modal Epistemic Friction." *Synthese*, vol. 198, Apr. 2021, pp. 1909–1935. *EBSCOhost*, doi:10.1007/s11229-018-1860-2.

While researching subjects for the "Intertextuality" assignment, I landed on a Stanford Encyclopedia of Philosophy article about imagination which had an entry on knowledge (which seemed relevant to my research subject: thought experiments). In that entry, it referenced applications of imagination to modal epistemology (also known as the epistemology of modality), so I thought I'd read up on modality in the SEP. There were multiple mentions of there being various types of modal judgments. While I understood necessity and possibility, I was curious about how essentialist claims fit into the picture. This ultimately led me to the source above, which analyzed the relationship between the epistemology of modality and the epistemology of essence.

The authors argued that the three main theories of modal knowledge must also provide an epistemology of essence. In other words, each account fundamentally generates modal knowledge by epistemic friction with essentialist claims. While the thesis is fascinating in its own right, I especially took note of the descriptions the authors gave for each theory. Namely, they described general procedures for generating modal knowledge via conceivability, counterfactuals, or deduction.

As far as I am aware, the closest account of thought experiments that resembles

the use of possible worlds is the mental-model account. However, this model is intended only for the benefit of probing questions about reality. If I want to account for all suitable thought experiments as experiments in possible worlds, I'll also need to consider contradictory thought experiments that violate conditions for being mere possibilities. This means researching the nature of impossible worlds and how one can go about using them.

* * *