Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research

Knowledge Under Threat

TOMAS BOGARDUS
Pepperdine University

Many contemporary epistemologists hold that a subject S's true belief that p counts as knowledge only if S's belief that p is also, in some important sense, *safe*. I describe accounts of this safety condition from John Hawthorne, Duncan Pritchard, and Ernest Sosa. There have been three counterexamples to safety proposed in the recent literature, from Comesaña, Neta and Rohrbaugh, and Kelp. I explain why all three proposals fail: each moves fallaciously from the fact that *S was at epistemic risk just before forming her belief* to the conclusion that *S's belief was formed unsafely*. In light of lessons from their failure, I provide a new and successful counterexample to the safety condition on knowledge. It follows, then, that knowledge need not be safe. Safety at a time depends counterfactually on what would likely happen at that time or soon after in a way that knowledge does not. I close by considering one objection concerning higher-order safety.

Introduction

Many contemporary epistemologists hold that a subject S's true belief that p counts as knowledge only if S's belief that p is also, in some important sense, *safe*. The idea common to each member of this family of views is that S's belief B is safe just in case the method S employed to arrive at B did not put S in serious epistemic danger, that is, serious danger of thereby arriving at a false belief. Being in danger is a modal condition: it concerns what easily could have happened. And so, therefore, is safety. Here are some members of this family of views:[1]

[1] The reader may wonder why I omit Timothy Williamson here. After all, he often sounds as though he means to place a substantive safety condition on knowledge. For example, (2000, 147): "If one knows, one could not easily have been wrong in a similar case." However, elsewhere he makes it clear that his talk of "reliability," "similarity" of cases, and the like are not intended to give a non-circular necessary condition on knowledge. Perhaps surprisingly, those turn out to be *technical* terms for Williamson, and their non-conventional senses are sculpted ultimately by our grasp of knowledge. As he says (2009, 305), "[W]ith the 'knowledge-first' methodology

Duncan Pritchard (2005, 163):

> "If a believer knows that p, then in nearly all, if not all, nearby possible worlds in which the believer forms the belief that p in the same way as she does in the actual world, that belief is true."

Ernest Sosa (1999a, 142):

> "[A] belief by S is 'safe' iff: as a matter of fact, though perhaps not as a matter of strict necessity, not easily would S believe that p without it being the case that p."[2]

John Hawthorne (2004, 56 n. 17):

> "Insofar as we withhold knowledge in Gettier cases, it seems likely that 'ease of mistake' reasoning is at work, since there is a very natural sense, in such cases, in which the true believer forms a belief in a way that could very easily have delivered error."[3]

---

of *Knowledge and its Limits*, we should expect to have to use our understanding of knowledge to determine whether the similarity to a case of error is great enough in a given case to exclude knowledge." In response to proposed counterexamples (cf. 2009, 305ff), Williamson seems happy to admit that there may be cases of knowledge that are very similar to cases of error in the ordinary English sense of "similar." So, evidently the idea is that one should not lean on one's understanding of the ordinary English word "similar" when evaluating whether Williamson's similarity requirement on knowledge is met. Rather, one should do something like evaluate whether any cases of error are *similar-enough-to-exclude-knowledge*. But if we judge a case C to involve knowledge, naturally we will not judge any cases of error to be similar-enough-to-C-to-exclude-knowledge-in-C. And if we judge cases of error to be similar-enough-to-exclude-knowledge to case C, naturally we will not judge C to be a case of knowledge. Therefore, Williamson's circular approach precludes counterexamples. Since I am interested in evaluating substantive, non-circular accounts of knowledge, I omit Williamson's work in this paper.

[2] This is Sosa's preferred definition of safety, and he repeats it in later work (2002, 274): "one would not easily have that belief without it being right." But elsewhere (1999a, 146) Sosa offers a definition of safety just in terms of the subjunctive conditional: "S's belief [B(p)] is safe iff B(p) $\rightarrow$ p," which would typically be expressed as "S's belief that p is safe iff were S to believe that p, p would be true." There's some question as to whether these two statements are equivalent in meaning. Compare, for example, the following two sentences:

(A) Not easily would the sun explode without it being the case that we're in trouble.
(B) If the sun were to explode, we would be in trouble.

To my ears, (A) and (B) sound equivalent in meaning. Similarly, it may be that Sosa's two statements of the safety condition are equivalent in meaning. But, in the spirit of being safe, let's assume here that (A) and (B) are non-equivalent and treat them separately. If there is a semantic difference, perhaps it is that *not easily would it be that A without it being that C* entails only that C holds in some sufficiently large proper subset of the closest worlds in which A holds, whereas *if it were that A then it would be that C* entails that C holds in *all* the closest worlds in which A holds, in which case (B) entails (A) but not conversely.

[3] Cf. also Sainsbury (1997, 907): "If you know, you couldn't easily have been wrong." And Luper (2006): "at time t, S knows p by arriving at the belief p through some method M only if: M would, at t, indicate that p was true only if p were true."

The view that knowledge must be safe has much to be said for it. Pritchard (2005, 147–52) argues that it can capture the intuitively attractive idea that knowledge is non-lucky true belief, the central dogma of popular anti-luck epistemologies.[4] Sosa (1999a) argues that the view that knowledge must be safe gives an excellent account of inductive and anti-skeptical knowledge. And, as Hawthorne mentions in the above quotation, the view seems poised to explain why the subject in standard Gettier-style cases lacks knowledge: the subject could so easily have been wrong. And so, as I say, the safety condition on knowledge is well-motivated: it promises a generous soil sown with the seeds of much philosophical fruit.

But knowledge need not be safe. That is what I will argue in this paper, anyway. I will present a counterexample to claim that a subject S's true belief that p counts as knowledge only if S's belief that p is also, in some relevant and important sense, safe. First, though, I will briefly sketch three recent proposed counterexamples to the claim that knowledge must be safe (see Comesaña 2005, Neta and Rohrbaugh 2004, and Kelp 2009). I will then explain why these proposals are unsuccessful. They all share a common failing: each moves recklessly from the fact that a believer was in epistemic danger just before she formed her belief to the conclusion that the believer formed her belief unsafely. And, of course, that is a fallacious inference. One may be perfectly safe even if she very nearly was not. Finally, I will propose a new counterexample to safety that avoids this fallacy.

Three Proposed Counterexamples

We will now examine three proposed counterexamples to the claim that knowledge requires safety. All three share a common failing, which I will diagnose in detail below. The first proposal is from Juan Comesaña (2005, 397), who takes aim specifically at Sosa's statement of safety:[5]

> **HALLOWEEN PARTY:** 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

[4] For a sample of recent work on anti-luck epistemology, see Pritchard (2007), Riggs (2007), and Coffman (2010). But see also Ballantyne (2011) who contends that the notion of luck may not be as central to the anti-luck epistemologists' project as is widely thought.

[5] Sosa's statement of the safety condition is, recall, as follows: "not easily would S believe that p without it being the case that p."

> 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.[6]

Comesaña (ibid., 399) says, in this case, I know that the party is down the left road, though my belief is not safe (at least in the primary sense defined by Sosa 1999a and 1999b), since "it *could* easily have happened that I had the same belief on the same basis and yet the belief was false." Though this proposal was aimed specifically at Sosa's account of safety, one assumes that it would work just as well against the accounts given by Hawthorne and Pritchard. I take it that someone persuaded by Comesaña might well think that, in **HALLOWEEN PARTY**, I form my belief in a way that could very easily have delivered error (and so Hawthorne's account is threatened), and that in very many nearby possible worlds in which I form my belief in the same way, the belief is false (and so Pritchard's account is threatened).

The second proposed counterexample is from Ram Neta and Guy Rohrbaugh (2004, 399–400).[7] Call this case **LUCKY DRINK:**

> 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. 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.

---

[6] Peter Baumann (2008) provides a case that is extremely similar to **HALLOWEEN PARTY**: Frank sees Nogood in disguise. Nogood's mask improbably falls, and Frank believes on that basis that the robber is Nogood. Frank knows this, but his belief is not safe, according to Baumann, since "there are close possible worlds" in which Frank is fooled by the disguise. This case is so similar to Comesaña's that I will only mention it here. What I argue with respect to **HALLOWEEN PARTY** applies equally to Baumann's case.

[7] Neta and Rohrbaugh take aim at Williamson's view of knowledge. For reasons that I gave in the first note, I think this is misguided. Williamson means to give at most a circular account of knowledge, which guarantees that no case of knowledge can be unsafe. Neta's and Rohrbaugh's attempt to describe a case of unsafe knowledge in Williamson's terms was, therefore, doomed to fail. I consider their **LUCKY DRINK** case in this paper with respect to Sosa's, Pritchard's, and Hawthorne's non-circular accounts.

Despite the fact that the "actual case and the envisaged possible case are extremely similar in all past and present phenomenological and physical respects," and 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."[8]

Once again, one might suspect that this proposal works equally well against each of the accounts offered by Sosa, Pritchard, and Hawthorne. I take it someone persuaded by Neta and Rohrbaugh might well think that, contra Hawthorne, in **LUCKY DRINK** I know and yet I could have easily gone wrong. And, contra Pritchard, I know even though in very many nearby possible worlds in which I form my belief in the same way it's false. And finally, contra Sosa, I know p despite the fact that I easily would believe p without it being the case that p.

Consider a third and final proposed counterexample due to Christoph Kelp (2009), which he takes to refute several species of the safety condition. Kelp asks us to imagine a variation on Russell's famous stopped-clock example. Call this case **GRANDFATHER 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… . 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. It is 8:22.

Kelp finds it intuitive that, since Russell forms his belief on the basis of a perfectly working clock, his true belief counts as knowledge. And yet, Kelp says, Russell's belief here is not safe, "since some of the possible worlds at which Russell comes down a minute earlier or later are among the very close nearby possible worlds (again, notice just how easily Russell may have stayed in bed a minute longer), it is not the case that at all very close nearby possible worlds at which he forms his belief in the same way he avoids forming a false belief."

---

[8] Neta and Rohrbaugh (2004) also offer a second case, involving a subject who *nearly* takes memory-hindering drugs (but doesn't) and then forms a belief (that they take to count as knowledge) on the basis of memory. Since it's similar in all the relevant respects, I won't rehearse it in detail here. My criticism of their first case applies equally to their second case.

This ends our tour of three recent proposed counterexamples to the safety condition on knowledge. I will soon explain why each one fails. To do that, we should first get clear on the proper methodology for refuting proposed statements of the safety condition on knowledge.

Methodology

In this section, I will lay out what I take to be the proper methodology for evaluating proposed statements of the safety condition on knowledge. I will begin with Pritchard's statement. Then, I will turn to Hawthorne's and Sosa's explicitly modal safety conditions.

When Pritchard states safety in terms of "nearby" worlds, I take it that he's invoking *technical* terms from the standard Lewis-Stalnaker semantics for subjunctive conditionals, or some other semantics in the neighborhood. The words are borrowed from ordinary English, but do not be misled: in this context, the sense of "nearby" is sculpted by theoretical semantic rules given for ordinary language subjunctive conditionals.[9] Ultimately, it is those ordinary language subjunctive conditionals that give sense to the talk of "similarity" or "nearness" of worlds, and not the other way around.

Therefore, if we would like to know whether a given belief was formed safely on Pritchard's account, we should not primarily consult our ordinary language intuitions about similarity or nearness relations among worlds or cases. Rather, our first order of business should be to consult our intuitions about the truth-values of the entities for which Lewis and Stalnaker proposed semantic rules, namely ordinary language subjunctive conditionals.[10] Otherwise, we

---

[9] For example, when Lewis (2001, 21), speaking of a counterfactual conditional *if it were that A, then it would be that C*, says that "…the conditional is true at a world W iff C is true at the A-world selected from the standpoint of W… . If one A-world is selected and another A-world is not, from the standpoint of W, that establishes a sense in which we may say that the first is *closer* to W." Notice that his proposed semantic rules—not the conventions of ordinary English—establish the sense of the word "closer." He's *not* using an ordinary sense of "closer" to give sense to his semantic rules. The same should go, I take it, with Pritchard's use of "nearer" and "nearby" in this context.

[10] I say "first order of business" here because you might think that, while such intuitions carry great weight, they are not unassailable and may be revised in order to preserve some virtues of a comprehensive semantic theory: consistency, simplicity, and the like. My point is just that we, like Lewis, ought to start with the data: our ordinary language intuitions about the truth-values of subjunctive conditionals. Some of these data are more secure than others. Some we would revise in light of a powerful theory of the similarity relation. So the theoretical similarity relation is not wholly passive in this process. Rather, through reflective equilibrium the data shape the theory, and a powerful theory may prompt us to reevaluate the data and even reject some more peripheral ordinary language intuitions.

may be led astray: our ordinary language intuitions about "nearness" or "similarity" of cases or worlds can easily come apart from our intuitions about the truth values of the relevant subjunctive conditionals.[11]

Take, for example, the subjunctive conditional "If Nixon were to press the button, there would be a nuclear holocaust."[12] If we evaluate that subjunctive conditional at the relevant time ("the darkest moment of the final days," as Lewis says), it strikes us as intuitively true. However, relying on intuitions concerning "similar" or "near" as they're used in ordinary English, one might think that any world in which there is a nuclear holocaust is extremely dissimilar and remote from the actual world, and so the "nearest" button-pushing world is not one in which there is a nuclear holocaust. (Rather, some minor miracles occur and the wire from button to bombs fails, or whatever.) And so, leaning on ordinary notions of "similar" and "near," one might be tempted to judge that subjunctive conditional *false*. But that's the wrong result. Therefore—as Lewis would agree—we can't count on our intuitions about the ordinary sense of "similarity" or of "closeness" to follow the true ordering of worlds or cases, or to track with safety.

In sum, then, when we evaluate whether a belief was formed safely on Pritchard's view, we should rely primarily on our linguistic intuitions concerning the truth-values of ordinary language subjunctive conditionals, and derive conclusions about the similarity or nearness of worlds or cases—if at all—only on the basis of those intuitions. This

---

[11] Lewis (2001, 21) agrees: "Is it useful to describe [the ordering of worlds] as a *similarity* ordering, saying that the selected A-worlds are the A-worlds most similar to W? We could mean… too much by that… if we meant that our immediate 'intuitions' of similarity could be relied on to follow the ordering." And earlier (1979, 466–7), Lewis says, of testing his proposed semantic analysis of counterfactuals, "The thing to do is not to start by deciding, once and for all, what we think about similarity of worlds, so that we can afterwards use these decisions to test [my proposed analysis]… Rather, we must use what we know about the truth and falsity of counterfactuals to see if we can find some sort of similarity relation—not necessarily the first one that springs to mind—that combines with [my analysis] to yield the proper truth conditions." The lesson for us is that, in evaluating claims about the "nearness" or "similarity" of worlds or cases, pride of place should be given to intuitions concerning the truth values of ordinary language subjunctive conditionals, and not intuitions concerning ordinary English senses of "nearness" or "similarity."

[12] This example is adapted from Lewis (1979, 467), who takes it from Michael Slote and Kit Fine.

is a point about methodology—too often overlooked in the literature—and here is how it will work in practice.[13] To evaluate whether a belief was formed in a way that satisfies Pritchard's account of safety, we should ask ourselves something like, "In the situation as described, were S to believe thusly, would she believe truly?" If not, then it is false that—in Pritchard's terms—in nearly all, if not all, nearby worlds in which the believer forms the belief that p in the same way as she does in the actual world, that belief is true. On the other hand, if in the situation as described it is true that *were S to believe thusly, she would believe truly,* then she believed safely according to Pritchard. That is one test we will run on these proposed counterexamples to the safety condition on knowledge. This test will reveal whether proposed counterexamples to safety are genuine cases of *unsafe* knowledge on Pritchard's view.

Sosa and Hawthorne both express their safety conditions in ordinary English, using explicitly modal terms. To evaluate whether a belief was formed in a way that satisfies Sosa's safety condition, we should ask ourselves, "Would S not easily believe that p without it being the case that p?" If so, then the belief was formed safely on Sosa's view. If not, it wasn't. And, for Hawthorne's account, we should ask, "Did S form her belief in a way that could very easily have delivered error?" If so, then the belief was not formed safely on Hawthorne's view. If not, then the belief was formed safely. These are two further tests we will run on proposed counterexamples to the safety condition on knowledge, in order to see if they really are cases of *unsafe* knowledge on Hawthorne's and Sosa's views.

One final point. To avoid trivializing the safety condition, we should resist the siren song of this inference: "S believes that p at t, and p is true at t. Therefore, were S to believe thusly at t, she'd believe truly." Otherwise, each and every true belief is formed safely, regardless of which method the believer employs.[14] But that weakens safety to

[13] Baumann (2008, 26) agrees: "It is remarkable that safety theorists or, more generally, epistemologists who propose a modal condition for knowledge usually don't even raise the question of what determines closeness of possible worlds." Baumann lets this "indeterminacy of closeness of possible worlds" stand as an objection to safety theorists. Kelp (2009, §3) relies on his ordinary language intuitions about the similarity and dissimilarity of possible worlds, a strategy which I've just argued is ill-advised. In general, discussions of "closeness" or "similarity" of possible worlds are a tangled mess in the literature. My hope is that this section will go some way toward rectifying that situation.

[14] I won't blame you if you take this to be one more nail in the coffin of Lewis' strong-centering assumption in his counterfactual semantics, an assumption often expressed by saying that the actual world is the nearest world to itself, and so if p and q are true, then if p were true, q would be true.

insignificance, and renders it unable to explain why any true believer—even Gettier's Smith—fails to know. So, when we evaluate whether a true belief was formed safely, we should take care to focus on the *way* in which the belief was formed, and whether this method *could* easily have led to error in spite of the fact that it *actually* did not, whether this method put the believer in epistemic danger, and so on. We should swear to ignore whether, in believing that p when p is true, the subject believes truly. Of course she does—given that she believes p when it's true—even if she forms her belief in the thick of epistemic danger. Our question is what the believer's method *would* produce, not what it *does* produce.

Why the Three Recent Proposed Counterexamples Fail

Using our tests described in the previous section, let's now evaluate the three proposed counterexamples to the safety condition on knowledge. Very importantly, the subjects in these scenarios were not in epistemic danger when they formed their beliefs. They were in danger just before they formed their beliefs, but the danger had passed by the time they formed their beliefs. Consider, for example, the tension between these two quotations from Comesaña concerning **HALLOWEEN PARTY**:

> It *could not* easily have happened that Judy said that the party is at the house down the left road to someone that doesn't look like Michael to her without it being so that the party is at the house down the left road. (ibid., 398)

> ...it *could* easily have happened that I had the same belief on the same basis and yet the belief was false. (ibid., 399)

Well, which is it? Could I or could I not have easily ended up with a false belief from Judy's testimony? It seems to me that the answer is obvious: *before I decided not to dress up like Michael*, I was at risk of gaining a false belief in the future from Judy's testimony. But *after I decided not to dress up like Michael*, I was no longer at risk. And so, by the time I formed the belief, I had averted epistemic danger by deciding against dressing up like Michael. But then, crucially, I was not at epistemic risk when I eventually formed the belief. I was then safe, if only by the skin of my teeth. And so I believed safely.

Let's think about a clearer, non-epistemic case. A slighted lover bent on revenge has released a poisonous gas into your house, while you lay unsuspecting on your sofa. If you were to breathe in with the gas in your room, you would die. Or, as Sosa might say, you wouldn't easily breathe in while the gas is in your room without dying. Or, as Hawthorne might say, breathing in with the gas in your room could

very easily cause death. Fortunately, you have a very effective gas mask. Unfortunately, you have misplaced it. And so you are in serious danger.

Alerted to the gas, you frantically try to find the mask and almost don't. But, just as the gas begins to slide under your door, you find your mask and put it on. You have narrowly avoided death. You were in grave danger, but you are not anymore. The poisonous gas fills your room, while your gas mask is on. What if you were to breathe in now, while the gas is in your room? Would you die? No. You would be just fine. You are safe. Your method of breathing no longer puts you in danger. As Sosa might say, you wouldn't easily breathe in without breathing in wholesome air. As Hawthorne might say, breathing in now could not easily cause death.

In this respect at least, breathing poisonous air is like believing false testimony. In **HALLOWEEN PARTY**, Judy has been instructed to give misleading directions to Michael. With these instructions, a poisonous epistemic gas has, as it were, been released into the environment. If you look like Michael when you breathe in Judy's testimony, you will end up stricken with a false belief. Fortunately, you do not look like Michael. Unfortunately, you are toying with the idea of dressing up like Michael and asking Judy for directions. And so you are in epistemic danger. You very nearly dress up like Michael. At the last moment, however, you decide not to. You have narrowly avoided believing falsely. You were in danger of gaining a false belief from Judy, but you are not anymore. You approach Judy, bearing no resemblance to Michael.

What if you were to ask her for directions now? Would you end up with a false belief? No. You would be just fine. For people who don't look like Michael—this now includes you—Judy's testimony is a smoothly paved path to the truth. You are safe, and no longer at serious risk of error. And so Hawthorne is off Comesaña's hook. You wouldn't easily believe what she tells you without believing truly thereby. And so Sosa is off Comesaña's hook. When you form your belief, it is true that *were you to believe what Judy tells you, you would believe truly*. Therefore, it is true that in nearly all, if not all, nearby worlds in which you form the belief that p in the same way as you do in the actual world, that belief is true. And so even Pritchard is off Comesaña's hook.

I conclude that, in **HALLOWEEN PARTY,** Comesaña fails to provide an example of unsafe knowledge. Though in this case I was in epistemic danger just before forming my belief—when I was seriously entertaining the idea of dressing up like Michael, while Judy intends to lie to Michael—but I was not in epistemic danger at the moment when

I, looking very unlike Michael, approached Judy. Comesaña has moved rashly from the fact that *S was at epistemic risk just before forming her belief* to the conclusion that *S's belief was formed unsafely*.

And the same considerations apply, mutatis mutandis, to both **LUCKY DRINK** and also to **GRANDFATHER CLOCK.** In **LUCKY DRINK**, I was at risk—epistemic and bodily—before the person next to me won the lottery. But once she won, I was no longer in danger, epistemic or otherwise. The drink before me is no longer threatened. Were I to drink the liquid before me, I would drink something wholesome. Similarly, were I to believe that the liquid is what it seems to be, I would believe truly. I am safe. And likewise with **GRANDFATHER CLOCK**: Russell was in a dark cloud of epistemic danger while he lounged in bed, deciding when to come downstairs. At that point, he easily could have stumbled into a false belief. But, once he decided to walk down the stairs at 8:22, he was safe. He was no longer at risk of believing falsely, since his arch-nemesis had by then already decided not to tamper with the clock, and so the danger had passed. Therefore, I conclude that—like Comesaña before them—Neta, Rohrbaugh, and Kelp have all moved hastily from the fact that *S was at epistemic risk just before forming her belief* to the conclusion that *S's belief was formed unsafely*.

The lesson is this: for a successful counterexample to the safety condition on knowledge, the subject must be at epistemic risk *when she forms the relevant belief and not merely before*. Comesaña's subject in **HALLOWEEN PARTY** was at serious risk of believing falsely only *before* she formed her belief, and not *when* she did. And the same goes for the subjects in **LUCKY DRINK** and **GRANDFATHER CLOCK**. In light of this lesson, let me now provide an example of unsafe knowledge, an example that will refute the proposed safety conditions on knowledge due to Hawthorne, Pritchard, and Sosa.

A Genuine Case of Unsafe Knowledge

I will now describe a case in which the believer is at substantial epistemic risk at the very moment she forms her belief that p, and yet she nevertheless knows that p. First, recall the standard stopped-clock case, discussed by Russell.[15] One morning, Russell looks at a clock that reads "8:22." Russell thereby forms the belief that it is 8:22 am. As a matter of fact, it is 8:22 am, but the clock stopped the previous evening at

[15] Russell (2009, 91): "'Knowledge' is sometimes defined as 'true belief', but this definition is too wide. If you look at a clock which you believe to be going, but which in fact has stopped, and you happen to look at it at a moment when it is right, you will acquire a true belief as to the time of day, but you cannot correctly be said to have knowledge."

8:22 pm. Most people judge that, in this case, Russell does not know that it is 8:22 am. Pritchard (2004, 207) agrees and explains: "the stopped clock case is clearly not an instance of knowledge, because there is a wide range of nearby possible worlds where the agent forms the same belief regarding what the time is on the same basis (i.e., by looking at the clock), and where his belief is false." Or, in ordinary English, it is false of Russell, as his eyes fall on the stopped clock before him, that were he to believe what the clock says, he would believe truly.[16] The method he employed could easily have led to error, even though it actually did not.

A slight variation of this standard stopped-clock case is a counterexample to the safety condition on knowledge. In this case—call it **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.[17]

---

[16] Remember to resist the allure of this inference: S believes that p at t, and p is true at t. Therefore, were S to believe thusly at t, she'd believe truly. That inference, recall, would trivialize safety, and render it unable to explain why Russell fails to know here. When evaluating the counterfactual, we need to focus on whether the belief is formed in a safe way, which it may not have been even if the belief is in fact true.

[17] Brueckner and Oreste Fiocco (2002) ask us to consider the situation of a generally well-informed citizen N.N. who in the actual world @ has not yet heard the news from the theater where Lincoln has just been assassinated. Let "t" be one millisecond before Lincoln dies, let "t + 1" be when Lincoln dies, and let "L" stand for the proposition that Lincoln is President. They say N.N. knows L at t in @. But now consider a distinct possible world w in which Lincoln dies at t instead of t + 1. They say: "If such a world w is indeed possible, then presumably w is very close to the actual world. In w, N.N. believes L while ~L. Thus... N.N. knows L

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.

Does Smith know that it is 8:22 am? In answering the question, both theoretical and intuitive considerations seem relevant. To support—*though by no means guarantee*—the claim that Smith knows, we will first count up the virtues of her belief and see if popular accounts of knowledge certify that we have here the genuine article. Then, we will consult our intuitions.

The available evidence supports Smith's belief, and she was within her epistemic rights to form that belief. At many levels of generality, her belief is formed by a reliable process. Her true belief manifests her intellectual powers, virtues, and abilities, so accounts of knowledge like those of Sosa (1991), Greco (2010), and Turri (2012) suggest that Smith knows. Her belief that p is causally connected in an appropriate way with the fact that p, satisfying Goldman's (1967, 369) analysis of knowledge. What's more, her belief results from properly functioning cognitive faculties working in a congenial epistemic environment according to a design plan successfully aimed at truth. And so a proper

---

in the actual conditions of the example, even though his belief of L is not safe from error… ." Set aside the fact that Bruekner and Oreste Fiocco are engaged in the hopeless task of counterexampling Williamson's circular safety condition on knowledge. Also set aside that the existence of *one* very close possible world in which the subject believes falsely on the same basis would not be sufficient to refute an account of safety like Pritchard's. The important question is this: is N.N.'s belief genuinely unsafe? I leave it to the reader to employ the proper methodology described in the previous section. When I run those tests, it seems far from clear that N.N's belief is unsafe. This is at least partly due to the fact that, assuming determinism, a miracle would have been required for N.N.'s belief in @ at t to have been false. And so the relevant safety conditional is less clearly false in this case than it is in **ATOMIC CLOCK**, where no miracle is required. (Cf. Lewis 1979 for a discussion of miracles. This point convinced me not to adapt the lottery mechanism of **LUCKY DRINK** for the task at hand; better to stick with a genuinely indeterministic mechanism to avoid miracles and make the relevant safety conditionals more clearly false.) Also, in this case it is the *truth* of N.N.'s belief that is in jeopardy, and not the *method* by which she formed the belief (a method which she employed long before time t). **ATOMIC CLOCK** is, therefore, an improvement: by imperiling the victim's belief *methods* with genuine *quantum indeterminacy*, the case I present is more clearly a case of unsafe knowledge than the case from Brueckner and Oreste Fiocco.

functionalist like Plantinga (1993) should say that Smith knows. Also, it is not an accident that the clock's reading is accurate and that Smith's belief is true, so Peter Unger's (1968, 159) analysis of knowledge rules that Smith knows.

We may further specify the case so that, like a well-cut gem, her belief shines with even more epistemic virtues. We may easily specify that her belief is "fully grounded," i.e. not based on any false grounds, satisfying Clark's (1963, 47) analysis of knowledge. It could easily be that the grounds for her belief do not include any falsehood F such that, if F were removed from her grounds, her belief would no longer be justified. We may also add that there is nothing that Smith believes or that she should believe, given her evidence, which would defeat her justification for her belief that it's 8:22 am.[18] Evidently, by the time Smith's belief in **ATOMIC CLOCK** is fully polished, many theories of knowledge certify that she knows. But, theories aside, you may find it intuitively obvious, as I do, that Smith knows.[19] I conclude, then, that Smith knows in @ that it is 8:22 am.

---

[18] There may be, however, a true proposition q such that if Smith added q to whatever justified her in believing that it's 8:22 am, she would no longer be justified in believing that it's 8:22 am. In this case, q might be something like *there is a soon-to-decay isotope near this sensitive atomic clock*. Many have taken the mere existence of such a "factual" defeater to preclude knowledge (see, for example, Klein 1971 and more recently Lackey 2003 note 11). But the mere existence of a factual defeater is not enough to preclude knowledge. Here's an example of knowledge despite the existence of a factual defeater: You believe on the basis of perception that there's a computer before you (call that proposition "p"). If you came to believe that *you were injected this morning with a drug that normally causes vivid hallucinations of computers iff there aren't any around* (call that italicized proposition "d"), you would no longer be justified in believing p. Unknown to you, d is true, and so there exists a factual defeater for your belief that p. However, there is an antidote to this drug, which completely reverses its effects. You were also injected with the antidote, and so the drug never had any effect on you. Everything was in proper working order when you came to believe that p. It sure looks like you know p despite the truth of d, a factual defeater. Therefore, Smith's having a factual defeater in **ATOMIC CLOCK** does not by itself preclude knowledge. In the face of proposed counterexamples, Klein (1976, 809) adds that factual defeaters must not be *misleading* in order to defeat knowledge. But in the case I describe d is not misleading as Klein defines the term: if you came to believe d you would no longer be justified in believing that p, but *not* "only because there is some false proposition f" that d justifies for you. Rather, if you came to believe that d you would no longer be justified in believing p *at least in part* because of d itself (which is true), together perhaps with your (false) belief that nothing will prevent this drug's normal effects.

[19] Neta and Rohrbaugh would, I think, agree on this point. In defense of the claim that **LUCKY DRINK** involves knowledge, they say (2004, 401) "the threats to knowledge… remain purely counterfactual: even though things *could* have gone epistemically less well, and almost did go epistemically less well, in point of fact, the threat was avoided and the actual case remains epistemically unproblematic." So too in **ATOMIC CLOCK**.

So Smith knows. But was her belief safe? Before we employ our proper methodology, suppose for a moment that the radioactive isotope in Smith's office will also, when it decays, trigger an atomic bomb under Smith's chair. The isotope is overdue to decay. Is *Smith* safe? If she were to take a seat in that chair in the situation as described, would she live happily ever after? No. She is in grave danger. And likewise with her method of belief formation, as our proper methodology will now reveal.

Since the isotope is very likely to decay, Smith would easily believe it is 8:22 am without it being 8:22 am. And so Sosa's account of safety rules that Smith's belief is unsafe. And Smith formed her belief in a way that could easily have delivered error. Therefore Hawthorne's account of safety rules that Smith's belief was unsafe. The accuracy of this clock was hanging by a thread, and so it is also false that, were Smith to believe what the clock says, her belief would be true. @ is a tiny island lost in a sea of nearby worlds in which Smith forms the same belief regarding what the time is on the same basis, and in which her belief is false. In a heaping spoonful of nearby worlds, the isotope has decayed and frozen the clock on the reading "8:22," Smith looks at the clock slightly earlier or slightly later than 8:22 am, and she forms the corresponding belief on that basis. In all of these very many nearby worlds, Smith's belief is false, though she forms the belief in the same way as she does in @. Therefore, according to Pritchard's account of safety, Smith's belief is not safe.[20]

So, Smith knows, and yet her belief is not safe. **ATOMIC CLOCK**, then, is a counterexample to the safety-based accounts of knowledge given by Hawthorne, Pritchard, and Sosa. *Pace* Sosa, Smith knows, though she easily would have believed that it's 8:22 without it being 8:22.[21] *Pace* Pritchard, Smith knows, though in very many nearby worlds she believes falsely by the same method. *Pace* Hawthorne, the method Smith employs could very easily have delivered error. Smith knows, despite the fact that she is at serious epistemic risk at the very

---

[20] This is so even on Pritchard's (2007, 292) more recent definition of safety, since that too requires that "in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world… the belief continues to be true."

[21] **ATOMIC CLOCK** also refutes Sosa's (2002, 275–6) more recent proposal, which is roughly that a subject knows that p on the basis of an indication only if either (a) the indication tracks the truth outright, or (b) the indication tracks the truth dependently on some condition that guides the subject. Sosa says (ibid., 272) that a subject who reads an accidentally working clock fails to know because neither condition (a) nor condition (b) is satisfied. In **ATOMIC CLOCK**, since the clock in Smith's office is at serious risk of malfunctioning, it also fails to meet both conditions (a) and (b), for the same reasons Sosa gives with respect to the accidentally working clock. However, in the case I've described, it's no accident that Smith's clock runs well, and so it's far clearer that **ATOMIC CLOCK** involves genuine knowledge.

moment she forms her belief and not merely before. Therefore, knowledge need not be safe.

Diagnosis

Before closing with an objection, let's collect a lesson or two from the preceding discussion. **HALLOWEEN PARTY**, **LUCKY DRINK**, and **GRANDFATHER CLOCK** all point out that one can know p at t even if something *nearly* happened before t that would have put one in an inferior epistemic position with respect to p at t. Despite that insight, these cases don't work as counterexamples to the safety condition on knowledge because what goes for knowledge here also goes for safety: one can be safe at t even if something *nearly* happened before t that would have put one in danger at t. With respect to *past* happenings, safety and knowledge march in lockstep. This is the fatal shortcoming of those three examples.

ATOMIC CLOCK, in contrast, points out a way in which knowledge and safety can indeed part ways. Smith can know that p at t even if either *some event E has a high chance of happening at t* or *E will almost certainly happen soon after,* where E would put Smith in an inferior position with respect to p. As long as E hasn't occurred *yet*, knowledge is still possible even using the threatened faculties or the imperiled methods. That is, one may know even under epistemic threat, so long as the threat is *as of yet* unrealized. However, the same doesn't go for safety. One can't be safe under threat, even if it's a mere threat.[22] And threatened faculties or imperiled methods can't form beliefs safely, even if the threat is *as of yet* unrealized. Unrealized threats always defeat safety, but they don't always defeat knowledge.

In this way, methods of acquiring knowledge are like bridges to one's destination. A bridge may have many virtues even if it is in serious danger of collapse. If Godzilla is rampaging in the area, for example, even the world's sturdiest bridge may be unsafe: it may be false that, were one to take the bridge, one would arrive at her destination. But if Godzilla has not *yet* hit the bridge, it remains as sturdy as you like. Similarly, a

[22] Of course, "threaten" has at least two senses in English, so that even a man who does not threaten me can still threaten me. For example, a man locked in a cage in a sinking submarine at the bottom of the ocean can still say to me (over the radio) with his last breath, "I'm gonna get you for this!" He threatens me, since he issues a verbal promise of harm. But he doesn't threaten* me, since he poses no significant danger to me. I take it here that only *threaten** is relevant to the current discussion of safety and knowledge, the sense in which there actually is substantial danger and not a mere promise of it. It is this second sense of "threat" that is compatible with knowledge but not safety. One can know via threatened faculties, but of course such knowledge would not be formed safely.

truly excellent method of forming beliefs—checking the world's most accurate clock, say—may be imperiled (by a radioactive isotope, say). And an imperiled method is an unsafe method: it is false that, were one to employ that method, one would arrive at the truth. Nevertheless, so long as the danger has not *yet* struck, that method may be as sturdy with epistemic virtues as you like (other than safety, of course). And evidently one *may* know via an unsafe method, just as one *may* arrive at her destination via an unsafe bridge.

The primary lesson, then, is this: safety at a time depends counterfactually on what would likely happen at that time or soon after in a way that knowledge does not. That, ultimately, is why knowledge need not be safe. This difference in counterfactual dependence is also why a sturdy bridge need not be safe: when we evaluate the sturdiness of a bridge at a time, we look only at its actual structural soundness at that time. But when we evaluate the safety of the bridge at a time, we embark on an extended survey of modal space at that time and future times. The same goes, it seems, with safety and knowledge.

Many epistemologists believe there is one unique quality that transmutes the lead of true belief into the gold of knowledge, and they eagerly chase after it. If there is such a quality, this paper shows that it does *not* depend counterfactually on what would likely happen at that time or soon after. And so philosophers with an interest in this epistemic alchemy would do well to turn their attention away from safety and towards features with the correct counterfactual profile.

In principle, our candidates include all of the theories I mentioned that certify **ATOMIC CLOCK** as a case of knowledge. But of course it is no secret that most (and likely all) of those theories have proven unsatisfactory for a variety of reasons.[23] Safety theorists now join the rest of us who keep calm and carry on among the smoldering ruins of

[23] In their diagnosis of why knowledge need not be safe, Neta and Rohrbaugh assume that knowledge must be an important, earned cognitive achievement as opposed to mere unearned success. And such achievements may be earned unsafely, they say. But this diagnosis sinks in a mire of controversy. No doubt knowledge is *often* an important earned cognitive achievement. But many philosophers believe that some knowledge is utterly trivial. (Much of the knowledge of the past one would gain by browsing a decades-old phone book, for example.) And it seems that some knowledge might be straightforwardly *unearned*: Jones is an enthusiastic mathematician. She tells me that through years of toil she's proven that there are an infinite number of twin primes (i.e. primes of the form $< p, p + 2 >$). Indeed she has, but I could not care less. In fact I hear of her triumph only because her voice manages to drown out my television. Nevertheless I can't help but form the true belief via her testimony that there are infinitely many twin primes; I just find myself believing that. Both of us now know this, but surely only one of us has earned that knowledge. A nice summary of this controversy may be found in Lackey (2009), though Lackey's proposed case of unearned knowledge involves a subject who actively seeks out reliable testimony, and so the resulting knowledge is less plausibly *unearned*.

our fallen theories of knowledge. One more failed analysis will perhaps hearten only those who take our concept of knowledge to be wholly nonamenable to traditional philosophical analysis, either because it is a primitine concept or because it is a cluster concept. For what it's worth, the long and fruitless post-Gettier discussion strongly inclines me in that direction. Perhaps we must rest content with a list of epistemic goods bearing a family resemblance to one another, no one of which is necessary for knowledge (aside from truth and confidence), and various combinations of which are sufficient for knowledge. But wouldn't it be nice should future research show otherwise?

Objection: Safe but not Safely Safe

In his diagnosis of **HALLOWEEN PARTY**, Comesaña says that while Judy's testimony is reliable, it is not *reliably* reliable: had I decided to dress up like Michael, which I very easily could have, Judy's testimony would not have been reliable. Thus his diagnosis: knowledge requires reliability but not reliable reliability, whereas safety requires reliable reliability. So Comesaña grants that one's belief is formed *reliably* in **HALLOWEEN PARTY**, but denies that it was formed *safely* (at least in Sosa's sense of safety). However, this is a frail distinction, one which we have seen crumbles upon examination: the subject *does* believe safely in **HALLOWEEN PARTY.**

But a safety theorist might adapt this thought from Comesaña as an objection to **ATOMIC CLOCK**:[24]

> In **ATOMIC CLOCK**, Smith's belief is safe, but not safely safe. You're confusing a lack of higher-order safety for a lack of safety *simpliciter*, just as one might confuse a lack of knowledge of knowledge for a lack of knowledge. But knowledge doesn't necessarily iterate: one can know without knowing that one knows. And likewise safety doesn't necessarily iterate: one's belief can be safe without being safely safe.[25] After all, one might truthfully remark, upon crossing a bridge in the vicinity of Godzilla, 'Thank goodness! Against the odds, I crossed safely'. Here, the bridge proved safe (luckily, and contrary to what you said above), though it could easily have proven unsafe. The same goes with Smith's method of belief formation in **ATOMIC**

[24] Several readers and auditors have raised just this concern, but I'm especially indebted to Mark Sainsbury for putting the concern particularly clearly. He should, of course, be held guiltless of any unclarity in the following discussion. Baumann (2008, 27) also wonders whether his subject's belief might be safe but not safely safe. His response is brief, and centers on this claim: "It seems obvious that [the subject's] belief is not safe here, not just not safely safe."

[25] See Sainsbury (1997, 910) and Williamson (2000, 124–5) for plausible examples showing that safety—at least understood in terms of easy possibilities—does not necessarily iterate.

> **CLOCK**. And so you've merely shown that knowledge does not require that one's belief be safely safe. For all you've said, *mere safety* might be a requirement on knowledge.

So ends the *Safe but not Safely Safe* objection. There are important issues for safety theorists to explore here, and I feel the draw of the objection. But there are also countervailing considerations, which I believe ultimately outweigh the pull of the objection.

The objection rests on several claims, but I will focus on the two that are least worthy of our confidence. First, that Smith's belief in **ATOMIC CLOCK** is in fact formed in a safe way. Second, that *nevertheless* Smith's belief in Atomic Clock is not formed in a safely safe way. I will argue that Smith's belief is *not* formed safely, and that—in this case at least—lower-order safety entails higher-order safety.

Is Smith's Belief Safe?

We have discussed Hawthorne's, Pritchard's, and Sosa's accounts of safety, and we've established procedures to test whether any given belief was formed safely on each of these accounts. It turns out that, in **ATOMIC CLOCK,** Smith's belief comes out as unsafe on each of these tests. So there are two alternatives: either the *Safe but not Safely Safe* objector has a *new* account of safety in mind, or the objector is using "safety" in a pretheoretical, non-technical sense.

So let's first consider whether there might be a new account of safety. Here is one line of thought that I believe merits exploration. Many epistemologists take it as given that knowledge precludes the type of luck featured in standard Gettier cases. Pritchard calls this "veritic" luck. He analyzes veritic luck in terms of safety, and he analyzes safety as we have seen. This was a mistake, as I have shown: that species of safety is not required for knowledge.

But perhaps the solution is to take things the other way around: to give an account of safety in terms of luck, and to leave *luck* unanalyzed. In a recent paper, Brent Madison (2011, 53) endorses Pritchard's nearby-worlds analysis of veritic luck, but he also cashes out the anti-luck condition on knowledge in this way: "there is no luck that what [subjects] believe is true, given their evidence." The idea is that while a subject may gain knowledge on the basis of evidence she was lucky to acquire,[26] a subject cannot gain knowledge if she was lucky to form a true belief on the basis of her actual evidence.[27]

---

[26] A detective might stumble upon the murder weapon, for example, and nevertheless know on that basis who committed the murder.

[27] I thank an anonymous referee of this journal for encouraging me to consider this suggestion.

As I say, I think this general strategy of cashing out safety in terms of an unanalyzed notion of luck is promising and ought to be explored. However, I don't think that *this* particular proposal inspired by Madison's distinction will help the safety theorist here. Consider **ATOMIC CLOCK** again. Given the evidence that Smith's method of belief formation delivered, was Smith lucky to form her true belief? I should think so. And I believe Madison would agree: in diagnosing the standard stopped-clock case, he says (ibid.) that the subject's evidence is "what the clock says." And he thinks that, in the stopped-clock case, one is lucky to form a true belief on the basis of this evidence since one could so easily have been misled by the clock: "Had the subject glanced at the clock a minute earlier, or a minute later, and believed that it was eight o'clock based on what the clock read, the belief would have been false." But our subject Smith could easily have been misled by the clock at any point in **ATOMIC CLOCK** as well. Yet there we have knowledge. And so this Madison-inspired proposal is not necessary for knowledge.

As I mentioned above, there is an alternative to searching for a novel analysis of safety. Perhaps the safety-theorist ought to leave "safety" unanalyzed. Our philosophical penchant for enlightening analyses rarely bears fruit, after all. So perhaps we should suppress the analytic urge in this case and say that a belief counts as knowledge only if it was formed *safely*, full stop.

But here too I believe that the safety theorist gets the wrong result in **ATOMIC CLOCK**. If the test for safety is just to ask, "Was this belief formed safely?" I should think that the answer is a resounding "No" in the Atomic Clock case. Secure the relevant concept in your mind via paradigm cases: suppose your chair rests upon an atomic bomb that would be easily triggered should a nearby radioactive isotope decay, which it easily might. Is sitting in that very chair now a *safe* way to relax? No. Suppose your nuclear submarine's main reactor would easily be disabled should a nearby radioactive isotope decay, which it easily might. Is that very submarine now a *safe* method of undersea voyage? No. Now suppose that your atomic clock would easily be disabled should a nearby radioactive isotope decay, which it easily might. Is this very clock now a *safe* way to form beliefs about the time? No. And so safety, even understood pretheoretically, is not necessary for knowledge.

The *Safe but not Safely Safe* objection contends that one's belief in the Atomic Clock case is actually formed safely. But this contention is unjustified: it *may* be true—I can't rule out every possible theoretical account of safety—but at present the evidence points the other way. Smith's belief in **ATOMIC CLOCK** is not formed safely according to

Hawthorne's, Pritchard's, and Sosa's accounts. Madison's account of safety in terms of luck is not promising, and neither is an account that leaves "safety" unanalyzed. This is enough to show that the objection fails, at least given the current state of the debate. But let's also consider the objection's second pillar of support, to see if it too fails.

Is Smith's Belief not Safely Safe?

The *Safe but not Safely Safe* objection contends that in **ATOMIC CLOCK** Smith's belief is, despite being safe, not safely safe. I will argue that this is impossible: in *this* case at least, lower-order safety entails higher-order safety.

"Safely safe" is a slippery notion. To get a grip on it, let's begin on the firmer ground of lower-order safety. Suppose Smith believes that p via some method, and p is true. And suppose we want to know whether this happened safely; we want to determine whether there is a tight modal connection between Smith's belief-forming method and the truth. Using Hawthorne's "ease of mistake" account, we evaluate whether Smith forms her belief in a way that could easily lead to error. According to Pritchard, as I've understood him, we should evaluate the subjunctive conditional: *were Smith to believe thusly, she would believe truly*. Or, using Sosa's subjunctive conditional, $B(p) \rightarrow p$: not easily would Smith believe that p without it being the case that p.

Now suppose that Smith believes that p via some method, and this method is safe. Suppose we want to know whether *this* happens safely, i.e., whether there is a tight modal connection between the belief-forming method and the safety condition. I take it that we're wondering whether, as Hawthorne might say, it could easily be that Smith forms her belief in a way that could easily lead to error. That is, we're wondering whether, as Sosa might say, *not easily would Smith believe that p without her belief being formed safely*. That is, we're wondering whether, as Pritchard might say, *were Smith to believe thusly, Smith would believe safely*. And so, if $B(p) \rightarrow p$ captures lower-order safety, I should think that higher-order safety is captured like so: $B(p) \rightarrow (B(p) \rightarrow p)$. This is the best theoretical sense I can make of the claim that a belief-forming method is safely safe.

The *Safe but not Safely Safe* objection asserts that in **ATOMIC CLOCK** Smith's belief is formed in a safe way, but not in a way that is safely safe. That is, $B(p) \rightarrow p$, but not $B(p) \rightarrow (B(p) \rightarrow p)$. I will soon show that this could be true only on a very implausible assumption. First, though, consider this inference pattern featuring subjunctive conditionals: $(A \rightarrow B)$, therefore $(A \rightarrow (A \rightarrow B))$. This is logically valid for

subjunctive conditionals, but only if we assume Lewis' strong centering.[28] I argued above that in order to avoid trivializing the safety condition, the safety theorist should not accept strong centering. So I will not appeal to strong centering in the following argument, and so we cannot move automatically from mere safety to safe-safety. Nevertheless, I believe that the description of **ATOMIC CLOCK** and the assumption that B(p) → p together entail B(p) → (B(p) → p). That is, there must be higher-order safety in **ATOMIC CLOCK**, contrary to the objection.

To see this, consider what would have to be true in **ATOMIC CLOCK,** according to the objection. The objection asserts that in the scenario as described (call it "@") Smith's belief is formed safely. So B(p) → p is true in @. And recall that Smith formed a true belief via the clock in @. So B(p) and p are true in @. So, setting aside Lewis' strong centering, let "w1" name the unique B(p) world selected from the standpoint of @, or if there is no such unique world, let "w1" name one of the B(p) worlds selected from the standpoint of @. (As Lewis et al. would say, w1 is among @'s *closest* B(p) worlds. I will temporarily adopt this talk in what follows as shorthand, though the same point could be made with subjunctive conditionals.)

So, B(p) is true in w1 and—since we're assuming B(p) → p in @—p is also true in w1. Therefore, in w1, Smith has formed a true belief by checking the clock. And it's false in @ that *were Smith to believe what the clock says, there would be no soon-to-decay isotope*.[29] Crucially, then, w1 is just like @ in every relevant respect: same subject, same clock, same soon-to-decay isotope. The only difference between @ and w1 is inconsequential to Smith's epistemic situation.[30]

The objection then asserts that in @, Smith's belief is not formed in a way that is safely safe. That is, B(p) → (B(p) → p) is false in @. For this to be so, B(p) → p must be false in at least one of the nearest worlds to @.[31] Let w1 be that world. So B(p) → p is false in w1 despite

---

[28] Without strong centering, there can be a world w1 with w2 among its closest *A* worlds. If w2 is an *B* world, but shares its own minimal sphere with a (*A* & ¬*B*) world w3, then (*A* → *B*) will hold at w1, but fail at w2, leading to the failure of (*A* → (*A* → *B*)) at w1. That would be a counterexample to this inference (cf. Bonevac, Dever, and Sosa 2006; they call this inference "Expansion"). For similar reasons, the inference fails even if we assume weak centering, and even if we assume, as we may in **ATOMIC CLOCK,** that (*A & B*) is true in w1.

[29] It's not even true that *were Smith to believe thusly and truly, there would be no soon-to-decay isotope*. So there is an isotope in w1, threatening to decay.

[30] Perhaps some inconsequential quantum event light years away from Smith unfolds differently in @ than it does in w1.

[31] Either the nearest B(p) worlds to w1 are worlds in which p is false, or there are (B(p) & p) and (B(p) & ¬p) worlds that are equidistant from w1.

the fact that in @—a world alike to w1 in every relevant respect—it is *true* that $B(p) \rightarrow p$.[32] Therefore, w1 and @ are alike in every way that is relevant to Smith's epistemic situation, and yet the objection has it that in @ Smith's belief is formed safely while in w1 it is not. That seems clearly impossible. Since the objection requires a distinction where there is no difference, we should reject the objection.

Conclusion

On the foundational assumption that knowledge requires safety, many philosophers have promised great theoretical rewards: Duncan Pritchard's account of the anti-luck platitude, Williamson's Epistemicism and Anti-Luminosity, Sosa's anti-skepticism, and Hawthorne's solution to the Gettier problem. This paper delivers good news and bad news. The bad news is that, since knowledge doesn't require safety, all of these impressive projects are built on sand. The good news is that the landscape is now wide open for new solutions to these perennial problems of philosophy.[33]

References

Ballantyne, Nathan 2011. Anti-luck epistemology, pragmatic encroachment, and true belief. *Canadian Journal of Philosophy* 41(4): 485–504.

Baumann, Peter 2008. Is knowledge safe? *American Philosophical Quarterly* 45: 19–30.

Bonevac, Daniel, Josh Dever, and David Sosa 2006. The conditional fallacy. *Philosophical Review* 115: 273–316.

Brueckner, Anthony and Marcello Oreste Fiocco 2002. Williamson's anti-luminosity argument. *Philosophical Studies* 110: 285–293.

Coffman, E.J. 2010. Misleading dispositions and the value of knowledge. *Journal of Philosophical Research* 35: 241–258.

---

[32] Or, to recast the objection in Hawthorne's "ease of mistake" terms, Smith must be safe: the clock could not easily mislead her. But this is so only if the clock is *not* genuinely threatened by the isotope. And at the same time Smith must lack higher-order safety: it could easily be that the clock could easily mislead her. But what could this easy possibility consist in, if we've already assumed that the clock is not genuinely threatened by the isotope? Surely the case could easily be fleshed out so that there is no such easy possibility. For this reason, I can't see how Smith's belief could be safe but not safely safe even on Hawthorne's ease-of-mistake account of safety.

[33] Acknowledgements: I owe much to conversations on this subject with David Barnett, E.J. Coffman, Jon Matheson, Andrew Moon, Duncan Pritchard, David Sosa, and Adam Pautz. I am especially indebted to the detailed and insightful comments I received from Nathan Ballantyne, Tim Pickavance, Mark Sainsbury, and two anonymous referees from this journal.

Comesaña, Juan 2005. Unsafe knowledge. *Synthese* 146: 395–404.
Gettier, Edmund 1963. Is justified true belief knowledge? *Analysis* 23: 121–123.
Goldman, Alvin 1967. A causal theory of knowing. *Journal of Philosophy* 64: 357–372.
Greco, John 2010. *Achieving Knowledge*. Cambridge University Press.
Hawthorne, John 2004. *Knowledge and Lotteries*. Oxford University Press.
Kelp, Christoph 2009. Knowledge and safety. *Journal of Philosophical Research* 34: 21–31.
Klein, Peter 1971. A proposed definition of propositional knowledge. *The Journal of Philosophy* 68: 471–482.
—— 1976. Knowledge, causality, and defeasibility. *The Journal of Philosophy* 73: 792–812.
Lackey, Jennifer 2003. A minimal expression of non-reductionism in the epistemology of testimony. *Noûs* 37: 706–723.
—— 2009. Knowledge and credit. *Philosophical Studies* 142: 27–42.
Lewis, David 1979. Counterfactual dependence and time's arrow. *Noûs* 13: 455–476.
—— 2001. *On the Plurality of Worlds*. Wiley-Blackwell Press.
Luper, Steven 2006. Restorative rigging and the safe indication account. *Synthese* 153: 161–70.
Madison, Brent 2011. Combating Anti Anti-Luck Epistemology. *Australasian Journal of Philosophy* 89: 47–58.
Neta, Ram and Guy Rohrbaugh 2004. Luminosity and the safety of knowledge. *Pacific Philosophical Quarterly* 85: 396–406.
Plantinga, Alvin 1993. *Warrant and Proper Function*. Oxford University Press.
Pritchard, Duncan 2004. Epistemic luck. *Journal of Philosophical Research* 29: 193–222.
—— 2005. *Epistemic Luck*. Oxford University Press.
—— 2007. Anti-luck epistemology. *Synthese* 158: 277–297.
—— 2008. Virtue epistemology and epistemic luck, revisited. *Metaphilosophy* 39: 66–88.
Riggs, Wayne 2007. Why epistemologists are so down on their luck. *Synthese* 158: 329–344.
Russell, Bertrand 2009. *Human Knowledge: Its Scope and Its Limits*. Taylor and Francis.
Sainsbury, R.M. 1997. Easy possibilities. *Philosophy and Phenomenological Research* 57: 907–919.
Sosa, Ernest 1991. *Knowledge in Perspective*. Cambridge University Press.

—— 1999a. How to defeat opposition to Moore. In J. Tomberlin (Ed.), *Philosophical Perspectives 13: Epistemology*. Blackwell, 141–54.
—— 1999b. How must knowledge be modally related to what is known? *Philosophical Topics* 26 (1/2): 373–384.
—— 2000. Skepticism and contextualism. In J. Tomberlin (Ed.), *Philosophical Issues* 10: 1–18.
—— 2002. Tracking, competence, and knowledge. In P. Moser (Ed.), *The Oxford Handbook of Epistemology*. Oxford University Press, 264–286.
Turri, John 2012. Is knowledge justified true belief? *Synthese* 184(3): 247–259.
Unger, Peter 1968. An analysis of factual knowledge, *Journal of Philosophy* 65(6): 157–70.
Williamson, Timothy 2000. *Knowledge and Its Limits*. Oxford University Press.
—— 2009. Replies to critics. In P. Greenough and D. Pritchard (Eds.), *Williamson on Knowledge*. Oxford University Press.