When, at the end of the last millennium, Ernest Sosa came up with the intriguing idea of defining knowledge as safe belief, i.e. true belief that could not easily have been false,[[1]](#footnote-1) a real breakthrough in the theory of knowledge was imminent. The safety account turned out to be superior to many competing accounts of knowledge. In contrast to the traditional JTB account, it licenses the correct verdicts about standard Gettier cases. For example, when one luckily deduces a true conclusion from a false premise, as in the original Gettier cases, this could easily have resulted in a false belief, so that one therefore lacks safe belief. In contrast to knowledge understood as *certain* belief, the safety account avoids massive skeptical consequences.[[2]](#footnote-2) This is because our methods need not rule out deception in very distant demon-worlds in order to guarantee truth in nearby worlds, as is required by safety. The safety account also seems to fare better than the main alternative *modal* account of knowledge, which understands knowledge as sensitive belief, i.e. true belief that would not have been believed if what is actually believed were false. Whereas safe belief is closed under competent deduction, sensitive belief seems to violate reasonable versions of the epistemic closure principle.[[3]](#footnote-3) Safety also wins the day when it comes to Gettier-like cases in which, as in the barn-façade case, the true belief depends on being lucky within an epistemically unfavorable environment. Consider Henry’s true belief that over there is a barn when he looks at the only real barn in fake-barn country. Intuitively, Henry does not know. However, his belief is sensitive.[[4]](#footnote-4) The closest worlds in which what he actually believes is false are worlds in which Henry looks at a place where there is no building at all. In those worlds he clearly would not believe that over there is a barn. Now, consider safety. Since there are nearby worlds in which Henry falsely believes that over there is a barn, namely those worlds in which Henry looks at barn façades, Henry’s belief is not safe. So, safety does better than sensitivity in ruling out Gettier-like cases. But there are also cases in which sensitivity seems to require too much. Consider Sosa’s garbage-chute case (Sosa 1999: 145; Pritchard 2009a: 33-34). In this case, Ernie drops his garbage down a garbage chute that he knows to be well-maintained. On this basis he believes that the garbage will soon be in the basement. And, as expected, the garbage makes its way. Intuitively, Ernie then knows that the garbage will soon be in the basement. But his belief is not sensitive. If the garbage bag were to snag on something on the way down, he would nevertheless believe that the garbage will soon be in the basement. However, if worlds in which the garbage bag snags on something on its way down are not too close, Ernie’s belief might well be safe. These short reflections may help to explain why the safety account of knowledge seemed very promising when it was introduced to the debate.

A few years later, the safety account of knowledge was integrated into a more general anti-luck perspective by Duncan Pritchard.[[5]](#footnote-5) Pritchard maintained that knowledge is incompatible with a specific type of luck. Although one can know of lucky events, as e.g. one’s own lottery win, and although one can also achieve knowledge on the basis of luckily acquired evidence, such as by accidentally reading something in the newspaper or by overhearing a conversation, one does not know when one’s belief is merely luckily true. In short, Pritchard claimed that knowledge rules out veritic luck.[[6]](#footnote-6) One can break down Pritchard’s early view into three distinct parts. (P1) Luck in general must be analyzed in modal terms along the following lines: An event e is lucky if in close worlds e does not obtain. In this sense, one’s winning and one’s losing the lottery are, at least under fair conditions, both lucky events.[[7]](#footnote-7) In the same sense, a belief is *not* luckily true only if there are no close worlds in which it is falsely believed. Consequently, the absence of veritic luck is the same as safety. (P2) Knowledge requires (as a necessary condition) the absence of veritic luck. (P3) The absence of veritic luck is sufficient for knowledge. (P2) expresses moderate anti-luck epistemology. Whoever accepts (P3) in addition is a proponent of an robust anti-luck epistemology.[[8]](#footnote-8)

The decline of safety began only a few years after its apparent rise. Philosophers came up with counterexamples to all three of Pritchard’s above claims. Some argued that although knowledge is incompatible with veritic luck, safety does not guarantee the absence of this kind of luck. According to them, there are cases of safe but luckily true beliefs that do not count as knowledge. Others came up with counterexamples to safety as a necessary condition for knowledge. According to them, some cases of knowledge involve luckily true beliefs. Finally, virtue epistemologists dispute robust anti-luck epistemology for the reason that the modal correlation between beliefs and facts is not sufficient for knowledge. The epistemic agent must also be creditable for the epistemic achievement of acquiring true beliefs in some way or other. Ironically, Sosa and Pritchard, who were among the main proponents of safety, now entertain a more distanced or even skeptical attitude to the safety account of knowledge. Whereas Pritchard still believes in safety as a necessary condition for knowledge,[[9]](#footnote-9) Sosa now disputes that safety has any relevance for knowledge at all.[[10]](#footnote-10)

In this paper I will offer a comprehensive defense of the safety account of knowledge against counterexamples that have been delivered along all three dimensions outlined above. In section 1, I will discuss different versions of safety, arguing that a specific variant of method-relativized safety is the most plausible. I will then use this specific version of safety to respond to counterexamples in the recent literature. In section 2, I will address alleged examples of safe beliefs that still constitute Gettier cases. In section 3, I will discuss alleged examples of unsafe (and in this sense lucky) knowledge. In section 4, I will address alleged cases of safe belief that do not constitute knowledge for non-Gettier reasons. My overall goal is to show that there are no successful counterexamples to robust anti-luck epistemology and to highlight some major presuppositions of my reply. I will conclude with some general implications of my argument.

*1. Varieties of safety: A shopper’s guide*

Safety requires a modal correlation between beliefs and facts. This modal correlation is limited to a proper subset of all possible worlds, namely to nearby ones. Safety requires that what is believed in the actual world is believed in nearby worlds only if it is true. We start with what is actually believed and then consider only those nearby possible worlds in which the agent believes what she believes in the actual world. If the target belief is true in all of these worlds, it counts as safe. In addition, we may distinguish different strengths of safety.[[11]](#footnote-11) The strongest version requires that there is no false belief in any relevant nearby world. Alternatively, we may require the belief to be true in most of the nearby worlds. A third version requires that false beliefs are ruled out in all very close worlds and in most close worlds. For simplicity, I will work with the strongest version of safety here, which excludes false beliefs from all nearby worlds.[[12]](#footnote-12) So far, these considerations suggest the following principle of safety:

(S1) S’s belief that p is safe if and only if the actual world and all relevantly nearby worlds in which S believes that p are p-worlds.[[13]](#footnote-13)

The problem with this principle becomes obvious when one keeps in mind that safety is motivated as a condition for knowledge. However, (S1) is not necessary for knowledge. One can know that p without satisfying (S1). Nozick’s grandmother case nicely illustrates this:

Granny […] is very good at being able to tell whether her grandson is well, provided that she gets to have a good look at him. One day her grandson, who is well, visits her and she gets a good look at him. She sees that he is well and so forms a true belief to this effect. Had her grandson not been well, however, then his family […] would have kept her grandson away from her and told her that he was well regardless. Moreover, Granny would have believed them. (Pritchard 2009a: 30–31)

Intuitively, Granny knows that her grandson is well, although her belief is not (S1)-safe. For she would continue to falsely believe the very same thing if her grandson were ill and her family lied to her about his well-being.

A natural reaction to this counterexample is to keep fixed not only what is actually believed but also *the actual basis* of this belief across nearby worlds. Consequently, one would consider only those nearby worlds in which the epistemic agent has formed the same belief on the same basis as in the actual world. A world in which Granny forms her belief on the basis of testimony instead of her own vision, as in the actual world, would, therefore, be irrelevant to safety. This suggests the following, modified safety principle:

(S2) S’s belief that p is safe if and only if the actual and all relevantly nearby possible worlds in which S believes that p *based on the same method as in the actual world* are p-worlds.[[14]](#footnote-14)

However, this version of safety, for a number of reasons, is too weak. On this understanding, safety would not be sufficient for knowledge. First, consider a case in which someone believes a non-obvious mathematical truth, e.g. 43+88=131, because she is reading it on the top side of a flipped coin that has the negation of the mathematical truth written on its other side. According to (S2), this belief is safe since true mathematical beliefs are true in all worlds even if they rely on absolutely crazy methods, as in this case. Nevertheless, we would not attribute knowledge here. Second, something analogous holds for cases in which someone believes a modally robust, though not absolutely necessary truth (i.e. a proposition that is true only in all nearby worlds) on the basis of a crazy method. Consider, e.g., someone who believes that there is water on earth, on the basis of tossing a coin. This belief is (S2)-safe as well. But it certainly does not constitute knowledge. Third, (S2) does not rule out all versions of the barn-façade case (see Hiller & Neta 2007): imagine that in fake-barn country Henry forms the de re belief that *that is a barn* when he is looking at the only real barn in the vicinity. Intuitively, Henry does not know, since his belief is only luckily true. But his de re belief about the real barn is (S2)-safe. Here is why: continuing to have the same de re belief in nearby worlds implies that the belief is about the same object. Since the real barn could not easily cease to be a barn, this belief will be true whenever it is believed in nearby possible worlds.

The problems with (S2) listed above teach us the following lesson: safety can do the job of ruling out epistemic luck and elucidating knowledge only if it does not keep the actual belief fixed across nearby possible worlds. What should instead be kept fixed is the general belief-forming method that constitutes the basis of the actual belief. Accordingly, a belief should be counted as safe if it is based upon a safe method that *in general* and in a modally robust manner results in true beliefs. Versions of safety that satisfy this condition will be classified under the label “method safety”.

One can distinguish between two versions of method safety. Whereas wide method safety takes as the basis of counterfactual evaluation all beliefs that result from the method-type underlying the target belief, narrow method safety considers only those of the resulting beliefs whose contents are similar to the target belief. This suggests the following definition of *narrow method safety*:

(S3) S’s belief that p is safe if and only if it is produced by a method M, such that all beliefs (i) that are produced by M in the actual and relevantly nearby possible worlds and (ii) whose contents are similar to p are true.[[15]](#footnote-15)

Even this version of safety is too weak to exclude all cases of epistemic luck. Consider, e.g., the following case:

 **Tombola**

 For some esoteric reason, Emma believes whatever she finds written on a card that she randomly draws from a tombola. This tombola contains a number of cards, each of which has a single statement written on it. Some of the cards display mathematical statements that are all necessarily true. Other cards display false contingent non-mathematical statements. These features of the Tombola are modally robust, i.e. the Tombola contains only true mathematical statements and false contingent statements in all nearby worlds. If Emma happens to draw a card with a mathematical statement on it, her corresponding belief will be (S3)-safe. This is true because all statements in the tombola that have a similar mathematical content are true. Nevertheless, Emma’s belief is only luckily true and therefore does not constitute knowledge.

In this respect, *wide method safety* fares better. Here is a working definition:

(S4) S’s belief that p is safe if and only if it is produced by a method M, such that all beliefs that are produced by M in the actual and relevantly nearby possible worlds are true.[[16]](#footnote-16)

According to (S4), Emma’s belief in a mathematical proposition is unsafe because relying on the tombola leads to false beliefs in nearby worlds. Hence, (S4) can explain why Emma, intuitively, does not know. Of course, (S4) leaves open many intricate questions. For example: what is the proper individuation of the relevant type of operative method (this is a version of the ‘generality problem’), and which nearby possible worlds are the relevant ones? But these questions do not point to problems that are specific to safety.[[17]](#footnote-17) In what follows, I will discuss alleged counterexamples to the safety account of knowledge on the basis of understanding safety along the lines of (S4). Let me first highlight that, so far, I have left one further crucial issue open, namely the question of how far the relevant methods of belief formation are extended. Do they start with the cognitive input to the agent’s psychological processes? Or can testifying other minds, external instruments, or external epistemic conditions be treated as their proper parts? I will come back to this issue of the internal or external individuation of methods in section 3. It will also turn out later (in section 4) that we have to elaborate (S4) a bit further by taking into account the safety of one’s reasons in order to deal with more intricate counterexamples.

*2. Is safety sufficient to exclude epistemic luck?*

Many epistemologists accept that knowledge rules out luckily true belief. But not all of them would agree that safety does the trick. In this section, I will discuss some cases that appear to be examples of safe beliefs that are nevertheless luckily true and, therefore, represent knowledge-precluding Gettier cases. I will argue that none of these cases is successful, if we apply the proper conception of safety, namely (S4).

Consider Lackey’s **Southernmost Barn Case**:[[18]](#footnote-18)

While driving through fake-barn country, Janice sees the only real barn in this area at the southernmost point of a field and forms the belief that there is a barn on this basis. The only real barn could not have been placed elsewhere in this area. Moreover, the circumstances and Janice’s motivation for looking at exactly the place where the only real barn is located are modally robust.

Lackey claims that this case is a Gettier case in which veritic luck occurs, although Janice’s belief is safe.

I think we can pose a dilemma for Lackey: if we understand the case in such a way that it constitutes a Gettier case, then Janice’s belief is not (S4)-safe either. If, however, we change the case in such a way that Janice’s belief becomes (S4)-safe, then it ceases to be a Gettier case. The Southernmost Barn Case can be read as a Gettier case. On such a reading, Janice has a modally robust preference or inclination to look at the modally robust location of the only real barn. However, since her preference does not fully determine her behavior, she might nevertheless have looked at some other place. For example, if she were distracted by the oncoming traffic for a moment, she might have looked at a barn façade, thus forming the false belief that there is a barn on this basis. In which case, Janice’s belief would not be safe either. Let us now change the case in such a way that it is guaranteed that Janice’s belief is (S4)-safe:

**Southernmost Residence Case**

Janice has always lived in the southernmost area of the field where the only visible barn-shaped building is a real barn. Janice cannot change her location, since she is severely disabled. All fake barns are out of her sight. She sees the real barn and forms her belief that there is a barn on this basis.

In this case, Janice’s belief is indeed (S4)-safe. But now it also seems much more intuitive to attribute knowledge to her. If this is true, Lackey’s case is not a good example for a safe but luckily true belief.

Let us consider a second case: Coffman’s **Presumptuous Secretary**:**[[19]](#footnote-19)**

An eminent historian, Hank, recently discovered that Abraham Lincoln was born not in 1809 (as most of us think) but in 1806. Hank writes a letter to his friend, Sandy, in which he asserts that Lincoln was born in 1806. When preparing the letter to be sent, Hank’s presumptuous new secretary assumes that he has made a careless mistake about Lincoln’s date of birth, and changes the text so that it says that Lincoln was born in 1809. Unbeknownst to all, the secretary’s printer has just developed the following glitch: when directed to print ‘9’, some other numeral besides ‘9’ is randomly selected and printed instead. As luck would have it, Hank’s letter gets printed as stating that Lincoln was born in 1806. When Sandy receives the letter, she consults it for the answer to the question of when Lincoln was born. Sandy comes to believe, justifiedly and truly, that Lincoln was born in 1806.

Coffman claims in his paper that this case is a Gettier case, although Sandy’s belief is safe.

He is surely right in claiming that this is a Gettier case. However, Sandy’s belief is not (S4)-safe either. Here is why: Sandy believes what he believes on the basis of *believing whatever he reads in the letter.[[20]](#footnote-20)* But now consider the relevant modal environment. In some nearby possible worlds the printer is directed to print ‘9’ but randomly prints some other numeral, e.g. ‘8’ or ‘7’. So it might easily have printed ‘1808’ or ‘1807’. In these cases, Sandy would have acquired a false belief about when Lincoln was born, namely that Lincoln was born in 1808 or that he was born in 1807. Hence it is not the case that all beliefs that are produced by the operative method in the actual and nearby possible worlds are true. Sandy’s belief may be (S2)-safe, but it clearly lacks the relevant (S4)-safety. Once again, this is not a successful counterexample.

Finally, consider Goldberg’s **Trick of Light Case**:[[21]](#footnote-21)

There is a rare optical illusion such that a particular bush on a hill visually appears to Sandy as a rock formation⎯but it can only occur if Sandy is looking at the bush from a particular position. Relative to this position, a real rock formation that could not easily have been located elsewhere lies⎯invisible to Sandy⎯behind the bush on the hill. Sandy comes to believe that there is a rock formation on that hill on the basis of this visual appearance.

Goldberg claims that in this case Sandy’s belief does not constitute knowledge, although it is safe.[[22]](#footnote-22)

Let us take a closer look at the Trick of Light Case. What would be the correct description of the method by which Sandy forms his belief about the rock formation? It seems natural to describe the operative method M as *basing a visual belief about rock formations on one’s visual experiences of those formations*. Under the circumstances stipulated by the case description, the following is true: (1) Sandy’s visual experience possesses an indexical content [that is a rock formation] that is false, since there is no rock formation (but a bush) in the place Sandy is looking at. (2) However, Sandy’s more general belief *that there is a rock formation on that hill* is true and (S2)-safe. But method M could easily have produced false demonstrative beliefs in the Trick of Light case. For example, Sandy might have falsely believed that *that* [the bush] is a rock formation or that a rock formation is over there [the place Sandy is looking at]. These beliefs are based on the same method as Sandy’s actual belief, yet they classify the perceived object, namely the bush, falsely as being a rock formation. Hence, Sandy’s belief is not (S4)-safe.

Let me wrap-up the results of this section: on closer scrutiny, the most prominent examples of luckily true, yet safe belief turned out to be illusive. The main reason why people falsely believe that the beliefs in the examples are safe is that they use some inappropriate conception of safety rather than (S4)-safety.

*3. Is safety necessary for knowledge?*

Whereas the critics of the last section claim that safety is too weak to account for knowledge, the critics discussed in this section attack safety from the opposite direction. According to them, the risk of believing something falsely, even if this risk is very high, does not always undermine knowledge. They believe that there are clear cases of luckily true or unsafe knowledge. In order to get a clearer view of what is at issue here, let me introduce the distinction between intervening and environmental luck.[[23]](#footnote-23) In a case of *intervening luck*, the reason that one uses is unrelated to the truth-maker of the belief based on that reason. It is a lucky coincidence that the situation is such that the truth-maker occurs when the reason is present. The world intervenes in turning an otherwise false belief into a true one. The standard example of a Gettier-case in which intervening luck obtains is the sheep case. In this case, a farmer believes truly that there is a sheep in the field, not by looking at a real sheep but rather by looking at a sheep-shaped dog. The belief is then made true by some real sheep in the field that is in fact invisible to the farmer. Contrast this with cases of *environmental luck*, in which nothing intervenes between the belief and the truth-making fact. It is just that the environment is epistemically hostile, such that the agent might easily have been deceived. The standard example of a Gettier-case involving environmental luck is the barn-façade-case. In that case, Henry sees a real barn and is thereby adequately related to the truth-maker of his belief. Nevertheless, his belief is luckily true.

The safety critics of this section claim that cases of environmental luck always involve unsafe belief, although they are not always Gettier cases. Let me distinguish between three different types of environmentally luckily true beliefs.

(Type 1) The *actual world* contains deceptive objects that would easily have misled the epistemic agent.

(Type 2) Close possible worlds *rather than the actual world* contain deceptive objects that would easily have misled the epistemic agent.

(Type 3) Close possible worlds contain defective instruments, deceiving informants, or unfavorable epistemic conditions that would easily have misled the epistemic agent.

The critics of safety now want to claim that only (Type 1), but not (Type 2) or (Type 3) cases of environmental luck constitute Gettier cases. If this were true, then (Type 2) and (Type 3) cases would be cases of unsafe knowledge. In what follows, I will argue that (Type 1) and (Type 2) cases should be treated alike, namely as cases of unsafe belief in which the agent also lacks knowledge. In contrast, (Type 3) cases can be understood as (S4)-safe beliefs in which environmental luck is absent as well. This becomes plausible if we individuate the operative methods in an externalist way. I will argue that this is the right way to go.

Although it is not beyond controversy that Henry does not have knowledge in the barn-façade case, I will treat (Type 1) cases as cases in which environmental luck destroys knowledge.[[24]](#footnote-24) What about (Type 2) cases? Neta and Rohrbaugh (2004) claim that cases in which deceptive objects do not exist in the actual world but only in very close possible worlds are compatible with actual knowledge. They take it that the following case illustrates this:

**Water**[[25]](#footnote-25)

Ram believes truly that he is drinking pure, unadulterated water on the basis of seeing and tasting it. However, Adele, who is standing next to him, would very easily⎯and unbeknownst to Ram⎯have polluted the water with a tasteless and colorless toxin, if she had not won the lottery. If she had lost the lottery, Ram would have falsely believed that he was drinking pure, unadulterated water.

Neta and Rohrbaugh take it that, intuitively, Ram does actually *know* that he is drinking pure, unadulterated water.[[26]](#footnote-26) They also have an explanation of why Ram but not Henry knows: whether deceptive objects have the power to prevent epistemic agents from possessing knowledge depends, according to them, on whether these objects exist in the actual world, as in the barn-façade case, or only counterfactually, as in the water case.[[27]](#footnote-27) The bad news for safety-proponents is that they cannot account for any epistemologically significant distinction between these cases. This is true even with respect to (S4)-safety. In both cases, the same method is used in the actual and the counterfactual situation. What varies are the registered objects. But on any reasonable account of belief-forming methods, they must be individuated independently of their target objects. Moreover, both Henry and Ram could easily have been misled by using their respective methods. So, from the perspective of safety, there is no significant epistemological difference between the barn-façade case and the water-case that could explain why we attribute knowledge in the latter but not in the former.

There is only one way to defend safety against this objection. One must dispute that Ram knows in the water-case.[[28]](#footnote-28) Of course, one could just bite the bullet here by claiming that Ram does not know despite one’s intuition to the contrary. But I don’t want to take this way out because I rely on the methodological assumption that intuitions about cases constitute a reliable guide to the analysis of knowledge. Instead, I want to use our intuition about a case that is fairly analogous to the water case, in order to clarify our intuitive verdict about the latter. Consider the following case:

**Playful Barn Demon**

Guy is driving through the countryside and sees a genuine barn under excellent observation conditions. On the basis of his vision he believes that there is a barn. In fact, all the other barn-shaped buildings in the vicinity are real barns. However, there is malicious demon watching him. He would have transformed all the real barns, including the one Guy is actually looking at, into barn- façades, if heads had won a coin-toss just performed. Does Guy know that there is a barn?

If you share my intuition that in this case Guy does not know, then you should also dispute that Ram has knowledge in the water case. This is because in both cases the deceiving objects exist only counterfactually.

Let us now consider (Type 3) cases, which critics take to be counterexamples to safety as being necessary for knowledge. All of these cases have a common structure: the epistemic agent believes truly that p, (a) on the basis of using a particular instrument, (b) on the basis of trusting a certain informant, or (c) on the basis of forming her belief under specific external conditions. Then a further element is added to the case description. In a very close world, the instrument actually used is defective, the informant who has been trusted is lying, or the external conditions under which the agent’s belief is formed are turned into bad ones. It is also true that if the epistemic agent were in the counterfactual situation, she would not notice the difference and, therefore, would form a false belief. The critics now make two crucial claims about cases like these. (1) Intuitively, we attribute actual knowledge to the epistemic agents in all these cases. (2) The counterfactual changes in the environment do not affect the identity of the method that is operative in the agent’s belief-formation in any of these cases. In other words, the methods remain the same across the counterfactual variations. If (1) and (2) are both true, then even (S4)-safety is not necessary for knowledge. This is so because each of the cases demonstrates that a method that would have easily resulted in a false belief nevertheless generates knowledge in the actual world. And this is in direct contradiction to what (S4)-safety claims.

Let me introduce the paradigm cases of (a), (b), and (c) starting with Kelp’s case of a *counterfactually defective instrument*:

**Russell’s Grandfather Clock**

The demon wants Russell to believe that it is 8:22 am no matter what the time really is. But the demon is lazy. So, when Russell comes down the stairs at 8.22 and relies on his old grandfather clock downstairs as always, then it need not intervene. But if Russell were to be a bit early or late, the demon would freeze the clock at 8:22 such that Russell then would falsely believe that it is 8:22. ‘Suppose, as it happens, Russell does come down the stairs at 8:22. It is 8:22. The grandfather clock is working reliably as always. Intuitively, I take it, Russell knows that it’s 8:22 upon reading the clock.’ [[29]](#footnote-29)

In his evaluation of this case, Kelp assumes without much ado that Russell uses the same method in the actual world and in the close possible world, namely “reading the clock.” Hence, Russell’s belief is not (S4)-safe.[[30]](#footnote-30)

Next, there are cases of *counterfactually lying informants*, such as Comesana’s

**Halloween Party Case**[[31]](#footnote-31)

A Halloween party is being held at Andy’s house and Juan is invited. The house is very difficult to find, so Andy hires Judy to stand at a crossroads and direct people towards the house. Unbeknownst to Juan, Andy doesn’t want Michael to go to the party, so he tells Judy that if she sees Michael she should immediately phone Andy so that the party can be moved to Adam’s house. Juan seriously considers dressing up as Michael for the party, but at the last moment doesn’t. When he goes crossroads, he asks Judy where he can find the party, and she tells him that he can find it down the road on the left. Juan truly believes what Judy tells him.

Comesana claims that in this case Juan knows where he will find Andy’s party. He also assumes that Juan uses the same belief-forming method in the actual and the considered nearby world, namely relying on what Judy tells him. But then again, this is a counterexample to the claim that (S4)-safety is necessary for knowledge.[[32]](#footnote-32) Why do I claim that this is a case of a counterfactually *lying* informant? Consider what would have happened if Juan had dressed up as Michael. Judy would believe that she was encountering Michael. She would have given him instructions for how to find Andy’s party. And she would then have let Andy move the party to a different place. Hence, she would have intentionally misinformed Juan. We can call this a “lie”.

Finally, here is Sosa’s paradigm example of cases of *counterfactually bad epistemic conditions*:

**Kaleidoscope**[[33]](#footnote-33)

Kyle sees a surface that looks red in ostensibly normal conditions. But it is a kaleidoscope surface controlled by a jokester who also controls the ambient light, and who might as easily have presented Kyle with a red-light+white-surface combination as with the actual white-light+red-surface combination. Does Kyle then know that the surface he sees is red when the jokester presents him with the good combination, despite the fact that, even more easily, he might have presented Kyle with the bad combination?

Sosa claims that Kyle knows under the good light conditions. Moreover, he maintains that Kyle’s belief is unsafe since the actually operative method (trusting one’s eyesight) might easily have led Kyle to form a false belief, namely that the surface is red when it is in fact a white surface illuminated by red light.

All three counterexamples rely on the assumption that the method of belief-formation is counterfactually robust, i.e., that the epistemic agents use the same method in the actual and the counterfactual situation. This assumption is not as innocent as it may at first appear. It presupposes that the relevant methods are individuated in an internalist manner. What is the difference between internally and externally individuated methods of belief-formation? A method is individuated internally if it supervenes on the agent’s dispositions to react to stimuli. On this view, if, e.g., the epistemic agent relies on an instrument no matter whether it is working properly or is defective, there is no difference in the operative method. In contrast, the method is externally individuated if it does not supervene on the agent’s reactive dispositions but is partly constituted by the actual nature of the external conditions. Applied to the above example, it would make a difference to the operative method whether the instrument on which the agent is relying is in fact working properly or broken.

One can easily see that this distinction is crucial for the evaluation of the above counterexamples. Let us take for granted that in all these cases the agent actually knows. First, consider Russell’s Grandfather Clock. When Russell comes down the stairs at 8:22 am, he is relying on a *properly working* clock. From an internalist perspective, it is crucial to this case that Russell is not sensitive to the proper working of this clock. If the clock were to be stopped by the demon, he would still rely on it. According to the internalist, Russell therefore actually relies on *reading the grandfather clock no matter whether it is properly working or not*. This method is clearly unsafe since it would very easily have led to a false belief in the current environment. Contrast this with the externalist perspective on the case. According to this view, it does not matter whether Russell is able to distinguish between the properly working and the stopped clock. Since he is actually relying on the properly working grandfather clock, its proper working should be kept fixed across the nearby worlds that are relevant to safety. Given this reading, Russell is actually relying on a safe method, namely the properly working grandfather clock. So the case would not constitute a counterexample to safety as a necessary condition for knowledge.

Next, consider the Halloween Party Case. In the actual situation, Juan is relying on the sincere Judy as his informant. From the internalist perspective, it matters that Juan is not sensitive to whether Judy is sincerely giving him directions to Andy’s party or not. Because of his insensitivity, the difference cannot play a role in the individuation of the relevant method. Accordingly, Juan uses the method of relying on whatever Judy is telling him. This method is clearly unsafe since it would easily have led to a false belief about the way to Andy’s party if Juan had dressed up as Michael. Now consider the same case from the externalist perspective. In fact, Juan is relying on an informant who is sincere about what she takes to be the correct location of Andy’s party. According to the externalist, we should keep Judy’s honesty fixed across the nearby possible worlds. And if we do so, then the directions given by the lying Judy, although they are very close to actuality, are not relevant because they change the externally individuated method. But then Juan is actually relying on a safe method.

Finally, consider the Kaleidoscope Case. In the actual situation Kyle is relying on his visual experience under good lightening conditions, namely surface-illumination with white light. From the internalist perspective, the lightening conditions do not matter for the individuation of the method as long as the agent is unable to notice the difference. Hence, the internalist judgment about the current case seems straightforward. According to the internalist, the method that is used in the actual situation and the method that is used in the counterfactual situation are the same, namely believing what one experiences visually. And this method is unsafe in the current case. Is the externalist better off with respect to the Kaleidoscope Case? At first, we might think that this is not so. For how can external observation conditions be part of a *cognitive* belief-forming method, even from an externalist point of view? However, the question of whether one really wants to count the external epistemic conditions as part of the belief-forming method or not is a purely verbal matter. What I take to be relevant here is whether we keep the actual observation conditions fixed across the nearby possible worlds when we evaluate the case with respect to safety. Whatever is kept fixed will then be called “method” in a purely technical sense.[[34]](#footnote-34) As long as one keeps this in mind, the externalist can now argue that in the Kaleidoscope Case we keep the actually good lightening conditions fixed in our evaluation of safety. If we do so, Kyle is actually relying on a safe method.

The previous discussion of counterexamples has a clear result: these examples are counterexamples to (S4)-safety as a necessary condition of knowledge only if the relevant methods are individuated internally. However, critics of safety have not argued for the truth of this condition. They simply presuppose it. Before we look more closely at the plausibility of external versus internal method individuation, let us take stock of what has been achieved. Take for granted, for the moment, that we can reply to the considered counterexamples by individuating the methods in an external way. Nevertheless, there seem to be two worries left. First, one still might worry that this strategy is not sufficient to rule out every possible counterexample in the vicinity. Consider, e.g., the following modification of the original Halloween Party Case:

**Fully Automatic Halloween Case[[35]](#footnote-35)**

There is a Halloween party at Andy’s house and Juan is invited. The house is very difficult to find, so Andy employs a device that stands at a crossroads and directs people towards his house. Unbeknownst to Juan, Andy doesn’t want Michael to come to his party, so he manipulates the device such that if it registers Michael (through a built-in camera), it sends a signal to Andy, who then will move his party to Adam’s house. Juan seriously considers dressing up as Michael, but at the last moment doesn’t. When he goes to the crossroads, he is directed to the party by the device. Juan truly believes that he will find Andy’s party at the location to which the device directs him.

Intuitively, in this revised case Juan knows where he will find Andy’s party, if he does not dress up as Michael. But in this case, no counterfactually lying informant is involved, as in the original case. Neither is Juan relying on a counterfactually defective or broken instrument. Hence, one might worry that in this case the reliance on an externally individuated method does not help with saving safety.

In response, let me first present a more abstract characterization of the Fully Automatic Halloween Case. In the actual situation, Juan relies on an instrument (the directing device) that correctly indicates the party’s location under the current conditions. If Juan had dressed up as Michael, he would have interfered with the good external conditions of the device’s correct indication of the party’s location. These good external conditions are violated if the epistemic agent’s, i.e. Juan’s, mere presence influences his belief’s objects unnoticeably. That is exactly what happens when Juan unintentionally induces the device to misrepresent him as Michael and to signal this misinformation to Andy. Here we have a case in which the agent unnoticeably influences the objects of his belief by his mere presence. On the basis of this abstract characterization, the externalist about method individuation may claim the following about this case: in the actual situation, Juan forms his belief by using an instrument under good working conditions. If he had disguised himself as Michael, this would have turned the good conditions into bad ones. However, from an externalist point of view, we should keep the actual observation conditions fixed. The fact that under different observation conditions Juan would have formed a false belief is, therefore, irrelevant to safety. Even the modified Halloween Party Case does not seem to be a good counterexample after all.

There is the second worry: by keeping the external conditions of the belief-formation fixed, one also rules out nearby worlds that, intuitively, seem to be relevant to undermining knowledge. Consider the ordinary barn-façade case. One might argue that the externalist must keep the perceptual relation to the genuine barn fixed such that nearby worlds in which Henry encounters fake barns are simply irrelevant to the epistemic evaluation. According to this view, Henry’s belief would be safe. But that would generate the wrong verdict about the case. Intuitively, Henry does not know in the barn-façade case. However, if safety is sufficient for knowledge, the externalist would have to conclude that Henry knows.[[36]](#footnote-36) It seems to me that this initial worry is unreasonable. Although conditions that are external to the epistemic agent’s dispositions can be relevant to the method-individuation from the externalist point of view, this does not include the belief’s object itself. Even for the externalist, the method of belief-formation must be individuated independently of the facts that are relevant to the method-evaluation. Otherwise true beliefs would automatically constitute knowledge. Hence, no matter where exactly the line between the method and its target objects has to be drawn, a perception of a genuine barn can never guarantee a safe method­⎯not even for the externalist.

As it turns out, all counterexamples to safety as necessary for knowledge may be rejected, if an external individuation of the operative method is appropriate.[[37]](#footnote-37) But does an external individuation of the relevant method have any plausibility? Let me start by giving a few examples of cases in which such an external individuation is very plausible. First, when we evaluate the reliability of devices or instruments, we do not check whether they indicate correctly in all nearby worlds. We rather test those devices under normal *external* conditions. When we evaluate, e.g., the reliability of a car engine, we do not check whether it runs under water or at -70 degree Fahrenheit. When we evaluate the reliability of a barometer, we do not test it at the top of Mount Everest. And when we evaluate the reliability of a compass, we do not test it in the vicinity of a strong magnet. So, we keep the normal external conditions fixed in our evaluation of those instruments. One might just as well say that the evaluated method is the use of an instrument under specific (normal) conditions. The same seems to apply to our evaluation of cognitive processes. When we evaluate vision we do not test its reliability in all nearby situations, but only under good epistemic conditions, such as, e.g., clear view, appropriate distance, and appropriate size of the object. This suggests that we evaluate those processes always relative to external conditions that we keep fixed.[[38]](#footnote-38) Second, consider a case in which Chris always looks at the grandfather clock when he wants to know what the actual time is. However, this clock might easily have been replaced by another fake clock that looks exactly alike but is not working properly. If it were to be replaced, Chris would not notice the difference. What is the correct description of Chris’ method when he looks at the genuine grandfather clock? I suggest that it would be intuitive to maintain that Chris then is relying on the grandfather clock rather than judging that he is relying on a clock that looks like the grandfather clock. If this is the intuitive judgment, it shows that we are inclined to individuate the operative method externally. Third, imagine that Olivia is in fake-clock country. She actually wants to consult a trustworthy clock and just happens to read the only properly working genuine clock in the vicinity. All the other clocks are similar-looking dummies. Olivia forms her belief upon reading the real clock, but she might easily have formed her belief on reading one of the dummies instead. Intuitively, Olivia relies on reading a genuine clock as her method in this case. Once again, this shows that the relevant methods are externally individuated.

If an external individuation of the relevant methods is intuitively plausible, the next question is which factors determine this external individuation. Since we are looking for the operative method *type*, it is not sufficient to refer to the particular token events that cause the target belief. Three different kinds of factors seem to be relevant here: (i) the normal external stimuli of the target beliefs; (ii) the intentions of the epistemic agent together with the distal causal chain that leads to the target belief; and (iii) teleological factors. Let me illustrate what is meant here using the examples from above. When we evaluate native cognitive processes such as, e.g., vision, we check their reliability under the external conditions for which they are (adaptively) designed. In these cases, teleological factors are dominant in determining the individuation of the relevant method-type. When, on the other hand, we claim that Chris in the above case is relying on reading the grandfather clock (although he cannot discriminate it from cleverly made dummies), then the normal external stimuli of his beliefs about the correct time play the dominant role. In fake-clock country, one cannot claim that reading a genuine clock is the normal cause of Olivia’s beliefs about the correct time. In this case, however, Olivia’s intention to rely on a properly working clock, together with her actually reading such a clock, determines the relevant method.

Let us now use these relevant factors to explain the correct individuation of the operative methods in the paradigm counterexamples. Consider again Russell’s Grandfather Clock. In this case, no actual intervention of the demon has taken place. Russell *always* reads the properly working grandfather clock when he wants to know what the time is. Here it seems quite natural to claim that the operative method is reading the properly working grandfather clock, because this reading is the normal cause of Chris’ beliefs about the correct time. Next, consider the Halloween Party Case. In contrast to the previous case, there is no historical background for normal causes of belief, like the one regarding where to find Andy’s party. Here the relevant method is partly determined by Juan’s expectations about Judy. He trusts her when she gives him instructions about where to find the party. So Juan is expecting Judy to be sincere in her communication. This expectation picks out Judy’s actual sincerity as a crucial feature of the relevant method. Finally, consider Sosa’s Kaleidoscope Case. In this case, one can use the information that illumination with white (neutral) light is the normal condition under which vision has been designed to function properly. This evolutionary perspective on native cognitive processes explains why these normality conditions, if they actually obtain, must be understood as constitutive parts of the actually operative method.

To sum up, three things should be noted. First, it turns out that the counterexamples to safety as being necessary for knowledge stand only if the operative methods of belief-formation are internally individuated. Second, there is ample evidence that an external individuation of the operative methods is plausible in many cases. Among the method-individuating factors are normal external causes, aspects of the belief’s actual history that are picked out by the agent’s intentions, and teleological aspects. Third, when we take these facts into account in our analysis of the counterexamples, the operative methods turn out to be safe in all cases. The strategy of using these counterexamples to refute the necessity of safety for knowledge, therefore, can be resisted.

*4. Is safety sufficient for knowledge?*

In section 2, I addressed the question of whether there are counterexamples to the view that safety is sufficient to rule out veritic luck. As it turned out, the alleged counterexamples are not successful. In this section, I will address the question of whether the absence of veritic luck is sufficient for knowledge. If this is the case, then safety is not only necessary but also sufficient for knowledge. Robust anti-luck epistemology then wins the day.

Before discussing the most eminent counterexample to robust anti-luck epistemology, let me quickly address a more straightforward objection to the claim that safety is sufficient for knowledge. One might just argue that since knowledge is defeasible and safety does not rule out defeaters, safety cannot be sufficient for knowledge in general.[[39]](#footnote-39) Suppose Lottie truly believes that her ticket will lose in a lottery that is in fact, unbeknownst to her, rigged such that there will be *no* winning ticket. In this case, Lottie’s belief seems to be safe since it could not easily have been false. Nevertheless, one intuitively hesitates to attribute knowledge to Lottie. What happens here is that Lottie takes the lottery to be a fair one although it is not. This false belief constitutes a misleading undercutting defeater for her knowledge. It falsely suggests that her target belief about her losing ticket is unsafe. And this misinformation destroys her knowledge.[[40]](#footnote-40) So safety cannot rule out defeating factors. One should not, therefore, claim that safety is sufficient for knowledge without any qualification. A more reasonable view would be that safety is sufficient for *undefeated* knowledge.

Pritchard has designed the following counterexample to refute even this qualified version of the sufficiency-claim:

**Temp Case**[[41]](#footnote-41)

Temp’s job is to keep a record of the temperature in the room that he is in. He does this by consulting a thermometer on the wall. As it happens, this way of forming his belief about the temperature in the room will always and in all nearby worlds result in a true belief. The reason for this, however, is not because the thermometer is working properly, since it isn’t⎯it is fluctuating randomly within a given range. Crucially, however, there is someone hidden in the room next to the thermometer who, unbeknownst to Temp, makes sure that every time Temp consults the thermometer the temperature in the room is adjusted so that it corresponds to the reading of the thermometer. This helper is present not only in the actual but also all nearby possible worlds.

Pritchard claims about this case that Temp’s belief about the temperature in the room is safe. Moreover, he finds it clearly intuitive that we do not ascribe knowledge to Temp in this case.[[42]](#footnote-42) Finally, he tries to explain this intuition by saying that the direction of fit between beliefs and facts is wrong. “[T]he facts are in effect responding to the agent’s beliefs rather than vice versa.”[[43]](#footnote-43)

In what follows I will pose a dilemma for Pritchard: either we take the Temp Case as it stands, or we revise it substantially. If we take the case as it stands, then the intuition that Temp does not know the temperature of the room is irresistible. But then his belief is not (S4)-safe either. Or we revise the case such that (S4)-safety is guaranteed. Then we cannot deny that Temp has knowledge. No matter which horn Pritchard chooses, the case does not pose a real threat for ambitious anti-luck epistemology.

Let me start by defending the first horn. To begin with, although, strictly speaking, the description entails that Temp’s beliefs about the temperature of the room counterfactually co-vary with the facts, the implementing mechanism creates the impression that Temp’s belief is not safe. The hidden person could easily make a mistake in adjusting the temperature of the room to the reading of the thermometer. In order to ensure that Temp’s beliefs are really safe, the hidden person must be substituted for a properly working device that adjusts the temperature of the room to the reading of the thermometer. This would turn the whole complex that is constituted by the broken thermometer and the supplementary device into a properly functioning thermostat that regulates the temperature of the room. Second, even beliefs that are based on the reading of a thermostat are not safe, since whenever the reading of the thermostat randomly changes, it will always take some time for the temperature of the room to adjust. During this interval, beliefs based on the reading will be false. This problem can be fixed by switching to an example in which the adjustment is made without any delay. Third, on an initially plausible understanding of the case, Temp takes the device on the wall to be a properly working thermometer. On the basis of this background belief he forms beliefs about the temperature in the room that correspond to the readings of the thermometer. Part of the operative method then is the false belief that the device on the wall is a properly working thermometer. The fact that the method involves a false belief is not by itself sufficient to demonstrate that the method is unsafe. But it seems obvious that the same method could have easily led to false output beliefs. On the basis of assuming that the thermometer is properly working and recognizing some particular reading of it, Temp might easily conclude that the sensor within the thermometer is registering the corresponding temperature. However, this conclusion would be false. Hence, Temp’s target belief is not (S4)-safe. This last problem can be fixed only if one assumes that Temp’s cognitive processing is very simple, such that it does not involve any beliefs about the proper working of the thermometer, but suggests believing whatever the device indicates.

We can design a case that guarantees the relevant type of safety only if we revise the original case along the following lines: (i) the hidden person must be substituted for a device; (ii) the facts must be instantaneously adjusted to the instrument’s readings; (iii) the agent’s cognition must not involve any false beliefs about the proper working of the instrument that is being used. This brings me to the second horn of the dilemma: if we revise the original case such that safety is guaranteed, we will no longer be inclined to deny knowledge. This can be illustrated by the following case:

**Mr. Light**

It is Mr. Light’s job to register whether the lights are on or off in the rooms of a big hotel. He consults a table with many little lights on it, each of which refers to a specific room in the hotel. On noticing that some little light is on, Mr. Light immediately believes that the lights in the respective room are on, and vice versa. Unbeknownst to Mr. Light, none of the little lights registers the status of the lights in the respective hotel room. On the contrary, their randomly activated on/off positions switch the lights in the respective rooms on/off. When Mr. Light sees that the little light for room #23 is on, he forms the true belief that the lights are on in room #23.

Mr. Light’s belief is (S4)-safe. Believing whatever the lights say could not easily lead to a false belief in this case. However, I also have the clear intuition that Mr. Light knows.

One might want to dispute that Mr. Light knows by using Pritchard’s argument that there is a wrong direction of fit between beliefs and facts in this case. As Pritchard puts it, “the facts are in effect responding to the agent’s beliefs rather than vice versa.”[[44]](#footnote-44) But this objection would be misleading. Neither in Mr. Light nor in the original Temp Case are the truth-making facts caused by the target beliefs. Strictly speaking, it is not the agent’s belief itself but the reasons for his belief that causes the belief’s truth-maker. In the original Temp Case, “the temperature in the room is adjusted so that it corresponds to the reading of the thermometer.”[[45]](#footnote-45) Hence, it is the reading of the thermometer, i.e. the basis of Temp’s belief, rather than this belief itself, that triggers the hidden person to adjust the temperature. In Mr. Light, it is the little light rather than the belief based upon it that causes the lighting status in the room, i.e. the truth-making fact. I do not see why these kinds of cases should pose a problem for the attribution of knowledge. Consider an analogous case: I believe that I will raise my arm in a second on the basis of my conscious decision to raise my arm in a second. This decision will cause the corresponding behavior that makes my belief true. If there can be any knowledge of future behavior, I do not see why we should hesitate to attribute knowledge in this case.

But maybe the real problem runs deeper than this here. What about a case in which the truth-makers are really caused by the agent’s beliefs rather than by her reasons? Consider the following modification of Mr. Light:

 **Mr. Light\***

A demon adjusts the lighting status of the hotel rooms to Mr. Light’s beliefs (about the light in the rooms), which are based on the random variation of the little lights on his table. When Mr. Light sees that the little light for room 23 is on, he forms the true belief that the lights are on in room 23.

 Intuitively, one would not attribute knowledge to Mr. Light, although his belief seems to be (S4)-safe.

In response to this counterexample, one can chose either of two strategies. According to the first, we finally have to move away from robust anti-luck epistemology and follow a competence-based virtue account of knowledge. Accordingly, knowledge requires that all beliefs that are produced by the actually operative method M in the actual and in relevantly nearby worlds are not only true, but are true *because* they are produced by M*.*[[46]](#footnote-46) It seems clear that the method used by Mr. Light does not explain why his beliefs about the lights in the hotel rooms could not easily have been false. However, this explanatory requirement cannot be motivated by a pure safety-account. Safety requires a modal correlation between beliefs and facts and is silent about explanatory relations.

Fortunately, there seems to be an alternative way of getting rid of the above counterexample. According to this alternative method, a safe belief requires that the operative method maps onto output beliefs only those reasons that exist in nearby possible worlds, only if the propositions that would be believed if the method were executed are true in those worlds. What is effectively required by this conception of safety? We start with the actually operative method. Next, we sort out which reasons are mapped onto which output beliefs using this method. We then consider whether this method produces false beliefs in any relevant nearby world (which is required by (S4)-safety) *and whether any of the reasons that are selected by the method in the actual and nearby possible worlds obtain in any nearby world in which the method is not executed but the proposition that would be believed if the method were executed is false.* If there are any such nearby worlds, then the target belief does not satisfy the current conception of safety. On this conception, safety requires operative methods that select *safe reasons* only. This is expressed by the following definition:

(S5) S’s belief that p is safe if and only if it is based on a belief-forming method M, such that any reason that is selected by M as input obtains in the actual or relevantly nearby possible worlds only if the proposition that is believed in those worlds, or would be believed if M were, contrary to the facts in those worlds, executed, is true in those worlds.

What is required by (S5) can be illustrated with an example. Suppose you rely on the readings of a thermometer when you form beliefs about a room’s temperature. To count as (S5)-safe, all beliefs that are based on these readings must be true in nearby worlds. In addition, even in those worlds in which one does not use the readings of the thermometer to form one’s beliefs about the temperature of the room, the readings must be correlated with the correct temperature of the room. But one can easily see why in Mr. Light\* the target belief fails to be (S5)-safe: whenever Mr. Light does not look at the little lights on his table and, hence, does not form a corresponding belief about the lights in the hotel rooms, the lights do not correlate with the lighting conditions in the rooms. This is because the demon intervenes only *after* Mr. Light has formed a belief. So (S5)-safety can explain why in Mr. Light\* the target belief is unsafe and does not constitute knowledge.

*4. Conclusion*

The safety-account of knowledge relies on the simple and intriguing idea that knowledge is nothing but true belief that could not easily have been false. It requires a robust correlation between belief and truth in the nearby modal environment. This paper has addressed three types of counterexamples to safety: (i) Gettier-like examples of epistemic luck that do not appear to be ruled out by safety; (ii) cases of unsafe knowledge; and (iii) counterexamples to robust anti-luck epistemology. It turns out that these counterexamples succeed only if one presupposes a very narrow understanding of safety. However, there are versions of safety that are immune to these counterexamples; and none of safety’s critics have persuasively argued that these versions of safety are untenable. One can therefore conclude that safety accounts of knowledge have not yet been refuted. More importantly, in engaging with the counterexamples, a viable conception of safety has been progressively delineated. A promising version of safety requires belief-forming methods that are widely safe, externally individuated, and that select only safe reasons. It also turns out that at least some of these qualifications are not simply ad hoc but are well motivated by a range of independent reasons. After all, the prospects of safety may not be as dim as many epistemologists, including Sosa himself, currently proclaim.

**References**

Alspector-Kelly, Marc 2011: “Why safety Doesn’t Save Closure”, in: Synthese 183: 127–142.

Bernecker, Sven 2012: “Sensitivity, Safety, and Closure”, in: Acta Analytica 27: 367–381.

Bogradus, Tomas 2012: “Knowledge under Threat”, in: Philosophy and Phenomenological Research 88: 289–313.

Bogradus, Tomas and Marxen, Chad 2014: “Yes, Safety is in Danger”, in: Philosophia 42: 321–334.

Bradley, Darren 2015: A Critical Introduction to Formal Epistemology. London: Bloomsbury.

Brancano-Berrocal, Fernando 2014: “Is Safety in Danger?”, in: Philosophia 42: 63–81.

Coffmann, E.J. 2010: “Misleading Dispositions and the Value of Knowledge”, in: Journal of Philosophical Research 35: 241–258.

Comesana, Juan 2005: “Unsafe Knowledge”, in: Synthese 146: 395–404.

Dodd, Dylan 2012: “Safety, Skepticism, and Lotteries”, in: Erkenntnis 77: 95–120.

Goldberg, Sanford 2015: “Epistemic Entitlement and Luck”, in: Philosophy and Phenomenological Research 91: 273–302.

Hetherington, Stephen 2006: “How to Know (that Knowledge-that is Knowledge-how)”, in: Hetherington, Stephen (ed.), Epistemology Futures, Oxford: Oxford University Press, 71–94.

Hiller, A. and Neta, R. 2007: “Safety and Epistemic Luck”, in: Synthese 158: 303–313.

Horvath, Joachim and Wiegmann, Alex 2016: “Intuitive Expertise and Intuitions about Knowledge”, in: Philosophical Studies 173: 2701–2726.

Kelp, Chris 2009: “Knowledge and Safety”, in: Journal of Philosophical Research 34: 21–31.

Kripke, Saul 2011: “Nozick on Knowledge”, in: Kripke, Saul: Philosophical Troubles. Collected Papers Vol. 1. Oxford: Oxford University Press, 162–224.

Lackey, Jennifer 2006: “Pritchard’s Epistemic Luck”, in: Philosophical Quarterly 56: 284–9.

Neta, Ram and Rohrbaugh, Guy 2004: “Luminosity and the Safety of Knowledge”, in: Pacific Philosophical Quarterly 85: 396–406.

Pritchard, Duncan 2005: Epistemic Luck. Oxford: Oxford University Press.

Pritchard, Duncan 2009a: Knowledge. London: Palgrave Macmillan.

Pritchard, Duncan 2009b: “Safety-Based Epistemology: Whither Now?”, in: Journal of Philosophical Research 34: 33–45.

Pritchard, Duncan 2015: “Anti-Luck Epistemology and the Gettier Problem”, in: Philosophical Studies 172: 93–111.

Sainsbury, Mark 1997: “Easy Possibilities”, in: Philosophy and Phenomenological Research 57: 907–919.

Sosa, Ernest 1996: “Postscript to ‘Proper Functionalism and Virtue Epistemology’”, in: Kvanvig, J. (ed.), Warrant in Contemporary Epistemology, Lanham: Rowman & Littlefield, 271–81.

Sosa, Ernest 1999: “How to Defeat Opposition to Moore”, in: Philosophical Perspectives 13: 141–54.

Sosa, Ernest 2007: A Virtue Epistemology. Apt Belief and Reflective Knowledge. Oxford: Oxford University Press.

Sosa, Ernest 2009: “Timothy Williamson’s Knowledge and its Limits”, in: Greenough, P. & Pritchard, D. (eds.), Williamson on Knowledge, Oxford: Oxford University Press, 203–216.

Williamson, Timothy 2000: Knowledge and its Limits. Oxford: Oxford University Press.

1. See Sosa 1999 for the introduction of the term. The idea is already present in Sosa 1996 and Sainsbury 1997. Williamson 2000 is another early proponent of the safety account, although he does not believe that one can give a non-circular definition of knowledge as safe belief. [↑](#footnote-ref-1)
2. See Sosa 1999: 146. For a dissenting view, see Dodd 2012. [↑](#footnote-ref-2)
3. See Sosa 1999: 143, Pritchard 2005: 168 for a defense of safety’s closure. For a dissenting view on safety, see Alspector-Kelly 2011. Bernecker 2012 claims that sensitivity does not violate closure. [↑](#footnote-ref-3)
4. See Kripke 2011: 167–68. [↑](#footnote-ref-4)
5. For the following, see Pritchard 2005, 2009a. [↑](#footnote-ref-5)
6. See Pritchard 2005: 145–152 for a more comprehensive analysis of veritic luck. [↑](#footnote-ref-6)
7. See Pritchard 2015: 96–97. [↑](#footnote-ref-7)
8. See Pritchard 2009a: 23 for this distinction. [↑](#footnote-ref-8)
9. See Pritchard 2009a: 51. [↑](#footnote-ref-9)
10. Sosa finds fault with safety because it does not require the belief to result from competences; he also disputes that safety is necessary for knowledge. See Sosa 2009: 206–7. [↑](#footnote-ref-10)
11. See Pritchard 2005: 161–173 for a discussion of the different strengths of the safety principle. [↑](#footnote-ref-11)
12. The best motivation for this version is that it explains why we do not know that we lose the lottery if we truly believe this on the basis of our general knowledge about the odds. [↑](#footnote-ref-12)
13. See Sosa 1999: 146. [↑](#footnote-ref-13)
14. See Pritchard 2005: 156; Sosa 2007: 26. [↑](#footnote-ref-14)
15. See Hiller and Neta 2007. [↑](#footnote-ref-15)
16. Pritchard 2009b suggests this version of safety when he says: “[A]ll we need to do is to talk of the doxastic result of the target belief-forming process, whatever that may be, and then not only focus solely on belief in the target proposition.” Williamson 2000 can be interpreted in a similar way. In contrast, Sosa addresses (S2)-safety only. [↑](#footnote-ref-16)
17. More specifically, a safety account of knowledge that uses (S4) might run into skeptical problems too easily. Consider a case in which you believe that there is a cup on the table in front of you on the basis of your current vision. If someone might easily have left a fake spoon on the table, you would, according to (S4), not know that there is a cup. Unfortunately, I do not have enough space to deal with this problem here. [↑](#footnote-ref-17)
18. This is an abridged version of Lackey 2006: 288. [↑](#footnote-ref-18)
19. Compare Coffman 2010: 246. [↑](#footnote-ref-19)
20. The operative method is italicized. [↑](#footnote-ref-20)
21. Compare Goldberg 2015:277–78. I present an abridged version of the original case. [↑](#footnote-ref-21)
22. See Goldberg 2015: 278: “I submit that the intuitively correct description of S is this: S’s true justified belief is formed through a safe method, yet the belief still fails to be knowledge (owing to the presence of epistemic luck).” [↑](#footnote-ref-22)
23. This distinction is made by Duncan Pritchard 2015: 105. It resonates with Hetherington’s earlier distinction between helpful and dangerous Gettier cases. Compare Hetherington 2006: 85–89. [↑](#footnote-ref-23)
24. See Horvath and Wiegmann 2016 for a recent experimental study that suggests that although the textbook view is that Henry does not know, the majority of real expert epistemologists does not share this view. [↑](#footnote-ref-24)
25. This is an abridged version of the original case in Neta and Rohrbaugh 2014: 399–400. [↑](#footnote-ref-25)
26. See Neta and Rohrbaugh 2014: 400; Sosa 2009: 207 shares this view. [↑](#footnote-ref-26)
27. Ibid.: 401. [↑](#footnote-ref-27)
28. See also Pritchard 2015: 104. [↑](#footnote-ref-28)
29. The case is a slightly abridged version of Kelp 2009: 27–28. [↑](#footnote-ref-29)
30. Bogradus 2012 presents a variant of this case with his Atomic Clock. When in this case Smith relies on the atomic clock, there is the persistent counterfactual risk that a radioactive isotope in the vicinity of the clock will interfere with its proper functioning. Bogradus argues that Kelp’s clock case does not establish a proper example of an unsafe belief since in the actual situation in which Russell has already decided to come down the stairs at 8:22 am, there are no longer nearby worlds in which the demon intervenes and Russell’s belief is false. In contrast, Bogradus’ own case is supposed to give an example of the continuing risk of forming a false belief. However, I do not think that the difference between the two cases has any epistemological significance. In particular, if one takes (S4)-safety to be the relevant notion here, we should not keep Russell’s decision to come down at 8.22 am fixed across nearby possible worlds. The only thing that one should keep fixed is the operative method. [↑](#footnote-ref-30)
31. This is a slightly revised version of the original Comesana case. See Comesana 2005: 397. [↑](#footnote-ref-31)
32. Kelp articulates the worry that the possible worlds in which Juan acquires a false belief on the basis of Judy’s misleading testimony might not be similar enough to the actual world to undermine safety (Kelp 2009:25). After all, these worlds are different from the actual world in many aspects: Juan decides to dress up as Michael; Judy believes that he is Michael; she calls Andy; the party is moved to Adam’s house. However, the distance between the worlds should not simply be measured by the number of differing facts. Otherwise, the consequent of the counterfactual conditional *If Nixon had pressed the button, a nuclear inferno would have been the result* could never be true in the next world in which Nixon presses the button. In fact, all the differences in worlds in which Juan dresses up as Michael are causally triggered by his disguise. The otherwise behaviorally relevant dispositions are already present in the actual world. So it is simply not true that in order to effect Juan’s false belief many facts that are independent of each other have to be different. In conclusion, I do not think that Kelp’s worry here is substantial. [↑](#footnote-ref-32)
33. The original case is from Sosa 2007: 31. Strictly speaking, Sosa claims that Kyle possesses animal knowledge rather than reflective knowledge. See Sosa 2007: 96, n.1. [↑](#footnote-ref-33)
34. If one wants to avoid method talk, one might also use the label ‘the basis of belief’ or ‘the way a belief is formed.’ [↑](#footnote-ref-34)
35. [footnote deleted] [↑](#footnote-ref-35)
36. Sosa seems to think, along these lines, that there is no relevant difference between the Kaleidoscope Case and the Fake Barn Case. See Sosa 2007: 96, n. 1. [↑](#footnote-ref-36)
37. See Broncano-Berrocal 2014 for another attempt to defend safety by individuating methods in an external way. However, Broncano-Berrocal’s sufficient account of method individuation runs into severe problems, as Bogradus and Marxen 2014 argue. Sosa 2007: 27 mentions, but does not develop, the idea of external methods. [↑](#footnote-ref-37)
38. Compare Brancano-Berrocal 2014: 73. [↑](#footnote-ref-38)
39. For the following see Bradley 2015: 205. [↑](#footnote-ref-39)
40. Interestingly, the defeater in this case is a defeater for knowledge without being a defeater for justified belief. This is so because the (mis)information that Lottie is participating in a fair lottery suggests that her target belief (that the ticket will lose) might easily be false, but it does not suggest that her belief is unreliable. [↑](#footnote-ref-40)
41. With some clarifications from Pritchard 2009: 49. [↑](#footnote-ref-41)
42. Pritchard 2009: 49. [↑](#footnote-ref-42)
43. Pritchard 2009: 50. [↑](#footnote-ref-43)
44. Pritchard 2009: 50. [↑](#footnote-ref-44)
45. Pritchard 2009: 49. [↑](#footnote-ref-45)
46. This is what Sosa requires for apt beliefs, namely that they are accurate (i.e. true) *because* they are adroit (i.e. are based on competence). [↑](#footnote-ref-46)