2022

Impossible worlds and the safety of philosophical beliefs

Zack Garrett

Zachariah Wrublewski

DOI: 10.1111/meta.12550

ORIGINAL ARTICLE

Impossible worlds and the safety of philosophical beliefs

Zack Garrett[1] | **Zachariah Wrublewski**[2]

[1]Excelsior Classical Academy, Durham, North Carolina, USA

[2]Department of Philosophy, University of Nebraska, Lincoln, USA

Correspondence
Zack Garrett, Excelsior Classical Academy, 4100 North Roxboro Street, Durham, NC 27704, USA
Email: zackgarrett127@gmail.com

Epistemological accounts that make use of a safety condition on knowledge, historically, face serious problems regarding beliefs that are necessarily true. This is because necessary truths are true in all possible worlds, and so such beliefs can be safe even when the bases for the beliefs are epistemically problematic. The existence of such problematically safe beliefs would undermine a major motivation for the condition itself: the ability to evaluate how well a belief tracks the truth. This paper argues that incorporating impossible worlds into the evaluation of beliefs solves this problem, but only if the relevant account of impossible worlds entails that many impossible worlds are incredibly similar to the actual world. Further, the paper argues that, as a result of including impossible worlds, some philosophical beliefs are unsafe, and many more are potentially unsafe. But, it argues, even if this is the case, we can still make philosophical progress.

KEYWORDS
compositionality, impossible worlds, metaphilosophical skepticism, philosophical progress, safety, similarity

1 | INTRODUCTION

When considering the necessary conditions on knowledge, there is intuitive appeal in accepting conditions that properly connect our beliefs to the truth of their objects. In large part, modal conditions on knowledge, such as the sensitivity and safety conditions, are motivated by the aim of ensuring that knowledge maintains this connection. Further, given some well-known problems with the sensitivity condition, the safety condition seems to be more plausible for this purpose. But, the safety condition on knowledge faces problems with maintaining this connection in specific cases—cases in which the relevant beliefs are necessarily true.

In this essay, we show that the problems for the safety condition arise because analyses involving the condition have only included possible worlds in the evaluation of safe beliefs. Further, we argue that including *impossible* worlds in these analyses enables the safety condition to avoid the problems mentioned above. Next, we offer and defend an account of impossible worlds, and the inclusion of impossible worlds in safety conditions. Lastly, we outline and consider what we believe are potentially substantive ramifications for philosophical knowledge: namely, that making such a move would mean that we should be skeptics about many potential instances of philosophical knowledge.

2 | SAFETY AND TRIVIALITY

2.1 | Sensitivity and safety

Given that the focus of our project is modal conditions on knowledge, it's prudent to start our discussion with a brief history of the most prominently discussed and supported modal criteria. Generally speaking, modal conditions on knowledge are meant to provide resources for epistemological theories to ensure that if one's belief amounts to knowledge, it properly tracks the truths related to the belief. In *Philosophical Investigations*, Robert Nozick argues for this connection straightforwardly, supporting the necessity of the *sensitivity* condition to the concept of knowledge. He formulates the sensitivity condition as follows:

> *Sensitivity:* If p weren't true and S were to use M to arrive at a belief whether (or not) p, then S wouldn't believe, via M, that p. (1983, 179)

Importantly, Nozick intends the conditional expressed in *Sensitivity* to be a subjunctive conditional. When understood in terms of a possible-worlds locution, satisfying *Sensitivity* requires it be the case that in the nearby possible worlds in which p is false and S uses M to arrive at a belief whether (or not) p, S doesn't believe, via M, that p.

The sensitivity condition as outlined above faces several well-known challenges.[1] But, Nozick's aim in attempting to formulate such a condition seems amicable. Intuitively, if a belief amounts to knowledge, it should track the relevant truths.

In order to account for this aim while avoiding the problems identified for the sensitivity condition, many theorists opt for a similar, but importantly different, condition on knowledge: the safety condition. An early formulation of this condition by Ernest Sosa goes as follows:

> *Sosa (SF):* If S were to believe that p, then p would be true. (1999, 141–53)

[1]For some discussion of general problems for sensitivity, see Vogel 1987; Sosa 1999; Kripke 2011; and the critical essays in Luper-Foy 1987.

As formulated by Sosa, this condition does not include reference to the method by which S comes to believe p. As many commentators subsequently accepting safety think this is an important feature of any such criterion, we can modify this original definition to include such a link:

Sosa (SF):* If S were to believe p via M, then p would be true.

As in *Sensitivity*, the conditional expressed here is meant to be a subjunctive conditional. Using a possible-worlds analysis, the basic idea would be as follows: in order for a belief to be safe, it must be the case that in the nearby possible worlds in which S believes p via M, p is true. Not only does this formulation avoid the problems for the sensitivity condition, it also (purports to) maintain the "truth-tracking" phenomenon that modal conditions are meant to ensure.

2.2 | Safety and necessity

As mentioned above, the sensitivity condition faces several well-known objections, and so in this essay we largely leave out discussions of the sensitivity condition and narrow our focus on the problems such beliefs pose for the safety condition. The problems that ultimately plague safety conditions on knowledge stem from the triviality associated with specific kinds of counterfactuals—in particular, *counterpossibles* and counterfactuals with necessarily true consequents.

A counterpossible is a counterfactual conditional in which the antecedent is necessarily false. The following are examples of counterpossibles:

1. If Hobbes had (secretly) squared the circle, all sick children in the mountains of South America at the time would have cared.
2. If Hobbes had (secretly) squared the circle, all sick children in the mountains of South America at the time would not have cared.[2]

Given that it's necessarily false that Hobbes squared the circle (and, more so, secretly!), the antecedent of both conditionals is necessarily false. Or, in a possible-worlds analysis, there are no possible worlds in which the antecedent is true. Accordingly, it seems, both of the conditionals (as well as all other counterpossibles) are *trivially true*.[3]

Similarly, counterfactual conditionals with necessarily true consequents face problems with triviality. For example, consider the following counterfactuals:

3. If Tokyo were the capital of Spain, then 2 + 2 = 4.
4. If I were to wear a red shirt tomorrow, then 2 + 2 = 4.

When analyzing (3) in terms of possible worlds, here's how the analysis goes. In order to determine whether this conditional is true, we look at the closest possible worlds in which Tokyo is the capital of Spain. While it's not exactly clear which of these worlds would be closer to the actual world than others—for example, is the world in which Spain takes control of Japan after World War II and moves the Spanish capital to Tokyo closer to the actual world than the world in which Spain colonizes Japan during the Sengoku period and relocates the Spanish

[2]Both of these counterpossibles are described in detail in Nolan 1997, 554.

[3]David Lewis refers to these as *vacuously* true. While there may be some important distinctions between triviality and vacuity made by some philosophers, none of these differences or distinctions will matter to the argument at hand.

capital to Tokyo (then called Edo)? It's not clear. But, this doesn't matter when considering the truth of such conditionals. Given that 2 + 2 = 4 is necessarily true (that is, true in all possible worlds), we know that in both of these worlds, 2 + 2 = 4. Because of this, we know that both (3) and (4) are trivially true. Furthermore, when we look at *any* possible world, we know that in this world 2 + 2 = 4. Accordingly, conditionals with 2 + 2 = 4 (or any other necessary truth) as a consequent are trivially true.

While potentially problematic for theories of counterfactuals generally, the trivial truth of the sorts of counterfactuals detailed above in possible-worlds analyses presents special problems for modal conditions on knowledge—specifically for our current discussion, the safety condition. Counterfactuals with necessarily true consequents cause the following problem for the safety condition, which we call the "Triviality Problem": the triviality involved in the truth of such counterfactuals undermines one of the central aims of the safety condition on knowledge—ensuring that beliefs which satisfy the safety condition track the relevant truths involved.

To see why this is, consider a case in which the relevant belief under scrutiny is a belief about something that is necessarily true. As formulated above, a general safety condition is as follows: If S were to believe p via M, then p would be true. Further, in the kind of case currently under consideration, p is necessarily true. This means that all possible worlds where S believes p via M are worlds in which p is true. Because p is true in every possible world, it *could* be that S's belief that p is safe, even when M is an epistemologically problematic methodology. The safety of such beliefs undermines what is supposed to be a crucial motivation for modal conditions on knowledge, generally: that they ensure some sort of connection between the relevant belief, the method of coming to that belief, and the truth of that belief.

Generally speaking, most theories of counterfactuals and counterpossibles suggest that we should consider only possible worlds when evaluating their truth-values. Because of this, such theories entail that counterpossibles and counterfactuals with necessarily true consequents are trivially true. But, opening up these theories to allow impossible worlds into the fold would allow for varying analyses of the truth-values of counterpossibles and counterfactuals with necessarily true consequents. To see how this works, consider the following examples from earlier:

1. If Hobbes had (secretly) squared the circle, all sick children in the mountains of South America at the time would have cared.
2. If Hobbes had (secretly) squared the circle, all sick children in the mountains of South America at the time would not have cared.
3. If Tokyo were the capital of Spain, then 2 + 2 = 4.
4. If I were to wear a red shirt tomorrow, then 2 + 2 = 4.

Earlier, we argued that if we consider only possible worlds in the scope of the relevant counterfactuals and counterpossibles, (1) and (2) are trivially true because there are no possible worlds in which Hobbes squares the circle, and (3) and (4) are trivially true because every possible world, regardless of antecedent conditions, would be one in which 2 + 2 = 4 is true. But, if we incorporate impossible worlds in our analyses of the relevant conditionals, *none of* (1), (2), (3), or (4) is trivially true. In the case of examples (1) and (2), the inclusion of impossible worlds in which Hobbes *does* square the circle would mean we would have to look at the closest impossible worlds in which this happens to see whether or not all sick children in the mountains of South America at the time care about this development. If they do care, then (1) is true, while (2) is false; if they don't care, then (1) is false, while (2) is true. Whichever ends up being the case, clearly neither conditional is *trivially* true. Similarly, for (3) and (4), we would have to look at the closest worlds (possible or impossible) in which

Tokyo is the capital of Spain and in which I wear a red shirt tomorrow, respectively. If 2 + 2 = 4 in the closest world in which Tokyo is the capital of Spain, then (3) is true; if not, and 2 + 2 ≠ 4 (that is, the closest world considered is an *impossible* world), then (3) is false. And, if 2 + 2 = 4 in the closest world in which I wear a red shirt tomorrow, then (4) is true; if not, and 2 + 2 ≠ 4, then (4) is false. Regardless of what the truth of the conditionals in (3) and (4) ends up being, clearly neither conditional is *trivially* true.

One way that theorists have attempted to defend the safety condition from the Triviality Problem *without* including impossible worlds is to argue for a "basis-relative" modification of the view. To see how this might go, consider Duncan Pritchard's version of the safety condition:

> *Pritchard (P-SF):* S's belief is safe iff in most nearby possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which S continues to form her belief about the target proposition in the same way as the actual world, her belief continues to be true. (2009, 34)

This condition, as expressed, doesn't quite address the triviality problem as such. But, Pritchard does suggest a way in which such an account might be extended to address the Triviality Problem:

> The way safety theorists like myself respond to this [the Triviality Problem] is to say that once we shift to a basis-relative formulation of safety … then our focus should not be on the particular proposition believed in the actual world, but rather on the doxastic output of the basis in the actual world instead. This means that while there is, of course, no close possible world where one falsely believes the necessary proposition that one actually believes, the kind of haphazard basis described above will lead to lots of false beliefs in close possible worlds (just not false beliefs in the proposition actually believed). As such it will be an unsafe basis. (2020, 209–10)

To summarize the strategy exemplified by Pritchard's discussion above, a basis-relative version of the safety condition holds that merely considering the relevant beliefs themselves is only part of the correct full analysis. In order for a belief to be safe, some argue, the method or basis used to form the belief must be such that it would not lead the agent to false beliefs in close possible worlds.[4]

This type of modification appears in many popular contemporary accounts of the safety condition, such as the following example of Timothy Williamson's formulation:

> *Williamson (W-SF):* If one knows, one could not easily have been wrong in a similar case. (2000, 147)

The short idea inherent in such an approach is this: in cases involving the potential problematic types of beliefs (that is, beliefs that are necessarily true), we should not merely look at the beliefs in question in nearby possible worlds; we should also be looking at the basis for the belief in order to determine whether the basis itself is a safe basis. In Pritchard's terms, we should look to see if one is "lucky" in having a true belief in the case; in Williamson's terms, we should look to see if one "could easily have been wrong" in the relevant case. If

[4]For more discussion of and objections to Pritchard's solution to the Triviality Problem, see Miščević 2007; Bernecker 2010; Melchior 2017; and Melchior 2021.

the belief is formed on a safe basis (that is, if one was not lucky or could not easily have been wrong, and so on) and is otherwise safe, then the belief is safe. If one or the other of these conditions is not met, then the belief is unsafe.

This sort of modification *does* help the safety theorist with some of the relevant potential counterexamples to the condition. For example, consider the following case and suppose that Goldbach's Conjecture is true:

> *Coin-basis:* Arnold formulates his belief that Goldbach's Conjecture (henceforth "G") is true on the basis of a coin flip. In coming to the belief, Arnold uses the following method: "If this coin lands on heads, I will believe G. If it lands on tails, I will believe not-G." Arnold flips the coin, the coin lands heads up, and, as a result, Arnold believes G.

In this example, a non-basis-relative safety condition that doesn't include impossible worlds would struggle with the Triviality Problem because there are no close possible worlds in which G is false. So, according to such a condition, the belief would be safe. But, a basis-relative safety condition can avoid this conclusion because there *are* close possible worlds in which the basis of belief (the coin flip) would lead Arnold to false beliefs—all of the nearby possible worlds in which the coin lands tails up!

While this does save the safety condition from the Triviality Problem in certain cases, it does not do so in *all* cases. Consider a very similar example:

> *Coin-basis*:* Arnold formulates his belief that G on the basis of a coin flip. In coming to the belief, Arnold uses the following method: "If this coin lands on heads or tails, I will believe G. If not, I will not form a belief." Also, suppose that Arnold will use *only* this method when considering whether or not to believe G, and will reliably use this method when confronted with the question of whether or not G. Arnold flips the coin, the coin lands heads up, and, as a result, Arnold believes G.

In this case, the belief-formation method is very similar to that of the first case. But, importantly, *this* method would never lead Arnold to a false belief. Further, Arnold would not *luckily* come to his belief—he would not be lucky to have a true belief (as he will form the belief G in almost all nearby possible worlds), nor would he be lucky to avoid forming a false belief, given that in the odd event that the coin does not land heads or tails up, Arnold will not form a belief at all. Similarly, it's not the case that Arnold *could easily be wrong* in his belief. There is no nearby false belief he would form on the basis of this method (as he employs the methodology only to form a belief about whether or not G), nor is there a similar methodology that Arnold might use to come to some false belief (as he uses this specific methodology reliably and only when considering whether or not G). Intuitively, this seems to be just as bad a method for forming the relevant belief as was the method in *Coin-basis*.[5] It's cases like these that remain a problem for safety conditions that merely range over possible worlds. And while the problematic cases are certainly fewer for these basis-relative safety conditions, the fact that there are problem cases at all—that is, that there are cases in

[5]According to Williamson, the idea of "not easily being wrong" involves looking not just at possible worlds in which an agent might use a particular methodology but also at different methodologies that are in some sense "close." While this response might seem to avoid the problem exemplified by *Coin-basis*,* the viability of this strategy is contentious. Unfortunately, we do not have the requisite space to fully treat this issue here. For more on this type of response from Williamson, see Williamson 2009, 325–28. For more on the potential problems with such a solution, see Hirvela 2019 and Zhao 2021.

which a belief like Arnold's is safe—continues to be a problem for basis-relative safety conditions that do not include impossible worlds.

Before moving on, we should say one more thing about the potential for using impossible worlds when considering the safety condition; in order to adequately consider the possibility of adding impossible worlds to the mix (without yet conceding that we *should* add impossible worlds), we have to consider (and alter) our understanding of the safety conditional and what it should mean to satisfy this conditional. When outlining safety conditions, historically safety theorists have suggested that what's important for safe beliefs is that the safety condition is not falsified—that is, that "if S believes p via M, then p" is not false. This idea lurks in the background of formulations of the condition like, for example, the condition argued for by Williamson—in the idea that one should not easily be wrong in holding safe beliefs. Importantly, safety theorists have also written as if the important part about satisfying the safety condition is that the conditional is true—such as in Sosa's initial formulation of safety. Until now, only worlds that are possible (that is, worlds that obey classical logic and, subsequently, are such that there are no real contradictions) have been considered—meaning, we haven't had to settle the issue of whether it's more important that the safety conditional not be falsified or that the safety conditional be true. But, now that impossible worlds may be added to the mix (as we argue), we *do* have to consider the matter—given that real contradictions may exist in impossible worlds. We suggest that what is important to safety is that the consequent of the safety conditional is true only (rather than both true and false) when the relevant belief is that p is true only. This is because understanding the conditional in this way achieves *both* of our epistemic goals—it ensures that in these cases we have the relevant true beliefs, while simultaneously avoiding relevant false beliefs. For example, in a world where S believes p, and p is both true and false, S has a false belief, falling short of the second epistemic goal. But, in a world where S believes p, and p is true only, S meets both the goal to have true beliefs and the goal to avoid false ones.

For the reasons outlined above, it seems that including impossible worlds in the scope over which safety-related conditionals range may help with the Triviality Problem (specifically considering beliefs that are necessarily true). If the necessary truth relevant to the belief *could be false* in some impossible world, and impossible worlds are included in the worlds we consider when evaluating safety, then these problematic beliefs would not be safe—they would be safe only when the nearby worlds considered are ones where the consequent is true and would be unsafe when the nearby worlds are ones in which the consequent is false. Thus, the central aim of modal conditions on knowledge like the safety condition—that of ensuring that our beliefs are properly connected to the truth—could be salvaged by such an analysis.

In "Sensitivity, Safety, and Impossible Worlds" (2021), Guido Melchior argues that the specifics of an impossible-worlds account will cause problems for using impossible worlds in analyses of the safety condition.[6] In broad strokes, Melchior's argument is as follows: A plausible constraint on the closeness of impossible worlds makes it the case that including impossible worlds will not change whether or not a given belief is safe. To kick off this argument, Melchior considers the Strangeness of Impossibility Condition (SIC).

> *SIC:* Any possible world is more similar (nearer) to the actual world than any impossible world. (Nolan 1997, 550)

If SIC is true, then clearly this is a problem for the impossible-worlds analysis of the safety condition because it would mean that each world we're investigating, if it really is a nearby world,

[6]Melchior further argues that the sensitivity condition escapes the Triviality Problem, and so we should accept a sensitivity condition on knowledge rather than a safety condition.

will be a possible world. So, the evaluations of counterfactuals with necessarily true consequents would be *unchanged*, even if impossible worlds are included in the scope of the worlds considered when evaluating such conditionals.

The other option Melchior considers is the case in which SIC is false. In short, he argues that SIC's falsity would entail that some impossible worlds are closer than some possible worlds and thus could potentially alter the evaluations of the conditionals. But, Melchior contends that we should accept SIC because accounts that reject SIC face problems with conditionals with contingent antecedents and consequents. The basic idea is this: If we reject SIC, then the potential closeness of impossible worlds will lead to counterintuitive results when we consider "normal" counterfactuals (that is, non-counterpossibles, and counterfactuals without necessarily true consequents). In light of these potential problems, Melchior concludes that we should accept SIC.

While Melchior does well to outline the general landscape, there's something further that should be said about SIC and its relation to the safety condition: not only would one have to reject SIC in order to save the safety condition, one would also have to *radically* reject SIC. One way to reject SIC would be to hold that there is *some* impossible world that may be closer to the actual world than some far-out, deeply odd possible world. Call this the *weak rejection of SIC*.

If one weakly rejects SIC, this still wouldn't be enough to salvage the safety condition in the face of the Triviality Problem. Because the safety condition only considers worlds near the actual world in which the antecedent is true, the addition of one far-out impossible world to the mix would never change the evaluation. In other words, if rejecting SIC is just a matter of accepting that one or two impossible worlds are closer than some extremely distant possible world, then these worlds won't be close enough to the actual world to change the evaluation of the conditional.

According to the *radical rejection of SIC*, not only is some impossible world closer to the actual world than some extremely distant possible world, *many* impossible worlds are closer to the actual world than *many* possible worlds. Further, this radical rejection of SIC is necessary for salvaging the safety condition because the only way in which an impossible-worlds analysis would differ from a traditional possible-worlds analysis of the same conditional is if the impossible worlds were both close enough and plentiful enough to change some safety evaluations. In what follows, we argue that we *should* radically reject SIC and, thus, contra Melchior, an impossible-worlds analysis of the safety condition is viable.

3 | IMPOSSIBLE WORLDS, CLOSENESS, AND SIC

3.1 | Starting assumptions

Before discussing our objections to SIC, we should elucidate our basic assumptions about impossible worlds. To begin, we follow Graham Priest in treating impossible worlds as ones that obey a different logic from one another.[7] "One might wonder, therefore, what makes a world impossible. Answer: an impossible world is one where the laws of logic are different from those of the actual world (in the way that a physically impossible world is a world where the laws of physics are different from those of the actual world)" (2014, xxiii). Treating impossible worlds as worlds with different logical laws is intuitively appealing. As Priest mentions, doing so mirrors our understanding of physically impossible worlds. Another benefit is that this definition subsumes many other definitions. For example, the view that Francesco Berto and Mark Jago call "contradiction-realizers" claims that impossible worlds are ones where sentences of the form φ

[7]Sandgren and Tanaka (2020) argue that there are two kinds of logically impossible worlds. Our set of impossible worlds is a subset of theirs, and so if our argument works here, then it works under their account as well.

and ~ φ both hold (2019, 32). A world that realizes a contradiction would, however, be one that obeys logical laws different from those the actual world obeys.[8]

As for what counts as a world, it is clear that we must accept some ersatz theory. For reasons described by Berto and Jago, realist theories struggle to handle impossible worlds (2019, 44–47). Berto and Jago argue for an ersatz theory on the grounds that it is the only way to accommodate impossible worlds, which prove to be incredibly useful philosophical tools. Ersatz theories of worlds treat worlds as maximally consistent sets of sentences.[9] A set of sentences is maximally consistent if, for any sentence φ, exactly one of φ or ~ φ is a member of the set. Since we are making use of impossible worlds, we can drop the "consistent" part of this definition. The definition of a world, then, would be a maximal set of sentences—a set of sentences for which at least one of φ or ~ φ is a member. One may even remove the maximal requirement. If the correct logic in a world *n* allows for truth-value gaps, then there may be some φ and ~ φ such that neither is a member of *n*. There could still be some constraints on which sets of sentences count as worlds. For example, in the next few paragraphs we argue that worlds are constrained by the meanings of the words that appear in the sentences that are their members.

Impossible worlds are not without controversy, and some may reject our approach to safety on the grounds that impossible worlds come with too much baggage. We feel that other philosophers like Daniel Nolan (1997) and Berto and Jago (2019) do a good job of defending the use of impossible worlds, but here we will briefly handle one objection that we feel has not received adequate treatment yet. Some reject the use of impossible worlds because they undermine compositionality. Impossible worlds appear to allow synonymous sentences to receive different truth-values. This would mean that the words that compose the sentences do not determine the meanings of the sentences. For example, since impossible worlds can include contradictions, we can have a man, Winston, for whom it is true that he is a bachelor and false that he is an unmarried man. Since "bachelor" is synonymous with "unmarried man," it would appear that two synonymous sentences receive different truth-values, and hence their truth-values cannot be a product of the components that make up the sentences. For many, the loss of compositionality is too costly.[10]

Luckily, we do not have to give up compositionality. Impossible worlds are not real metaphysical entities like David Lewis's concrete possible worlds. Ersatz theories of worlds often treat them as linguistic entities, like sentences or sets of sentences. As linguistic entities, impossible worlds are constrained by what language can sensibly describe—more specifically, they're constrained by analytic entailment. A sentence φ analytically entails ψ iff ψ follows from φ by virtue of the meanings of the words in φ and ψ. Amie Thomasson describes analytic entailment as follows: "It is in part constitutive of the meaning of 'house' that all houses are buildings, so that the truth of 'X bought a house' is sufficient for the truth of 'X bought a building': if we know the truth of the first, the meanings of the terms, and have reasoning abilities, we can infer the truth of the second claim on that basis alone" (2007, 28). "Winston bought a house" analytically entails "Winston bought a building." The former cannot be true when the latter is false, but this isn't because of the logical structures of the sentences. Instead, the entailment comes from the meanings of "house" and "building." Because worlds in an ersatz theory are linguistic entities, they will make use of words like "house" and "building." They will be constrained by analytic entailments.

[8]We are assuming that the actual world obeys classical logic.

[9]The view described here is linguistic ersatzism. Instead of using sentences, one could use combinations of objects and universals—that is to say, states of affairs. We do not intend to take a stand on the metaphysical status of the components of worlds. For our purposes here it will not matter if worlds are sets of sentences or sets of states of affairs.

[10]This is because it is normally thought that a necessary condition for a language to be learnable is that the meanings of sentences are determined by the meanings of their parts.

Something's being water analytically entails its being H_2O. So, a world could contain the sentence "Water is not H_2O," but it must also contain "Water is H_2O."[11] If one were to claim that a world could include the former without the latter, then one simply is not using the word "water." Consider the people on Hilary Putnam's Twin Earth. When they say that water is XYZ and not H_2O, they are using a homophone of "water," not the word "water." "[I]n the sense in which it is used on Earth, the sense of $water_E$, what the Twin Earthians call 'water' simply isn't water" (Putnam 1974, 285). A world containing the sentence "Water is not H_2O" but not its negation would be like Twin Earth—the sense of "water" is different. The worlds, both possible and impossible, that we are describing in this essay, however, are composed of sentences that get their meanings from the meanings of the words in the actual world, not from a Twin Earth world. So, whenever "water" appears in a sentence that is a member of a world, it means "H_2O." In every world, "Water is H_2O" is true. Worlds that allow for contradictions, however, may also include the sentence "Water is not H_2O." Such a world would contain both "Water is H_2O" and its negation.

To summarize this point, if worlds (possible or impossible) are linguistic entities, they must be closed under analytic entailment. The only viable accounts of impossible worlds treat them as linguistic entities, and so they must be closed under analytic entailment. Given analytic entailment, the world where Winston is a bachelor and not a bachelor is one where he is an unmarried man and a married man. Therefore, "Winston is a bachelor" is both true and false, and "Winston is an unmarried man" is also both true and false. The meanings of the sentences are being determined by their components, and so compositionality is saved. In his recent attack on the use of impossible worlds, Williamson writes: "[I]f we want to take impossibilities seriously, why exclude the impossibility in which A but not B is true, where those sentences are in fact synonymous?" (Williamson 2020, 344). The answer to Williamson's question is that such worlds are not meaningful descriptions, since they fail to follow through on some analytic entailment.

Even with a solution to potential compositionality problems in hand, some still might wonder whether, and why, we ought to incorporate impossible worlds into our consideration of the safety condition given that impossible worlds are, well, *impossible*. That is, one might think that the fact that impossible worlds could never be realized, or that we could never have counterparts in impossible worlds, or, relatedly, that the safety condition was initially conceived of as relying on a subjunctive conditional (which by definition considers only possible worlds) is intuitive evidence against the idea of considering impossible worlds in this way. But, these concerns miss one of the central ideas of the account of impossible worlds we've laid out—that these "worlds" are merely useful tools (on our account, tools composed of sets of sentences) with which we can potentially get a better understanding of how well our beliefs track the relevant truths involved in a variety of situations and scenarios. These tools themselves need not include realizable worlds, as on many accounts of possible worlds the relevant worlds are also not realizable (because they are, for example, physically impossible, epistemically impossible, and so forth). In addition, impossible worlds, according to our account, are not radically different from the actual world, so it's not clear that we *couldn't* have counterparts in them. It seems that what matters most is whether or not the worlds are *close*, rather than that they are realizable, and so on.

[11]For every instance of something that is water, it must also be H_2O. So, the waters that are not H_2O must also be H_2O (that is, it is both true and false that they are H_2O) for the analytical entailment to be preserved.

3.2 | Worlds with different laws

There is a strong intuition that worlds with different laws are by default very dissimilar from the actual world (and thus our beliefs in these worlds might not affect the safety of our beliefs in the actual world). This intuition isn't wholly wrong. "Similar" is context sensitive (see, e.g., Nolan 1997, 551). We make use of counterfactuals every day to consider how things might have gone if some mundane fact about ordinary objects had been different. For example, "What would have happened if I had drunk water instead of coffee this morning?" and "If Janet had been vaccinated, she wouldn't have ended up in the hospital" are examples of such counterfactuals. In the context of such "everyday" counterfactuals, differences in laws should make worlds more dissimilar. Our concern in everyday counterfactuals is what would happen with different non-modal facts but the same laws. Consider how a chess player pondering what would have happened had she made a different move will consider only sequences that follow the rules of chess she was using. Consider also the kind of counterfactual described by Kit Fine. "If Nixon had pressed the button, there would have been a nuclear holocaust" should be true. Fine claims that it is false on Lewis's original account of counterfactuals because there is a world where a small miracle stops the signal from the button to the bomb (1975, 452). Lewis alters his account to weight law differences more heavily, thereby "pushing away" worlds with small violations of physical laws (1979, 472).

Importantly, weighting differences in laws more heavily in the context of everyday counterfactuals removes one of Melchior's primary concerns with rejecting SIC. He was concerned that rejecting SIC would give us the wrong truth-values for counterfactuals, but because "similar" is context sensitive, SIC can be true in the context of everyday counterfactuals and false in the context of safety.

Though we do weight law differences more heavily in the context of everyday counterfactuals, we should not in the context of safety. With everyday counterfactuals, we do not care about how things would go if the natural laws were different. With safety, we do care, since, were the laws different, some of our beliefs might be false even when we formed them in the same way that we did in the actual world. It would be epistemically problematic if we formed a belief in some law on the basis of an experiment and would have believed the same thing via the same basis in a world where that law is false. That is to say, if a world is similar enough to the actual world that our evidence remains the same and our belief is false, that would indicate that there is a problem with our belief-forming method. A belief that is safe according to a view that does not weight law differences more heavily is better than one that is safe merely because worlds with different laws are treated as too dissimilar to consider.

Not all differences in laws will have a large impact on the world. Consider the debate about Humean supervenience in the philosophy of science. Humean supervenience claims that "the whole truth about a world like ours supervenes on the spatiotemporal distribution of local qualities" (Lewis 1994, 473). A number of counterexamples to Humean supervenience have emerged. Consider the following example from Michael Tooley:

> Suppose that there are ten different kinds of fundamental particles. So, there are fifty-five possible kinds of two-particle interactions. Suppose that fifty-four of these kinds have been studied and fifty-four laws have been discovered. The interaction of X and Y particles have not been studied because conditions are such that they never will interact. Nevertheless, it seems that it might be a law that, when X particles and Y particles interact, P occurs. Similarly it might be a law that when X and Y particles interact, Q occurs. There seems to be nothing about the local matters of particular fact in this world that fixes which of these generalizations is a law (1977, 669).

In Tooley's example we have two putative laws, and nothing about the particles themselves seems to count in favor of either of the laws. So, if the laws were to change, there would be no discernible changes in the local matters—in the non-modal facts. This would be problematic for theories of Humean supervenience. For our purposes here, Tooley's example gives us a case where differences in laws do not require differences in the non-modal local facts of the world. Two worlds with different laws need not differ with regard to other facts, contrary to the intuition that differences in laws make worlds significantly dissimilar.[12]

Now, imagine that we had entitled this essay "Mission: Impossible Worlds." Such a difference would require changes to various non-modal facts about the world, such as our personalities and the prospects for the essay's publication. Each of these non-modal differences would bring with them other changes both to the history of the world prior to this essay and to the future of the world after the essay. Differences between worlds with regard to non-modal facts about ordinary objects will often require a substantial number of other differences. We are involved in enough chaotic systems in our daily lives that even seemingly mundane differences between worlds can have significant effects. We have argued already that differences in laws do not necessitate significant other differences. So, we could have a world with different laws—different modal facts—without much difference from the actual world with regard to non-modal facts. We could also have a world with drastically different non-modal facts but the same laws as the actual world. For the purposes of safety, the world with minor differences in laws is more similar to the actual world than the world with significant non-modal differences.

The intuition that differences in laws are more significant than others can be explained by the fact that many law differences *are* very impactful. Were the gravitational constant different, the world would be a drastically different place. This is not because we are changing a law but instead because altering the gravitational constant would require significant changes to the non-modal facts of the world.

3.3 | An argument against SIC

Once we've acknowledged that law differences should not be weighted more heavily than other differences, rejecting SIC is simple. A world with different physical laws need not be too different from the actual world. To find a similar phenomenon among logically impossible worlds, we need only find logics that preserve large swaths of classical logic. Luckily logicians have given us a nice set of such logics.

Imagine a world just like the actual world except that it obeys the logic of paradox; let's call this world *p*. All of the atomic sentences in *p* get the same values that they do in the actual world.[13] According to the truth-tables (see Priest 1979, 226–27) for the logic of paradox, all negations, conjunctions, disjunctions, and so forth get the same values as they do in the actual world, assuming that these sentences do not contain modal operators.[14] This means that *p* does not actually contain any contradictions. Since it is governed by the logic of paradox, contradictions are possible but need not actually happen. Of course, there are some notable differences. For example, true contradictions are possible in *p*. Everything else is the same. The world

[12]For staunch Humeans, note that for our purposes here we only need it to be the case that there could be some difference in law without a *drastic* difference in the non-modal local facts. To get this, one could tweak Tooley's case to allow for some small local difference between a world with one law and a world with the other. Since *X* and *Y* particles never actually interact, a minor change to *X* particles between the worlds need not cause drastic changes to the other fifty-four kinds of two-particle interactions.

[13]Because the logic of paradox treats contradictory values as designated, inferences that would lead to explosion are blocked. A world where the logic of paradox is true would be impossible, but it need not be a world where everything is both true and false.

[14]In the logic of paradox, the truth-functional logical operators output classical values when the inputs are classical.

works like the actual world. This essay still gets written. Joe Biden still wins the 2020 election. Alpha Centauri is still four light-years away from Earth.

Compare *p* to the world described earlier where the title of this essay is "Mission: Impossible Worlds." Let's call this possible world *z*. Here are each world's differences from the actual world (hereafter "@"):

Differences from @	*z*	*p*
Logic	None	Allows contradictions, invalidates MP, MT, DS, and so on
Non-modal facts	Moderate differences	None
Physical laws	None	None
Modal facts	Different counterfactuals	Possible contradictions

As we have already mentioned, there will be modal differences between *p* and @. But, there will also be modal differences between *z* and @. If we are using an S5 modal logic, in which the accessibility relation between worlds is an equivalence relation, then what is possible in @ will be possible in *z*, and what is necessary in @ will be necessary in *z*. World *z* will have different nearby possible worlds, however, and so some counterfactuals in *z* will receive truth-values different from those they do in @. Since the possible worlds near *p* will be physically and metaphysically identical to the possible worlds near @, unlike *z*, these same counterfactuals will get the same value in *p* that they do in @.[15]

The big differences between @ and *z* concern non-modal facts. Differences in the title of this essay require differences in our personalities, which require additional physical differences between @ and *z*. Seemingly minor differences in the physical world can have drastic effects on the truth-values of other non-modal facts due to the chaotic nature of many physical systems.

How about the differences in the logic? World *z* has the same logic as @, but *p* does not. We have argued that they should not be weighted significantly more heavily than other differences, but what exactly are the differences here? Of course, the possibility of assigning both truth-values to a sentence is a difference. We should be wary, however, of double counting. The fact that the logic of paradox allows a sentence to be assigned both the true and the false values can be construed as the modal fact that true contradictions are possible. Other differences in the logics include the loss of *modus ponens*, disjunctive syllogism, and *modus tollens* as valid inferences. These can also be construed as modal differences. *Modus ponens* puts a constraint on what is possible. A world that obeys classical logic is one that could not have been such that $\{\varphi \rightarrow \psi, \varphi\}$ are true and ψ is false. A world that obeys the logic of paradox is one that could have been such that $\{\varphi \rightarrow \psi, \varphi\}$ both get designated values and ψ is false.

Ultimately, many of the modal differences are identical to the differences in the logic, so they should not be counted twice. World *z* diverges from @ with regard to both modal and non-modal propositions, while *p* only diverges from @ when it comes to modal propositions. More things are possible in *p*, but $\Diamond\varphi$'s being true in *p* doesn't entail that φ is true in *p*. One may think that invalidating *modus ponens* would lead to substantial differences in the non-modal facts between *p* and *z*. But, remember that not all differences in laws require substantial non-modal differences. The possibility of a counterexample to *modus ponens* does not require that one is actual. In all close worlds, every instance of $\{\varphi \rightarrow \psi, \varphi\}$ may be accompanied by ψ. There is just some distant possible world where $\{\varphi \rightarrow \psi, \varphi\}$ are true and ψ is not. Note also that a world where contradictions are possible does not have to be a world where contradictions

[15]Note that we are considering these modal differences because some beliefs we have are modal, and so we need to know about the modal features of nearby worlds.

are realized. A world where the logic of paradox is correct needs only a distant possible world where a contradiction is realized.

Since, in the context of safety, we should not weight differences in laws more heavily than other differences, the moderate non-modal differences between z and @ are weighted similarly to the modal differences between p and @. It is not clear that p is closer to @ than z is, but it is also not clear that z is closer than p is. So, p will be closer to @ than many possible worlds with differences more significant than a change to the title of this essay.

The existence of a single impossible world that is closer to @ than some possible world is enough for the weak rejection of SIC, but it is not enough to show that impossible worlds can play a role in safety. We need to show also that there are sufficiently many impossible worlds that are close enough to @ to affect the safety of our beliefs. That is to say, we need to establish the radical rejection of SIC.

To get the radical rejection of SIC, take the set of close possible worlds. For each world, u, in the set, consider an impossible world u' that is just like u except that the correct logic in u' is the logic of paradox. As we saw, a world that merely differs from @ with respect to its logic is relatively close. So, each u' is relatively close to the u from which it is generated. Since those parent worlds are close to @, the generated impossible worlds are also somewhat close. They will be closer to @ than many possible worlds with more significant differences in non-modal facts than u. Many of these close impossible worlds will be relevant in evaluating the safety of our beliefs about necessary truths.

4 | PHILOSOPHICAL BELIEFS

If there can be close impossible worlds, then we can properly evaluate the safety of our beliefs in necessary truths, even in the cases where our coming to our belief was not lucky. Returning to *Coin-basis*,* suppose that we include worlds where intuitionistic logic is correct in our evaluation of the safety of Arnold's belief. The worlds under consideration here are ones where excluded middle is not always true and where double negation is invalid. In intuitionistic worlds where G has not been proven, it is neither true nor false. Whether or not Arnold's belief is safe will turn, in part, on what he would believe in these worlds and whether they are close enough to the actual world. We hold that Arnold's belief is not safe. After all, intuitionistic logic does not change the truth-values of most propositions about the world. The fact that some mathematical propositions are neither true nor false does not lead to absurdities like $2 + 2 = 5$. With the exception of unproven mathematical propositions and various modal facts about, for example, double negation, not much else is changed. Many possible worlds with different non-modal facts about ordinary objects will be more distant than these merely intuitionistic worlds. These worlds are close, Arnold flips his coin in them, and he comes to believe that G. Unfortunately for Arnold, his belief is not true in the intuitionistic worlds, and so it is not safe.

This analysis of safety is not restricted just to mathematical beliefs. It finds its primary usefulness when analyzing philosophical beliefs. For the remainder of this essay we consider philosophical beliefs that, if true, would be necessarily true. There are two main ways that philosophical beliefs could come out safe. First, our belief-forming methods could be such that they "push away" impossible worlds where the belief is false. Second, the belief could be such that its not being true is sufficient to make the world too dissimilar from @ to play a role in safety.

The efficacy of philosophical methodologies is a topic of much discourse in metaphilosophical circles. Some of the central controversies focus on the viability of abduction and the evidentiary status of intuitions. Interestingly, the efficacy of these methodologies will have an effect on the safety of philosophical beliefs on an account that includes impossible worlds.

First, we consider intuitions and the safety of philosophical beliefs that are arrived at by intuition. After this, we consider abduction and the safety of philosophical beliefs arrived at on its basis.

Though philosophers make use of deductive and inductive arguments, the bottom-level premises of our arguments are usually supported only by abduction or intuition. Herman Cappelen (2012, 112) treats this role of being a bottom-level justification as a central feature of intuitions. Though Cappelen thinks that it is a feature of intuitions that they play this role, he also argues that philosophers don't actually make much use of intuitions. What we are going to treat as intuitions, here, are the basic premises of arguments that are not supported by deduction, induction, or abduction, whether they receive the name "intuition" or not. Our targets are the starting assumptions that come from reflections on or graspings of the meanings of concepts, whether immediate or formed through a mediate process. Unless our premises reduce to logical truths, at some point the support for them must connect to the world, and when it does, the support for the premises must come from either empirical data or our understandings of the concepts involved. Our understandings or graspings of the concepts are the intuitions we are targeting.

We do, to some extent, weigh in on the debates about the efficacy of intuitions, but only insofar as we consider whether these methods are safety conducive with regard to philosophical beliefs. We are not concerned with the general reliability of intuitions. We proceed by considering a particular philosophical belief—the belief that classical logic is the correct logic—and assume that this is a true belief. That is to say, we assume that the logic that is correct in the actual world is classical logic. For example, worlds where sentences get values other than true or false, where *modus ponens* is invalid, and where excluded middle is not a tautology are all impossible.

Without keeping fixed the facts about belief formation, there are many close worlds where classical logic is not the correct logic. Consider the u' worlds that were used to show that SIC is false. Many of those worlds are close enough that they factor into evaluations of safety. Suppose that someone believes that classical logic is the correct logic by means of intuitions. They think about excluded middle and intuit that it must always be true. In order for this person's belief to be safe, it must be the case that in the u' worlds their intuitions would have given them a different belief.

Now, the close impossible worlds where the logic of paradox is correct are spatiotemporal matches for @, and so our intuitions about the correct logic will not change from what they are in @. There are numerous accounts of intuitions, but for almost all of them it is the case that having different intuitions requires having different brain chemistry. If one's brain chemistry is exactly the same in @ and in p, then φ cannot seem true to one in @ and false to one in p. There are, however, views of intuitions that would entail that changing the logic of a world would require changing our brain chemistry. If intuitions give us direct access to truths, then a world where the correct logic is the logic of paradox would be one where our intuitions (and brain states) are different. Ethical intuitionism is such a view. Ethical intuitionism holds that ethical properties are real and that they can be known non-deductively. In the context of logic, Kurt Gödel has a view like this (see Parsons 1995, 62). He claims that mathematical intuitions are direct perceptions and need no further justification. Such views, however, are extremely controversial. We do not intend to take a stand on the nature of intuitions here, but if intuitions are direct in the way that Gödel, for example, claims they are, then we stand a better chance of getting safe beliefs. After all, changing the relevant concept should change our perception of it. As mentioned above, however, such a view is controversial, and so pinning one's hopes for safe philosophical beliefs on these kinds of intuitionism is dubious.[16]

Abduction does not fare much better. Suppose someone believes that classical logic is correct on the basis of abduction. As Williamson puts it, "Classical semantics and logic are vastly

[16]Note that these kinds of intuitionism are not intuitionistic logic.

superior to the alternatives in simplicity, power, past success, and integration with theories in other domains" (2002, 186).

If we look at the many close impossible worlds where the logic of paradox is correct, the theoretical virtues of classical logic are unchanged. It fits the evidence we have in those worlds, it is simpler than the logic of paradox, and it is more powerful. So, if, in @, abduction was the basis of our belief that classical logic is correct, then it should give us the same belief in *p*.

The problem for both intuitions and abduction is that they sometimes do not track the relevant truths. In the case of intuitions, they will only track differences in truth-value across worlds if a difference in the truth-value requires the kinds of changes to our experiences that would result in a *proper* change in intuition. As for abduction, changing which theory is correct doesn't have to impact the theoretical virtues of that theory, as the example above shows. With so many close impossible worlds, a failure of tracking can lead to unsafe beliefs.

So far we have looked only at one philosophical belief—the belief that classical logic is the correct logic. This belief, whether it is arrived at via intuition or abduction, is not safe. Other philosophical beliefs could fare better. There are two ways this can happen. First, the impossible worlds where the belief is false may all be distant from @. Second, intuitions or abduction may do a better job in some domains of philosophy than they do in others.

Let's consider the belief that phenomenal zombies are possible and assume for the sake of argument that it's true. Assuming that S5 is the correct modal logic, if the belief about phenomenal zombies is true, then it is necessarily true. This is because it is a belief that zombies are possible, and $\Diamond\varphi$ entails $\Box\Diamond\varphi$ in S5. In the situation where the belief is false, there is no world containing a phenomenal zombie. Would an actual believer in the possibility of phenomenal zombies have a safe belief? Let's consider the reasons that people actually believe zombies are possible. The basic version of the argument for the possibility of zombies is the argument from conceivability.

P1. Phenomenal zombies are conceivable.
P2. Everything that is conceivable is possible.
C. Phenomenal zombies are possible.

Many object that it is unclear that zombies are conceivable. In defense of the conceivability of zombies, David Chalmers argues like this: "I confess that the logical possibility of zombies seems equally obvious to me. A zombie is just something physically identical to me, but which has no conscious experience—all is dark inside. While this is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description. In some ways an assertion of this logical possibility comes down to a brute intuition" (1996, 96). Many nonphysicalists may find that their own arguments against physicalism ultimately rest on this kind of brute intuition that there is no logical contradiction in the existence of zombies.

Should we expect that these intuitions about the conceivability of zombies would change if zombies were impossible? We think not. Consider the non-modal facts about a world where zombies are impossible. Such a world could be physically identical to @. World @ may be a world where physicalism is false but be a world that lacks zombies. All qualia in @ and nearby worlds may be the result of physical processes. So, in the world where zombies are impossible, our brains would work in the same way as they do in @ and we would come to the same intuitions. But, in a world where zombies are impossible, our beliefs about their possibility would remain the same.

Are the worlds where zombies are impossible close to worlds where they are possible? One feature of zombies is that they function indistinguishably from non-zombies. The only required difference between a zombie-possible world and a zombie-impossible world is that in one of them the physical facts do not determine the mental facts. It may still be the case that in both worlds the physical and the mental facts are the same. Accordingly, these two worlds can be *extremely* similar to each other. Since the impossible worlds where zombies are impossible are

close and our intuitions would not change in them, a belief that zombies are possible on the basis of intuition is not safe.

There are numerous philosophical beliefs that can be evaluated using impossible worlds in an account of safety. For each of them, we need to consider how different the world would be were the belief false, and whether or not certain philosophical methodologies would work better than they do in the cases described above.

Of course, philosophers don't always use intuitions or abduction. Some areas of philosophy are more empirically based than others. For example, were someone to come to a belief that the world is not gunky on the basis of an empirically supported physical theory, then the belief may very well be safe. Were the belief false, then the observations that led to the empirically supported physical theory would change. Alternatively, were someone to come to a philosophical belief through pure deduction using only inferences that are valid or quasi-valid in reasonable logics, then that belief has a good chance of being safe.[17] Unfortunately the project of evaluating various philosophical beliefs will have to be left for a different time. Here are some interesting avenues for discussion:

- The belief that utilitarianism is true.
- The belief that properties are tropes.
- The belief that indiscernibles are identical.
- The belief that the world is gunky.

One might worry that a substantial number of philosophical beliefs will turn out to be unsafe. But, even if we cannot get safe philosophical beliefs in many domains by aiming for knowledge, we could still make philosophical progress by acquiring *safer* beliefs (by employing more empirical methodologies) or by aiming for rational belief or reflective equilibrium, as various metaphilosophers have advocated (see Beebee 2018).

5 | CONCLUSION

Accounts of safety that do not include impossible worlds have problems with beliefs in necessary truths. Adding impossible words to an account of safety allows us to give a more interesting and adequate account of the safety of beliefs in necessary truths. Because impossible worlds can be extremely similar to world @, to know whether or not a belief is safe we need to consider what beliefs one would form in these close impossible worlds. Many philosophical beliefs are such that we would still believe them in close worlds *where we are wrong*. At best, this is epistemically troubling; at worst, it supports a deep metaphilosophical skepticism.

ACKNOWLEDGMENTS

Special thanks to Reina Hayaki and Adam Thompson for reading and providing helpful comments on early drafts of this essay, as well as to the anonymous reviewers for *Metaphilosophy*, whose comments and suggestions helped strengthen the piece.

REFERENCES

Beebee, Helen. 2018. "Philosophical Scepticism and the Aims of Philosophy." *Proceedings of the Aristotelian Society* 118, no. 1: 1–24.

Bernecker, Sven. 2020. "Against Global Method Safety." *Synthese* 197, no. 12: 5101–16.

[17]A quasi-valid inference is one that would be valid were we to restrict the premises to sentences that receive classical values. See Priest 1979, 232.

Berto, Francesco, and Mark Jago. 2019. *Impossible Worlds*. Oxford: Oxford University Press.
Cappelen, Herman. 2012. *Philosophy Without Intuitions*. Oxford: Oxford University Press.
Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
Fine, Kit. 1975. "Critical Notice." *Mind* 84, no. 335: 451–58.
Hirvela, Jaako. 2019. "Global Safety: How to Deal with Necessary Truths." *Synthese* 196, no. 3: 1167–86.
Kripke, Saul A. 2011. "Nozick on Knowledge." In *Philosophical Troubles: Collected Papers, Volume 1*, 162–264. Oxford: Oxford University Press.
Lewis, David. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13, no. 4: 455–76.
Lewis, David. 1994. "Humean Supervenience Debugged." *Mind* 103, no. 412: 473–90.
Luper-Foy, Steven, editor. 1987. *The Possibility of Knowledge: Nozick and His Critics*. Lanham, Md.: Rowman and Littlefield.
Melchior, Guido. 2017. "Epistemic Luck and Logical Necessities: Armchair Luck Revisited." *Thought Experiments Between Nature and Society: A Festschrift for Nenad Miščević*, edited by Bojan Borstner and Smiljana Gartner, 137–50. Cambridge, Mass.: Cambridge Scholars.
Melchior, Guido. 2021. "Sensitivity, Safety, and Impossible Worlds." *Philosophical Studies* 178, no. 33: 713–29.
Miščević, Nenad. 2007. "Armchair Luck: Apriority, Intellection and Epistemic Luck." *Acta Analytica* 22, no. 1: 48–73.
Nolan, Daniel. 1997. "Impossible Worlds: A Modest Approach." *Notre Dame Journal of Formal Logic* 38, no. 4: 535–72.
Nozick, Robert. 1983. *Philosophical Explanations*. Cambridge, Mass.: Harvard University Press.
Parsons, Charles. 1995. "Platonism and Mathematical Intuition in Kurt Gödel's Thought." *Bulletin of Symbolic Logic* 1, no. 1: 44–74.
Priest, Graham. 1979. "The Logic of Paradox." *Journal of Philosophical Logic* 8, no. 1: 219–41.
Priest, Graham. 2014. *One: Being an Investigation into the Unity of Reality and of Its Parts, Including the Singular Object Which Is Nothingness*. Oxford: Oxford University Press.
Pritchard, Duncan. 2009. "Safety-Based Epistemology: Whither Now?" *Journal of Philosophical Research* 34: 33–45.
Pritchard, Duncan. 2020. "Anti-Risk Virtue Epistemology." In *Virtue Theoretic Epistemology*, edited by Christoph Kelp and John Greco, 203–24. Cambridge: Cambridge University Press.
Putnam, Hilary. 1974. "Meaning and Reference." *Journal of Philosophy* 70, no. 19: 699–711.
Sandgren, Alexander, and Koji Tanaka. 2020. "Two Kinds of Logical Impossibility." *Noûs* 54, no. 4: 795–806.
Sosa, Ernest. 1999. "How to Defeat Opposition to Moore." *Philosophical Perspectives* 13: 141–53.
Thomasson, Amie. 2007. *Ordinary Objects*. Oxford: Oxford University Press.
Tooley, Michael. 1977. "The Nature of Laws." *Canadian Journal of Philosophy* 7, no. 4: 667–98.
Vogel, Jonathan. 1987. "Tracking, Closure, and Inductive Knowledge." In *The Possibility of Knowledge: Nozick and His Critics*, edited by Steven Luper-Foy, 197–215. Lanham, Md.: Rowman and Littlefield.
Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.
Williamson, Timothy. 2002. *Vagueness*. London: Routledge.
Williamson, Timothy. 2009. "Reply to John Hawthorne and Maria Lasonen-Aarnio." In *Williamson on Knowledge*, edited by Patrick Greenough and Duncan Pritchard, 313–29. Oxford: Oxford University Press.
Williamson, Timothy. 2020. *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford: Oxford University Press.
Zhao, Bin. 2021. "A Dilemma for Globalized Safety." *Acta Analytica*. doi:10.1007/s12136-021-00478-w. Accessed December 30, 2021.

How to cite this article: Garrett, Zack, and Zachariah Wrublewski. 2022. "Impossible worlds and the safety of philosophical beliefs." *Metaphilosophy* 53: 344–361. https://doi.org/10.1111/meta.12550.