KRITIKE VOLUME FOURTEEN NUMBER TWO (DECEMBER 2020) 54-75

Article

Where Epistemic Safety Fails

Mark Anthony L. Dacela

Abstract: In a previous paper, I briefly profiled unsafe beliefs as either: (1) beliefs formed using a method that is conditionally reliable and (2) beliefs formed using a method with unstable reliability. I dubbed these profiles as B-type and C-type, respectively. Extending this analysis, I will demonstrate how these belief types operate and why they fail in some notable counterexamples to safety offered by Neta and Rohrbaugh, Cosmesaña, Baumann, Kelp, Bogardus, and Freitag. Examining these cases also motivate my thesis that a method's conditional reliability or instability does not render a belief formed by an actually reliable method unjustified; its epistemic worth remains intact, unsafe as it may be.

Keywords: epistemology, safety, knowledge, possible worlds

Introduction

In a previous paper, I briefly profiled unsafe beliefs as either: (1) beliefs formed using a method that is conditionally reliable and (2) beliefs formed using a method with unstable reliability.[1] I dubbed these profiles as B-type and C-type, respectively. Extending this analysis, I will demonstrate how these belief types operate and why they fail in some notable counterexamples to safety offered by Neta and Rohrbaugh,[2] Cosmesaña,[3]

[1] See Mark Anthony L. Dacela, "Are Modal Conditions Necessary for Knowledge?" in *Kritike: An Online Journal of Philosophy*, 13:1 (2019), 101–121.

[2] See Ram Neta and Guy Rohrbaugh, "Luminosity and the Safety of Knowledge," in *Pacific Philosophical Quarterly*, 85 (2004), 396–406.

[3] See Juan Cosmesaña, "Unsafe Knowledge," in *Synthese*, 146 (2005), 395–404.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

Baumann,[4] Kelp,[5] Bogardus,[6] and Freitag.[7] Examining these cases also motivate my thesis that a method's conditional reliability or instability does not render a belief formed by an actually reliable method unjustified; its epistemic worth remains intact, unsafe as it may be.

This paper is divided into the following sections: first, a quick review of the safety condition and its use of possible worlds; then, a discussion of my profiling of *unsafe beliefs*; and finally, an analysis of *unsafe* beliefs in the counterexamples cited above.

Safety in Brief

Sosa offered safety as a necessary condition for knowing.[8] He stated this condition, where "S" stands for subject and "*p*" stands for proposition, as:

> S's belief is safe = df. S would believe that *p* if it were so that *p* or alternatively S would not believe that *p* without it being the case that *p*.

We may simplify this condition using this subjunctive conditional:

> If S were to believe *p*, it would be the case that *p.*

And employing the possible-worlds account of subjunctive conditionals, we can modify this to:

> S's belief is safe = df. In the closest possible worlds in which S believes *p*, *p* is true.

Sosa's analysis of knowledge can then be expressed as:

---

[4] See Peter Baumann, "Is Knowledge Safe?" in *American Philosophical Quarterly*, 45:1 (2008), 19–30.

[5] See Christoph Kelp, "Knowledge and Safety," in *Journal of Philosophical Research*, 34 (2009), 21–31.

[6] See Tomas Bogardus, "Knowledge under Threat," in *Philosophy and Phenomenological Research*, 88:2 (2014), 289–313.

[7] See Wolfgang Freitag, "Safety, Sensitivity and 'Distant' Epistemic Luck," in *Theoria* 80:1 (2014), 44–61.

[8] See Ernest Sosa, "How to defeat Opposition to Moore," in *Philosophical Perspectives* 13 (1999), 141–153.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

> S knows that *P* = Df. (1) *p* is true, (2) S believes that *p*, (3) in the closest possible worlds in which S believes *p*, *p* is true (safety condition).

To understand how safety works, we need to review the semantics at play. From here on I will use the terms "subjunctive conditional" and "subjunctive" interchangeably. Also, note that moving forward "'pq'" represents the subjunctive: *If it were p then it would have been q*. Now a brief note on to Lewis's and Stalnaker's accounts of subjunctives: these theories were offered as ways of determining the truth condition of subjunctives. The question that these theories try to answer can then be stated as, "When do we judge statements in the form 'pq' as true?"

Consider first Stalnaker's account. (Let "@" stand for actual world, and, "*p*-world" for world where the antecedent is true):

> **STL**: 'pq' is true in @ = Df. 'pq' is true in the closest *p*-world to @.

Stalnaker asks us to consider the world closest to the actual world, which for him refers to the world which "differs minimally" from the actual world, and in which the antecedent (*p*) is true: If 'pq' is true in that world then 'pq' is true in @.[9] So, given a set of *p*-worlds we only check the *p*-world which differs minimally to @. Stalnaker also tells us that requiring a world that "differs minimally" implies that:

> [T]here are no differences between the actual world and the selected world except those that are required, implicitly or explicitly, by the antecedent . . . [and] among the alternative ways of making the required changes, one must choose one that does the least violence to the correct description and explanation of the actual world.[10]

These further conditions recognize that fact that different situations may obtain in worlds where the antecedent of a given subjunctive is true. And that there is a *degree of variance*, such that one world is more similar to the base world (i.e., the actual world) than another world. We check only the *p*-world

[9] Robert Stalnaker, "A Theory of Conditionals," in *Studies in Logical Theory*, ed. by Nicholas Rescher (Oxford: Basil Blackwell, 1968), 102.

[10] *Ibid*., 104.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

that is most similar in that it "differs minimally" from the actual world. Lewis calls this the Stalnaker assumption:[11]

> For every world @ and antecedent *p* that is accessible to @, there is a sphere around @ containing exactly one p-world.

Lewis rejects this assumption and offers this revised account:[12]

> **LEW**: 'pq' is true in @ = Df. Some world in which *p* and *q* are true is closer to @ than any world in which *p* and not-*q* is true.

Again, *closeness* refers to the similarity relation of worlds. Unlike Stalnaker, Lewis does not limit the set of *relevant worlds* to only one member. Lewis's account asks us to compare worlds in which '*p* and *q*' obtains and worlds in which '*p* and not-*q*' obtains. If at least one member of the first set is more similar to the base world (or the actual world) than any member of the second set, then the subjunctive 'pq' is true.

By adding a temporal element to the equation, Lewis further qualifies the set of relevant worlds, where "w_1" stands for a possible world:[13]

> **LEW*:** 'pq' is true at @, where *p* is entirely about affairs in a stretch of time t_1 = Df. (1) *p* is true at w_1; (2) w_1 is exactly like @ at all times before the transition period beginning shortly before t_1; (3) w_1 conforms to the actual laws of nature at all times after t_1; and (4) during t_1 and the preceding transition period, w_1 differs no more from @ than it must to permit *p*.

LEW* tells us just how similar the worlds should be. (1) limits the relevant set to *p*-worlds (same as Stalnaker), which we take as the initial similarity test. (2) to (4) set a *similarity range*: in (2) the worlds should be exactly similar from all times before the transition period which begins shortly before *p* obtains (t_1); in (3) the laws of nature should be similar at all times; and in (4) and onwards, the difference should be no more than what it is required for *p* to obtain.

---

[11] David Lewis, *Counterfactuals* (Oxford: Blackwell Publishers, 1973), 78.

[12] *Ibid*., 82.

[13] David Lewis, "Counterfactual Dependence and Time's Arrow," in *Noûs* 13:4 (1979), 462.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

Lewis also offers a priority list in weighing factors for similarity (in order of significance): (1) avoid big, widespread, diverse violations of law (large miracles); (2) maximize the spatiotemporal region throughout which a perfect match of particular facts prevails; (3) avoid even small, localized, simple violations of law (small miracles); and (4) secure approximate similarity of particular facts.[14] (1) to (4) tell us that a perfect match of small facts for an extended time counts more than the absence of small miracles in weighing overall similarity. However, the absence of large miracles outweighs these two factors.

Going back to safety, note that this condition is in the form of a counterfactual. And as discussed, to check if a counterfactual is true, we need to check possible worlds in which the antecedent is true and see if the consequent holds there as well. Safety thus requires us to check close possible worlds where the subject believes the proposition and see if in those worlds the proposition is true. Then alternatively, in the close possible worlds where the subject does not believe the proposition, the proposition is false.

Unsafe Beliefs

I offered two profiles of unsafe beliefs: B-type and C-type.[15] B-type beliefs are formed using a conditionally reliable method, while C-type beliefs are formed with unstable reliability. Developing the notion introduced by Goldman,[16] we can consider a method conditionally reliable if in case there is a possible circumstance in which it fails to produce a true belief; and unstable if at any given instance in can produce a false belief. We can say then that methods with unstable reliability are also conditionally reliable, but not all conditionally reliable methods are unstable.

To appreciate the difference, it is helpful to think of this in modal terms. A method is conditionally reliable if there are worlds in which it produces false beliefs. If these worlds are extremely close to the actual world, the method is unstable. We can also think of it in terms of probability. If the probability that the method will produce a false belief is high, the method is unstable. If there is a possibility that it will produce a false belief but the probability is low, then the method is conditionally reliable.

I also identified other features of both B-type and C-type beliefs. First, both beliefs are internally justified. This means that what justifies the belief is within the conscious grasp of the subject. In other words, the evidence is

[14] *Ibid*., 472.

[15] Dacela, "Are Modal Conditions Necessary for Knowledge?" 104.

[16] See Alvin I. Goldman, "What is Justified Belief?" in *Justification and Knowledge*, Philosophical Studies Series in Philosophy, Vol. 17, ed. by George Sotiros Pappas (Dordrecht: Springer, 1979), 1–23.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

known. Second, both beliefs are factually defeated. Following Steup, a factual defeater is a true proposition, hidden from the subject, and either weakens the justification of a belief or renders it completely unjustified.[17]

Counterexamples to Safety

Now we take a closer look at some notable counterexamples to safety to see how these beliefs operate and why they fail to meet the safety requirement. As a way of framing my analysis note that at least four sets of possible (epistemic) worlds are at play in these cases: {} worlds in which the proposition is true and the subject believes it, {} worlds in which the subject falsely believes the proposition, {} worlds in which the proposition is true but the subject does not believe it, and {} worlds in which the proposition is false and the subject does not believe it:

{} Bsp. p
{} Bsp. p
{} Bsp. p
{} -Bsp, p

In each case, the crucial step is determining if these sets are included or excluded in the set of relevant or close worlds {}. The similarity criterion states that any member of {} is similar to the actual world (@):

> For a given world, call it the actual world @, and a possible world #, # is a member of {} iff # is similar to @.

Note that a belief is safe if and only if '*if it were that the subject believes the proposition, the proposition is true*'. The safety condition then limits the set of relevant or close worlds to {} worlds in which the proposition is true and the subject believes it; and excludes {} worlds in which the subject falsely believes the proposition:

Belief is safe iff: (3) {} includes members of {} and
(4) {} excludes members of {}.

(3) and (4) are necessary conditions.

[17] Matthias, Steup, *An Introduction to Contemporary Epistemology* (Upper Saddle River, NJ: Prentice-Hall, 1996), 14.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

Gottit and Nogood (Baumann)

Baumann presents a case that exposes the safety condition's lack of clarity and straightforwardness.[18] Consider this version first:

> **MASK**. The following story is from Milleville, a small town in the Wild West. Two notorious bank robbers have been doing business in the area for some time: Frederick P. Nogood and Wilbur Gottit. Their faces are on "Wanted" posters all over the place. They are rivals and don't like each other at all. When Nogood goes to the bank, he uses a perfectly deceptive Gottit mask; when Gottit goes to the bank, he uses a perfectly deceptive Nogood mask. Nobody but they themselves know this. One day, Frank is walking around in the streets of Milleville when he suddenly sees a bank robber leave the bank with a bag full of money on his back, shooting back at the bank. Frank happens to look at him and there is no doubt for him: It is Nogood. But it really is Gottit with his Nogood mask on. However, by sheer coincidence Gottit's Nogood mask slips at that very moment, and Frank notices all this. This is extraordinary because something like that only happens this one time to Gottit and never to Nogood.

So, Frank forms the belief that *Gottit is the robber* (*p*). And clearly, Frank knows *p*. However, Baumann claims that Frank's belief does not satisfy the safety condition, since there are close worlds where he falsely believes *p*. In these worlds, Frank believes that Gottit just robbed a bank when it was really Nogood wearing his Gottit mask. The counterfactual 'S would believe that *p* only if it were so that *p*' does not hold in this case.

Now a safety theorist might question just how close the world where Nogood is wearing his Gottit mask (w_1) to the actual world where Frank notices Gottit's Nogood mask slip (@). She might say that only worlds where Frank sees Gottit's Nogood mask slip should be counted as close worlds. That is, worlds where everything is the same with the actual world except for one small epistemically irrelevant detail (e.g., Gottit has one less hair on his right leg), or something slightly different happening far elsewhere that does not have anything to do with Frank, Gottit, or Nogood. But Baumann questions just how defensible this notion of closeness would be.

[18] Baumann, "Is Knowledge Safe?" 20.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

Baumann identifies and evaluates possible determinants of closeness. He started with this general condition, where @ stands for the actual world and w_1 for the possible world in question:

> **D1** w_1 is close to @ = Df. The differences between w_1 and @ are epistemically irrelevant (enough).

D1 recognizes that some differences are *epistemically irrelevant* while others are not. Whether or not the difference is relevant depends on how much it varies the *epistemic situation* of the subject in @. That Frank has one less hair on his right leg in W_1 does not change the epistemic situation. So we consider this difference epistemically irrelevant. Thus, we can modify D1 to:

> **D1*** w_2 is close to @ = Df. The epistemic situation of S is the same in both w_1 and w_a.

Baumann still finds D1* unsatisfactory, since it does not tell us what an epistemic situation is, and, more importantly, what makes an epistemic situation the same or different. He also finds other versions problematic:[19]

> S's epistemic situation is the same in w_1 and w_a =
>
> **D2** Df. S holds the same belief in w_1 and @.
> **D3** Df. The truth value of S's belief is the same in w_1 and @.
> **D4** Df. The relevant facts are the same in both w_1 and @.
> **D5**Df. The initial conditions are the same in both w_1 and @.

Baumann claims that they are either trivially true (D2), or they make safety trivial (D3 to D5). He then closely examines three promising versions. Consider this first:

> **D6** S's epistemic situation is the same in w_1 and @ = Df. S's warrant for believing *p* is the same in w_1 and @.

Baumann notes that *warrant* here is taken in its broadest sense.[20] It includes the *reasons or justification* the subject might have for the belief and the *methods of belief acquisition*. D7 involves sameness of *subjective* evidence. It asks us to

[19] *Ibid*., 23.
[20] *Ibid*.

ISSN 1908-7330

consider the way the evidence appears to the subject. We can restate it this way:

> **D6*** S's epistemic situation is the same in w_1 and @ $=_{Df.}$
> (1) Subjectively speaking, S has the same evidence for her belief in w_1 and @.

Now consider these two worlds:

> @ Frank notices Gottit's Nogood mask slip (e_a), then forms the belief *that Gottit is the robber* (*p*).
> w_1 Frank sees Nogood wearing his Gottit mask (e_1), then forms the belief that *Gottit is the robber* (*p*).

Note that as far *seeing* goes, Frank's evidence in both worlds is the same: Gottit's face. Given D6 then, Frank has the same epistemic situation in both worlds. D6 does not restrict the set of close worlds to worlds where Gottit's mask slip. So, D6 does not work if the idea is to exclude worlds like W_1.

Now consider this definition:

> **D7** S's epistemic situation is the same in w_1 and @ $=_{Df.}$ Objectively speaking, S has the same evidence for believing *p* in w_1 and @.

Given D7, the safety theorist can claim that Frank's epistemic situation in w_1 and @ vary. In @, Frank actually sees Gottit's face. While in w_1, he is in fact seeing Nogood in his Gottit's mask. Frank may not be able to tell the difference, but objectively speaking, his evidence in W_1 is different from his evidence in @. However, Baumann finds D8 too strong and not very illuminating.[21] Typically, the subject's evidence for his belief *p* in a world where *p* is true would differ from his evidence for the same belief in a world where *p* is false. This is the case with worlds @ and w_1. In @, Frank's belief that *Gottit is the robber* (*p*) is true, and his evidence confirms this, while in w_1, Frank's belief is false but his evidence misleads him to believe otherwise. D7 thus excludes worlds in which the subject's belief is false. Baumann worries that this would trivialize the safety account.[22]

Finally, consider this definition:

[21] *Ibid*., 24.
[22] *Ibid*.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

> **D8** S's epistemic situation is the same in w_1 and @ = $_{Df}$. S's belief forming method is the same in w_1 and @.

D8 does not work with the set of worlds we have. It does not tell us how Frank's method in @ is relevantly different from his method in w_1. Baumann thinks D8 run into similar problems in the argument from sameness and differences of evidence or reasons (see discussion above). On the one hand, if you consider *seeing* or *perceiving* as Frank's belief forming method in @, then the difference in method does not seem relevant. In w_1 Frank's belief is formed via *perception* as well (only he's actually seeing Nogood's Gottit mask). On the other hand, if we construe method in the externalist sense, then they only differ in terms of the truth-value of the proposition. Everything else would be the same (Frank uses the method of looking at the person's face in both worlds) except that in the @ the belief is true, and in w_1, false. This leads to the exclusion of worlds where the subject's belief is false, which threatens to trivialize safety.

Now consider version two of Baumann's case:[23]

> **FAKE**. Many people do robberies in the Milleville area. All of them (including Nogood) wear non-slipping perfect Gottit masks, except Gottit who usually wears a Nogood mask, except today. Frank happens to see Gottit without his mask (he forgot to bring it to work today).

So, Frank forms the belief that *Gottit is the robber* (*p*). Baumann thinks it is uncontroversial to claim in this case that Frank does not know *p*. And this is consistent with safety. There are close worlds where Frank *falsely* believes *p*. The subjunctive conditions 'S would believe *p* only if it were so that *p*' does not hold in this case. Frank's belief is not safe. Notice that those worlds where another person wears a Gottit mask are considered close in this case. While in *Mask*, the world where Nogood wears a Gottit mask, arguably, is not included in the set of close worlds. Baumann wonders why this is so. *Mask* and *Fake* differ in two ways: (1) There are more robbers in *Fake* not just Gottit and Nogood and (2) Gottit wears a slipping mask in *Mask* but not in *Fake*.[24] Bauman argues that neither of these explains why the set of close worlds or the *ceteris paribus set* varies in *Mask* and *Fake*.

Baumann considers (1) negligible.[25] You can easily modify *Mask* to include many masked robbers. This would not significantly change the result. Frank still knows that *Gottit is the robber* (*p*) yet Frank's belief remains *unsafe*.

---

[23] *Ibid*., 25.
[24] *Ibid*.
[25] *Ibid*.

ISSN 1908-7330

(2) does not solve the puzzle either. It would explain why the *ceteris paribus set* varies in *Mask* and *Fake* only if it would imply a difference in either reasons or methods. But even if you grant these differences, it is not clear why such qualitative differences would have implications on the *ceteris paribus set* (this argument would parallel the ones discussed above). Without an argument to explain why these differences relevantly vary the *ceteris paribus set*, safety theorists should assume that in both *Mask* and *Fake* either: the masked worlds are included or excluded in the *ceteris paribus set*. Baumann asserts that either way the counterexample would hold.

Halloween Party (Cosmesaña)

Now consider Juan Cosmesaña's example:

> **HALLOWEEN**: There is a Halloween party at Andy's house, and I am invited. Andy's house is very difficult to find, so he hires Judy to stand at a crossroads and direct people towards the house (Judy's job is to tell people that the party is at the house down the left road). Unbeknownst to me, Andy doesn't want Michael to go to the party, so he also tells Judy that if she sees Michael she should tell him the same thing she tells everybody else (that the party is at the house down the left road), but she should immediately phone Andy so that the party can be moved to Adam's house, which is down the right road. I seriously consider disguising myself as Michael, but at the last moment I don't. When I get to the crossroads, I ask Judy where the party is, and she tells me that it is down the left road.[26]

In *Halloween*, Cosmesaña knows that *the party is down the left road* (*p*), but he would have believed this even if it weren't true. The subjunctive conditional '*S would believe that p only if it were so that p*' does not hold in this case. Cosmesaña's belief is not safe, yet he knows *p*.

But how do we motivate the intuition that Cosmesaña knows *p* in this case? One way is to point out that his basis for belief *p*, Judy's testimony (t), is at least *actually* reliable, although it *possibly* is not. A method or a belief-basis is reliable if it is knowledge conducive. Basis *t* would have been unreliable in a possible world where he is disguised as Michael. Call this *possible unreliability* to distinguish it from *actual* reliability. In this possible

[26] Cosmesaña, "Unsafe Knowledge," 397.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

world, he would have falsely believed *p* on the same basis, *t*. This makes his belief unsafe. So, if Cosmesaña considered disguising himself as Michael, *t* would have been questionable. But this is not the case. So, *t* is *actually* reliable. This makes Cosmesaña's belief justified. And that warrants the intuition that he *knows p*.

Cosmesaña's example demonstrates that knowledge tolerates this sort of weak reliability where a belief-basis would have easily been unreliable though it actually is not. In this case, safety requires that my basis for believing *p*, *t*, produce a true belief in all the close possible worlds (where I use *t* in forming belief *p*). But Cosmesaña claims that this is not necessary for knowledge.[27]

Russell's Clock (Kelp)

Now here is a case that involves a *counterfactual intervener*, offered by Christoph Kelp:

> **CLOCK.** Suppose Russell's arch-nemesis has an interest that Russell forms a belief (no matter whether true or not) that it's 8:22 by looking at the grandfather clock when he comes down the stairs. Russell's arch-nemesis is prepared to do whatever it may take in order to ensure that Russell acquires a belief that it's 8:22 by looking at the grandfather clock when he comes down the stairs. (Since we are concerned with a conceptual claim here, Russell's arch-nemesis may have means available to do so that we can imagine only in our wildest dreams. For instance, he may be an evil-demon who can set the clock to 8:22 with his invisible hand a second before Russell looks at it.) However, Russell's arch-nemesis is also lazy. He will act only if Russell does not come down the stairs at 8:22 of his own accord. Suppose, as it so happens, Russell does come down the stairs at 8:22. Russell's arch-nemesis remains inactive. Russell forms a belief that it's 8:22 (p). It is 8:22. The grandfather clock is working reliably as always.[28]

Kelp claims that Russell knows it is 8:22 (*p*) since,

[27] *Ibid*., 402.
[28] Kelp, "Knowledge and Safety," 27–28.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

> [H]e looks at a perfectly working clock, he has the ability to read the clock, exercises his ability and hits upon the truth through the exercise of this ability. Moreover, his belief is true. It is in fact 8:22.

But he also points out that Russell's belief is not safe since at all the nearby worlds he would have falsely believed *p*.[29] These are worlds where Russell comes down a minute earlier or later prompting his arch-nemesis to intervene and change the clock's setting to 8:22. In these possible worlds, Russell would have still believed that it's 8:22 (p), even if it weren't. The subjunctive "S would have believed *p* only if it were so that *p*" does not hold in this case. And yet, Russell knows *p*.

There are striking similarities between *Halloween* and *Clock*. In both examples, the subject's basis for forming the belief in question has a weak sort of reliability similar to Cosmesaña's belief-basis in *Halloween*. That it is actually reliable motivates the intuition that the subject knows. That it would have been unreliable makes the belief unsafe. Russell's grandfather clock is *actually* reliable since his arch-nemesis did not intervene in the actual world, though he would have in close possible worlds where Russell comes down earlier or later. This latter bit makes the grandfather clock possibly or potentially unreliable.

Kelp however claims that his counterexample is more plausible than Cosmesaña's; in fact, the latter's argument strikes him as unconvincing.[30] He points out that in *Halloween*, a lot of things had to be different for the subject to have a false belief: Cosmesaña has to decide to disguised himself as Michael, he must have successfully done so, Judy must be convinced that he is Michael, she must have phoned Andy, Andy must have moved the party elsewhere. While in *Clock*, Kelp argues that all it takes is that Russell stays in bed a minute longer or he comes down a minute earlier.[31]

But is Kelp correct in claiming this? Consider these three worlds:

> w_1 Russell comes down at 8:21, looks at the clock then forms the belief that it's 8:22 (*p*).
> @ Russell comes down at 8:22, looks at the clock then forms the belief that it's 8:22 (*p*).
> w_2 Russell comes down at 8:23, looks at the clock then forms the belief that it's 8:22 (*p*).

[29] *Ibid*., 27.
[30] *Ibid*., 25.
[31] *Ibid*., 28.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

Kelp is claiming that the only difference between @ and w_2 is that, in w_2, Russell stays in bed a minute longer. But why would Russell stay in bed a minute longer? Definitely something else would have to change. It could either be something internal, viz., Russell is not as eager or motivated to wake up and start his day, or something external, viz., his alarm clock was set at 8:23. Any of this would imply some other changes. For instance, how would you explain Russell's unwillingness to get off his bed? Perhaps he has a meeting with someone he does not really like. Or there is some chore he has to do that day. This holds true with the other possibility; a lot of things need to vary to explain why Russell set his alarm clock at 8:23 instead of 8:22.

The same can be said about @ and w_1. Kelp thinks that the only difference between these two worlds is that, in the latter, Russell would have come down a minute earlier.[32] But this would certainly imply other things as well. Perhaps this time Russell is motivated to start his day, or he set his alarm clock at 8:21. And both would imply other changes too. Moreover, in w_1 and w_2, Russell's arch-nemesis decided to intervene, and have done so successfully. And, both worlds, the clock's hand is pointing at a different number. The point here is that what varies in worlds is never just one small detail. However, without a clear way to determine which worlds are close, Kelp's counterexample would still hold against safety.

Water and Flashes (Neta and Rohrbaugh)

We now turn to two cases, call them *Water* and *Flashes*, presented by Ram Neta and Guy Rohrbaugh:[33]

> **WATER**. I am drinking a glass of water which I have just poured from the bottle. Standing next to me is a happy person who has just won the lottery. Had this person lost the lottery, she would have maliciously polluted my water with a tasteless, odorless, colorless toxin. But since she won the lottery, she does no such thing. Nonetheless, she *almost* lost the lottery. Now, I drink the pure, unadulterated water and judge, truly and knowingly, that *I am drinking pure, unadulterated water* (p). But the toxin would not have flavored the water, and so had the toxin gone in, I would still have believed falsely that I was drinking pure, unadulterated water. The actual case and the envisaged possible case are extremely similar in

[32] *Ibid*.

[33] Neta and Rohrbaugh, "Luminosity and the Safety of Knowledge," 399–400.

ISSN 1908-7330

> all past and present phenomenological and physical respects, as well as nomologically indistinguishable. (Furthermore, we can stipulate that, in each case, I am killed by a sniper a few moments after drinking the water, and so the cases do not differ in future respects.) Despite the falsity of my belief in the nearby possibility, it seems that, in the actual case, I know that I am drinking pure, unadulterated water.
>
> **FLASHES.** I am participating in a psychological experiment, in which I am to report the number of flashes I recall being shown. Before being shown the stimuli, I consume a glass of liquid at the request of the experimenter. Unbeknownst to either of us, I have been randomly assigned to the control group, and the glass contains ordinary orange juice. Other experimental groups receive juice mixed with one of a variety of chemicals which hinder the functioning of memory without a detectable phenomenological difference. I am shown seven flashes and judge, truly and knowingly, that *I have been shown seven flashes* (p). Had I been a member of one of the experimental groups to which I was almost assigned, I would have been shown only six flashes but still believed that I had been shown seven flashes due to the effects of the drug. It seems that in the actual case I know that the number of flashes is seven despite the envisaged possibility of my being wrong. And yet these possibilities are as similar in other respects as they would have to be for the experiment to be well designed and properly executed.

In both cases, I know *p*, yet my knowledge is not safe: there is a nearby world in which I *falsely* believe *p*.

In *Water*, in the close possible world where the person next to me lost the lottery, she would have spiked my drink with a phenomenologically and physically undetectable toxin. And this would have falsified *p*, yet I would have still believed it. Similarly, in *Flashes*, in the close possible world where I am assigned to one of the experimental groups, I would have shown only six flashes. This would have falsified *p*, yet due to the effect of the drug given to me, I would have still believed it.

Neta and Rohrbaugh claims that in both cases the possible worlds in which the subject *falsely* believes *p* are initially similar in just about every

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

aspect except for the truth of *p* to the *actual* world in which he knows *p*.[34] We should note, however, that other changes are at play here. In *Water*, it seems clear that these things would also vary between the actual world and the nearby world: the lottery result, the subject's mood, and as a consequence, the action of the person next to me, and the quality of the water I drank (this falsifies my belief *p*). Notice how these changes are linked. In the close possible world where I falsely believed *p*, call these W_{2a}, the person next to me does not have the winning ticket, and that makes her unhappy (perhaps bitter is more accurate) so much so that she spikes my drink with toxin. Her actions, obviously, compromises the quality of my drink. In *Flashes*, what vary are the grouping assignment, the quality of my drink, the reliability of my memory, and the number of flashes (this falsifies my belief that *p*). These are not isolated changes either. In the close possible world where I falsely believed *p*, call these W_{2b}, I was assigned to a group which members are asked to drink a spiked orange juice, which then compromises the reliability of my memory, making me believe that I saw seven flashes, forgetting that I only saw six.

But are these changes enough to disqualify w_{2a} and w_{2b} as nearby worlds to the actual ones, in both cases, where I know *p*? Let's check for nearby possible worlds more similar to the ones considered as actual in both cases. In *Water*, a more similar world to the actual than w_{2a} is one in which the lottery result is the same. Similar to the actual world, the person next to me wins the lottery. And this leads to a series of events that make my belief *p* true. Any other changes would be inconsistent with the realities we have established in describing the actual world, and this too would warrant other changes that would make this possible world significantly different from the actual one. The same goes in *Flashes*. A closer world to the actual one than W_{2b} is a world where I was assigned to the same group. This also triggers a series of events that eventually make my belief *p* true.

Notice then that in these examples, the actual events are closely linked to each other, viz., a slight change in the initial conditions would vary the truth-value of the proposition. The lottery result and the assigning of groups are both crucial, since these events determine what happens next, and, eventually, whether my belief is true or false. If the person next to me wins, I would have truly believed *p*. Otherwise I would have been mistaken. If I were assigned to the non-experimental group, I would have truly believed *p*. If I were assigned to the other group, *p* would have been false. Also, note that these are the conditions set in the actual world described in both cases. In *Water*, it was stated that the person next to me did not poison my drink because she won the lottery. Had this not been the case, I would have falsely

[34] *Ibid*., 399.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

believed that *I am drinking pure, unadulterated water* (*p*). While in *Flashes*, it was stated that, if I were assigned to the experimental group, I would have falsely believed that *I saw seven flashes* (*p*).

These conditions will help us identify some non-relevant worlds. For instance, in *Water*, we've established that *she would pollute my drink only if she did not win the lottery*. So we consider non-relevant the possible world in which she pollutes my drink after winning the lottery. The condition we've set in describing the actual world gives us reason to think that this could not have easily been the case. So a world in which this obtains is not a relevant world. In the same way, a world in which she does not pollute my drink after losing the lottery is also non-relevant.

What are these relevant worlds then, so far, we've identified the following:

> *Water*
> w_1 The person next to me happily wins the lottery and leaves my drink toxin free.
> w_2 The person next to me loses the lottery then spikes my drink with toxin.
>
> *Flashes*
> w_3 I was assigned to the non-experimental group, made to drink a pure orange juice, and was shown seven flashes.
> w_4 I was assigned to the experimental group, was drugged and shown six flashes.

In w_1 and w_3, *p* is true. While in w_2 and w_4, *p* is false. The possible worlds considered as relevant in both cases are w_2 and w_4. But are these worlds really closer to the actual one than w_1 and w_3? Obviously, the answer is no. Worlds w_1 and w_3 are more similar to the actual worlds described in both cases. Perhaps a little too similar, in fact my belief *p* is true in both worlds, like in the actual worlds. So, if we limit the set of close worlds to these worlds, my belief will be safe in both cases. However, if we limit the set of close worlds to worlds similar with respect to the truth of *p*, safety will be a trivial condition. On the other hand, if we include worlds 2 and 4 the counterexamples will hold.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

Atomic Clock (Bogardus)

Tomas Bogardus offers this counterexample:

> **ATOMIC CLOCK**. The world's most accurate clock hangs in Smith's office at a cereal factory, and Smith knows this. The clock's accuracy is due to a clever radiation sensor, which keeps time by detecting the transition between two energy levels in cesium-133 atoms. This radiation sensor is very sensitive, however, and could easily malfunction if a radioactive isotope were to decay in the vicinity (a very unlikely event, given that Smith works in a cereal factory). This morning, against the odds, someone did in fact leave a small amount of a radioactive isotope near the world's most accurate clock in Smith's office. This alien isotope has a relatively short half-life, but—quite improbably—it has not yet decayed at all. It is 8:20 am. The alien isotope will decay at any moment, but it is indeterminate when exactly it will decay. Whenever it does, it will disrupt the clock's sensor, and freeze the clock on the reading ''8:22.'' (Don't ask why; it's complicated.) Therefore, though it is currently functioning properly, the clock's sensor is not safe. The clock is in danger of stopping at any moment, even while it currently continues to be the world's most accurate clock. Smith is quite punctual, and virtually always arrives in her office on workdays between 8:20 and 8:25 am, though no particular time in that duration is more likely than any other to see her arrive. Upon entering her office, Smith always looks up at her clock and notes the time of her arrival. Today, in the actual world, that alien isotope has not yet decayed, and so the clock is running normally at 8:22 am when Smith enters her office. Smith takes a good hard look at the world's most accurate clock—what she knows is an extremely well-designed clock that has never been tampered with—and forms the true belief that it is 8:22 am (p).[35]

[35] Bogardus, "Knowledge under Threat," 12–13.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

In *Atomic Clock*, Smith's belief *p* has several epistemic virtues. First, it is supported by evidence. Smith reasonably forms that belief after looking at a clock that is known to be the world's most accurate. Second, there is no *defeating evidence*. In fact, there is no defeater of any sort. Smith's belief is justified, true, and undefeated. Third, it's not grounded on any false belief. And lastly, it is formed via a reliable process. At 8:22 am in the actual world, the clock is still very accurate. These, among other things, warrant the intuition that Smith knows *p*.

But is Smith's belief safe? Bogardus says that it is not.[36] Remember that if the *alien isotope* decayed before or around the time Smith formed her belief, the clock would have malfunctioned, and her belief would have been false. And at the time Smith formed the belief the isotope is very likely to decay. So, there is a *close world* where the isotope decayed, the clock malfunctions and erroneously reads 8:22 am. In this possible world, Smith falsely believes that it's 8:22 (*p*). Smith would have easily believed *p* even if it were false. So, Smith knows, but her belief is not safe.

Bogardus claims that *Atomic Clock* succeeds where other counterexamples failed, particularly those offered by Cosmesaña, Neta and Rohrbaugh, and Kelp (see my discussion of these cases above).[37] The difference is, in those examples, the subjects are no longer at epistemic risk when they formed their beliefs, while in *Atomic Clock*, the subject remains to be epistemically threatened at the time that she formed her belief.

Recall that in *Halloween* (Cosmesaña), *Water and Flashes* (Neta and Rohrbaugh), and *Clock* (Kelp), the subject *nearly* got into a situation where they would have falsely believed the proposition in question. But they actually avoided these situations. This happened in *Halloween* when Cosmesaña decided not to disguise himself as Michael; in *Water* when the person standing next to me won the lottery; in *Flashes*, when I was assigned to the control group; and in *Clock* when Russell came down the stairs at 8:22. In other words, in these examples, when the subject actually formed the belief in question, she was no longer in a situation where she could have falsely believed it. Bogardus's main contention is that: "One can be safe at t even if something nearly happened before t that would have put one in danger at t."[38] He argues then that the beliefs in these examples are *safe.*

In contrast, the subject in *Atomic Clock* is in an actual situation where she could have easily formed a false belief. The threat is real and live, so to speak. At any time, the isotope could decay. It could have decayed before Smith came in, right before she looked at the clock, and even while she was looking at it. The clock could have easily malfunctioned. She could have

[36] *Ibid*., 12.
[37] *Ibid*., 16.
[38] *Ibid*.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

easily falsely believed *p*. And she would have in a set of close possible worlds. And this makes her belief unsafe.

3/6 Clock (Freitag)

Lastly, we turn to a case presented by Wolfgang Freitag:

> **3/6 CLOCK**. The clock malfunctions . . . and shows either 3:00 or 6:00. It shows 3:00 at 3:00 and at all times between 6:00 and 11:58 (a.m. and p.m.). At all other times, it shows 6:00. Jim, not aware of the clock's malfunctioning, looks at the clock at 3:00, thereby picking up the true belief that it is 3:00 (*p*).[39]

In *3/6 Clock*, we have another malfunctioning clock; it shows the right time only twice a day, at 3 in the morning and in the afternoon. The chance for Jim's belief to be true is only 1/360 given that the clock shows 3:00 only 12 hours a day. Intuitively then, Jim does not know *p*. But Freitag claims that Jim's belief is safe. Let me demonstrate his argument.

Note a few things first about *3/6* Clock. First, what we have here is not your usual stopped clock. It shows 3:00 at 3:00 (a.m. and p.m.) and between 6:00 to 11:58 (a.m. and p.m.). From 3:01 to 5:59 (a.m. and p.m.) it shows 6:00, then again from 11:59 to 2:59 (a.m. and p.m.). Let's represent this on a table for easy reference:

TIME (A.M. and P.M.)	**WHAT THE CLOCK SHOWS**
11:59 – 2:59	6:00
3:00	3:00
3:01 – 5:59	6:00
6:00 – 11:58	3:00

So Jim, luckily, looks at the clock at 3:00 (whether it is a.m. or p.m. is not important since in either case the clock will show the correct time) and forms the true belief that *it is 3:00* (*p*). Notice that Jim would have falsely believed *p* if he had looked at the clock at any time between 6:00 and 11:58. But, he would not have formed the same belief (and so not believe the same belief falsely) if he had looked at the clock a minute earlier or a minute later at 2:59 or at 3:01. In fact, the only other time he would have formed the same belief is between 6:00 and 11:58. If he looked at the clock at any other time, he would have formed an equally false but different belief, *it is 6:00* (*q*).

[39] Freitag, "Safety, Sensitivity and 'Distant' Epistemic Luck," 11.

ISSN 1908-7330

Now, from the qualifications given, we can identify at least four sets of possible worlds:

w_1 3:00 worlds
w_2 3:01–5:59 worlds
w_3 11:59–2:59 worlds
w_4 6:00–11:58 worlds

Recall again that Jim would have formed belief *p*, by looking at the clock, in w_1 and w_4. Belief *p* is true in w_1, and false in w_4 (call this the failure worlds). Jim would not have formed belief *p*, by looking at the clock, in w_2 and w_3. Instead, he would have falsely believed *q* in these worlds. If we order these worlds in terms of similarity, *ceteris paribus,* the worlds close to the actual world clearly belongs to w_2 and w_3 (3:01 and 2:59 worlds). These worlds are so much closer to the actual world than any of the failure worlds, w_4 (6:00 and 11:58). Thus, Freitag claims that in the nearby worlds (w_2 and w_3), Jim does not falsely believe *p*.[40] The subjunctive *"S would believe that p only if it were so that p"* obtains. Jim's belief is safe. The failure worlds (worlds 4) are not close worlds.

Freitag shows that safety cannot account for cases that involve what he calls *distant* (non-close) *failure worlds*, provided that, *all things being equal*, in the nearest possible worlds, the subject would not have formed the same belief she forms in the actual world, using the same method she used in forming her belief in the actual world. Proponents of the modal theories can provide an analysis of *closeness* that would include the *distant* failure worlds in the set of close worlds. However, Freitag, similar to others (see discussion above), notes that it is difficult to provide a "consistent and convincing set of criteria" for closeness ranking.[41] And instead of tinkering with the given semantics or the intuitive similarity ordering, Freitag proposes that we fix safety by "searching for a different way of selecting relevant possible worlds."[42]

Conclusion

The counterexamples cited above further demonstrate why B-type and C-type beliefs fail the safety test. The dilemma as I already noted is that the conditional reliability or the instability of a method does not take away the epistemic worth of justified and true belief formed via an actually reliable

[40] *Ibid*.
[41] *Ibid*., 16.
[42] *Ibid*.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330

method.[43] And since what makes the method actually reliable is a relevant epistemic detail, safety theories cannot respond to these objections by adjusting the similarity ranking without trivializing safety.

Department of Philosophy, De La Salle University, Philippines

References

Baumann, Peter, "Is Knowledge Safe?" in *American Philosophical Quarterly*, 45:1 (2008).

Bogardus, Tomas, "Knowledge under Threat," in *Philosophy and Phenomenological Research*, 88:2 (2014).

Cosmesaña, Juan, "Unsafe Knowledge," in *Synthese*, 146 (2005).

Dacela, Mark Anthony L., "Are Modal Conditions Necessary for Knowledge?" in *Kritike: An Online Journal of Philosophy* 13:1 (2019).

Freitag, Wolfgang, "Safety, Sensitivity and 'Distant' Epistemic Luck," in *Theoria*, 80:1 (2014).

Goldman, Alvin I., "What is Justified Belief?" in *Justification and Knowledge*, Philosophical Studies Series in Philosophy, Vol. 17, ed. by George Sotiros Pappas (Dordrecht: Springer, 1979), 1–23.

Kelp, Christoph, "Knowledge and Safety," in *Journal of Philosophical Research* 34 (2009).

Lewis, David, "Counterfactual Dependence and Time's Arrow," in *Noûs* 13:4 (1979).

_____, *Counterfactuals* (Oxford: Blackwell Publishers, 1973).

Neta, Ram and Guy Rohrbaugh, "Luminosity and the Safety of Knowledge," in *Pacific Philosophical Quarterly* 85 (2004).

Sosa, Ernest, "How to defeat Opposition to Moore," in *Philosophical Perspectives* 13 (1999).

Stalnaker, Robert, "A Theory of Conditionals," in *Studies in Logical Theory*, ed. by Nicholas Rescher (Oxford: Basil Blackwell, 1968), 98–112.

Steup, Matthias, *An Introduction to Contemporary Epistemology.* Upper Saddle River, NJ: Prentice-Hall, 1996).

[43] Dacela, "Are Modal Conditions Necessary for Knowledge?" 114.

https://www.kritike.org/journal/issue_27/dacela_december2020.pdf
ISSN 1908-7330