PRINCIPIA **22**(2): 301–319 (2018) doi: 10.5007/1808-1711.2018v22n2p301
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

DO SAFETY FAILURES PRECLUDE KNOWLEDGE?

J. R. FETT
Pontifical Catholic University of Rio Grande do Sul [PUCRS], BRAZIL
jrfett01@gmail.com

Abstract. The safety condition on knowledge, in the spirit of anti-luck epistemology, has become one of the most popular approaches to the Gettier problem. In the first part of this essay, I intend to show one of the reasons the anti-luck epistemologist presents for thinking that the safety theory, and not the sensitivity theory, offers the proper anti-luck condition on knowledge. In the second part of this essay, I intend to show that the anti-luck epistemologist does not succeed, because the safety theory fails to capture a necessary requirement for the possession of knowledge. I will attack safety on two fronts. First, I will raise doubts about whether there is any principled safety condition capable of handling a kind of case, involving inductive knowledge, that it was designed to handle. Second, I will consider two cases in which the safety condition is not met but the protagonist seems to have knowledge nonetheless, and I will vindicate my intuitions for thinking that those are in fact cases of knowledge by contrasting them with traditional, well-known Gettier cases. I want to conclude, finally, that safety failures do not necessarily prevent one from acquiring knowledge.

Keywords: Gettier • safety • sensitivity • knowledge • anti-luck epistemology

RECEIVED: 06/10/2017 REVISED: 05/01/2018 ACCEPTED: 27/03/2018

Since Plato, we have been told that a mere true belief cannot be known if it happens to be true just by chance, because the luck in play is a knowledge-preventing element. And we have also been told, since Peter Unger, at least, that a correct analysis of knowledge should include a clause capable of ruling out the lucky achievement of a true belief. Unger's proposal suffers from many flaws, but his central idea has inspired epistemologists to find a better alternative. Among these epistemologists, we find Duncan Pritchard. He developed a theory he baptized anti-luck epistemology, built on the premise that knowledge excludes luck and that we can solve the Gettier problem by finding the correct anti-luck condition on knowledge. The anti-luck epistemologist holds that the intuition that knowledge excludes luck should be somehow turned into a necessary condition in our analysis of knowledge. Since the anti-luck epistemologist typically understands luck as a modal notion, he thinks the best way to offer an anti-luck condition on knowledge is to offer a modal condition on knowledge, such as the sensitivity condition or the safety condition. In the first part of this essay, I intend to show one of the reasons the anti-luck epistemologist presents for thinking that

the safety theory, and not the sensitivity theory, offers the proper anti-luck condition on knowledge. In the second part of this essay, I intend to show that the anti-luck epistemologist does not succeed, because the safety theory fails to capture a necessary requirement for the possession of knowledge. I will attack safety on two fronts. First, I will raise doubts about whether there is any principled safety condition capable of handling a kind of case, involving inductive knowledge, that it was designed to handle. Second, I will consider two cases in which the safety condition is not met but the protagonist seems to have knowledge nonetheless, and I will vindicate my intuitions for thinking that those are in fact cases of knowledge by contrasting them with traditional, well-known Gettier cases. I want to conclude, finally, that safety failures do not necessarily prevent one from acquiring knowledge.

1. Gettierization and epistemic luck

Let us imagine the following scenario — we will call it *Stopped Clock*. Wondering what time it is, *S* checks her analogical clock, which reads 5p.m. *S* has an accurate perception and has some beliefs about the reliability of the clock. Thus, *S* forms a justified belief that it is 5p.m. Surprisingly, though, the clock is not working; it broke during the night. But the clock stopped exactly twenty-four hours ago, so that it is *in fact* 5p.m, as it shows to *S*. She has a justified true belief that does not seem to count as knowledge. This case — which had been put forward by Bertrand Russell (1948, p.140) before Edmund Gettier's seminal paper was published (though with a slightly different narrative) — goes against a purportedly millennial view of the concept of knowledge, harking back to Plato's Meno and Theaetetus, according to which knowledge is *justified true belief*.[1]

Now let us consider a different Gettier-type case that has been called *Fake Barns*.

> *S* is in Phony Barn Country, but she doesn't suspect it. All of the barns in the vicinity [...] are phony except the barn *S* is looking at. Given that the real barn and the phony ones are indistinguishable from *S*'s perspective and that S has no grounds for thinking something is amiss, she comes to believe truly and justifiably that there is a barn. (Klein, 2012, p.159)[2]

Here, again, *S* has a justified true belief about the presence of a barn in front of her that does not seem to count as knowledge, given the inhospitable circumstances in which she forms her belief. Another blow to the justified true belief view of knowledge.

The most popular diagnosis people offer when faced with Gettier cases suggests that the gettierized belief falls short of knowledge because *it is just a matter of luck that it is true*. In *Stopped Clock*, *S* looks at the clock during the only time of the day in which it is correct. She could very easily end up with a false belief, given the

misleading evidence she acquires by reading the clock. And an analogous explanation seems adequate concerning the *Fake Barns* case. There, it is a coincidence that *S*'s belief about the presence of a barn is true in that hostile environment, since she could very easily have looked at a phony barn that she would not be able to distinguish from a genuine barn.

In a desire to do justice to these common reactions I just expressed about what is wrong with the Gettier cases, *anti-luck epistemology* was born. It had Peter Unger (1968) as its precursor, but its success as an epistemological methodology is due to Duncan Pritchard's work.[3] According to Pritchard, the definitive solution to the Gettier problem, and to some other problems epistemologists face, lies in a theory of knowledge built on the concept of *epistemic luck*.

Anti-luck epistemology suggests that gettierization involves what has been called *veritic epistemic luck*.[4] This kind of luck is supposed to reveal the disconnection between the gettierized subject's justified belief and what makes it true. As we noticed above, both the misleading evidence S acquires by checking the stopped clock and the abnormal environment of the *Fake Barns* scenario S finds herself in make it likely that she will form a false belief. In this sense, it is a matter of (*veritic*) luck that she formed a true belief instead.

Pritchard thinks that the best way of fleshing out this anti-luck intuition I just exposed is through a *modal condition* on knowledge. The literature furnishes us with two major alternatives, namely, the *sensitivity theory* and the *safety theory*. In the remainder of this essay I will, firstly, examine the sensitivity theory and a strong objection to it. Secondly, I will focus on the safety theory and will put pressure on it by (i) raising doubts about whether there is any principled version of the safety condition capable of handling an important case involving inductive knowledge that the theory was designed to handle, and (ii) by considering two cases of unsafe belief that, I will argue, are in fact cases of knowledge, thus showing that safety failures do not necessarily prevent one from acquiring knowledge.

2. Sensitivity

Among the many attempts to solve the Gettier problem, we find the so-called *tracking theory* advanced by Robert Nozick (1981). Nozick's central idea, which was probably influenced by Armstrong (1973) and by Dretske's (1971) epistemology, was that knowers are like thermometers in that their output, so to speak, is well-connected with reality, which knowers express through their beliefs and thermometers express through their readings. According to Nozick's epistemology, if a subject *S* knows that it is raining outside, for instance, then if it continued to rain and she attended to that event, she would believe it is raining outside; and if it were not raining any longer,

she would not believe that it was raining outside. In this sense, knowers are supposed to *track the truth* of the propositions they know in a very similar way to the way that thermometers track the temperature in their environments.[5]

Nozick's complete analysis of knowledge can be stated as follows:
S knows that *P* if and only if

(i) *P*

(ii) *S* believes, via method or way of coming to believe *M*, that *P*.

(iii) If *P* weren't true and *S* were to use *M* to arrive at a belief whether (or not) *P*, then *S* wouldn't believe, via *M*, that *P*.

(iv) If *P* were true and *S* were to use *M* to arrive at a belief whether (or not) *P*, then *S* would believe, via *M*, that *P*. (Nozick 1981, p.179)

Conditions (iii) and (iv) are the centerpieces in his tracking theory, and they have been called *sensitivity condition* and *adherence condition* respectively.

Despite there being a very interesting debate in the literature regarding the adherence condition,[6] we will focus here solely on the sensitivity condition — which is largely considered to be the core idea in Nozick's epistemology. The sensitivity condition carries much intuitive appeal. It requires that, in order for you to know that *P*, you must be sensitive to changes in the truth-value of *P* in a specific range of counterfactual situations; if *P* were not true, you would not believe that *P*. Take the *Stopped Clock* case. It is pretty clear that *S* is not sensitive to the possible variations with respect to the truth-value of her target-belief in slightly different situations. If it were false that it is 5pm, she would believe that it is 5pm anyway, since she would continue to be informed by a stopped clock that reads the same thing any moment of the day.

As you may have already noticed, the idea behind the sensitivity condition is expressed through a counterfactual conditional which states a *modal claim*. Usually, we understand that claim by making use of the possible worlds heuristics advanced by Lewis (1973) and Stalnaker (1968). We just need to know a brief story about the notion of a possible world and a story about the similarity criterion for measuring the distance among possible worlds.[7] Roughly put, possible worlds are ways a world might be, as Stalnaker's slogan suggests. One of these ways a world might be is the way our world *actually* is: that is exactly what is called *the actual world*. Possible worlds, except the actual one, can be understood through the description of counterfactual situations (Greco, 2012). For instance, in the actual world I am wearing a blue T-shirt and I am Brazilian, but I might be wearing a red T-shirt and I might be Argentine. In fact, there is a possible world, which is not the actual world, where I am wearing a red T-shirt and I am Argentine; this counterfactual situation describes that possible world. Possible worlds are ordered by a *similarity relation* to the actual

world. The more different from the actual world a possible world is, the more distant from the actual world it is; conversely, the more similar to the actual world a possible world is, the closer to the actual world it is. Equipped with these notions, sensitivity theory suggests that *S* knows that *P* only if in the closest worlds (that is, in the worlds that are more similar to the actual world) in which *P* is false, *S* does not believe that *P*. We have here a recipe for a sensitivity test. First, we look at the possible world which is closest to the actual world but in which *P* is false. Then we examine the most likely doxastic behavior of S in that world concerning the target-proposition *P*. If in that world *S* would believe that *P* (via the same belief-forming method she employed in the actual world), then she is not sensitive to the falsehood of *P* and, hence, does not have knowledge in the actual world.

Let us test the sensitivity condition with the *Fake Barns* case. Imagine the closest possible world where it is false that there is a barn in front of S. This is an easily conceivable world, since it is exactly any of the worlds in which *S* looks at one of the many fake barns surrounding her. In this world where she is looking at a fake barn, would *S* not believe that what she sees is a barn? It is reasonable to think she would, for she would have the same barn-percepts she has in the actual world. Her belief is insensitive and therefore, Nozick argues, does not amount to knowledge. Gettierization detected and eliminated, and the same verdict generalizes to every Gettier case.

Notwithstanding its intuitive appeal and ability to handle a vast range of Gettier cases, the sensitivity principle has been accused of a number of shortcomings. The strongest blow against it comes from an objection that Ernest Sosa (1999a) — inspired by Jonathan Vogel (1987) — put forward in order to show that the principle is too strong in incorrectly excluding cases of inductive knowledge. This is Sosa's famous counterexample:

> [CHUTE]: On my way to the elevator I release a trash bag down the chute from my high rise condo. Presumably I know my bag will soon be in the basement. But what if, having been released, it still (incredibly) were not to arrive there? That presumably would be because it had been snagged somehow in the chute on the way down (an incredibly rare occurrence), or some such happenstance. But none such could affect my predictive belief as I release it, so I would still predict that the bag would soon arrive in the basement. My belief seems not to be sensitive, therefore, but constitutes knowledge anyhow, and can correctly be said to do so. (Sosa 1999a, pp.145–6)

In fact, if we look at the closest world where the bag does not get to the basement, you will see that I would still have believed that it did, for I would still have a good inductive basis for believing that. How can a clear case of knowledge be counted as a case of ignorance? The verdict the sensitivity theory delivers here strikes us as very

implausible. While inductive basis may be fragile enough to be properly condemned by the sensitivity condition, it is still the basis of a large amount of the knowledge common sense holds we have. Are we disposed to give up the strong intuition that *Chute* and other cases of induction can be instances of knowledge?

Things get worse, the objectors say, when we notice that the objection Sosa advanced is generalized through the observation that the sensitivity condition suggests the wrong procedure to detect ignorance. It recommends, according to them, that we look at what happens in the closest not-*P* world; *but the closest not-P world may still be a very distant world*. In this respect, Pritchard, for instance, highlights that while we look at a distant not-P world, we may fail to take into account what happens in a vast number of *P*-worlds in the neighborhood of the actual world (that is, the possible worlds that are most similar to the actual world) — see Pritchard (2014, p.156). Thus, in *Chute*, it seems to be true that in the closest world in which the bag did not arrive in the basement, *S* would believe falsely that it did, but the neighborhood of the actual world is full of close worlds in which the bag does arrive in the basement and *S*'s doxastic attitudes regarding this proposition would turn out to be correct.

Detractors of the sensitivity condition argue that it wrongly takes *modal sensitivity in the closest not-P possible world, no matter how far it is from the actual world*, to be the master intuition about knowledge. The correct intuition to be captured, they say, is that *we are highly intolerant of error in those worlds that are close to the actual world*, even though we are tolerant of it in those distant worlds in which very strange things happen — see Pritchard (2014, p.157).

Based on the considerations we have made about the sensitivity condition's being the wrong procedure to detect ignorance and about what the right modal intuition about knowledge is, it seems that the right test to check the possession of knowledge should be formulated as follows: In the close possible worlds in which *S* believes that *P* (in the same way as she does in the actual world), does *S* believe *truly* that *P*? The epistemology Sosa proposes in place of the sensitivity principle delivers exactly this test. It corresponds to the *safety principle* and it is the topic of the next section.

3. Safety

The safety principle, though it dates back to Luper (1984) at least, was made famous by Ernest Sosa, who put forward the following condition on knowledge:

> Call a belief by S that p "safe" iff: *S* would believe that *p* only if it were so that *p*. (Alternatively, a belief by *S* that *p* is "safe" iff: *S* would not believe that *p* without it being the case that *p*; or, better, 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*.) Safety: In order to (be said correctly

> to) constitute knowledge a belief must be safe (rather than sensitive). (Sosa 1999a, p.142)

When a subject S has a safe belief that P, she would not easily be mistaken as to whether P. The safety principle is a modal claim, like the sensitivity condition, so it is useful to avail ourselves of the possible worlds heuristics for subjunctive conditionals once again and, as we did with the sensitivity condition, we should relativize safety to the basis on which the subject forms her belief in the actual world.[8] In order for us to see whether or not a certain belief P of S is safe, we need to look at close possible worlds in which S believes that P in the same way as in the actual world and then we need to find out whether she believes *truly* that P in these worlds.

Take the *Stopped Clock* case in order to test the safety principle. Consider one of the many close worlds where S believes that it is 5p.m. by looking at the stopped clock. Consider, for instance, the world in which everything stays the same except for the fact that it is 4p.m. In this world, would S's belief about the time have been true? Clearly not, after all S would have looked at the stopped clock and, as a result, S would have believed that it is 5p.m. S's belief is unsafe, thus not amounting to knowledge.

Now let us see whether safety fares better than sensitivity when it comes to the troublesome *Chute* case. Let us call the actual world '@' and the relevant possible world 'w_1'. In *Chute*, though S's target-belief is true in @, it is false in w_1, because in w_1 the bag does not get to the basement. Nonetheless, the modal stability of S's target-belief is not threatened by this failure, since w_1 is not within the range of close possible worlds, the worlds that matter to assess the safety of S's belief. S may fail to believe truly in distant worlds, as long as she believes truly in all the close ones. Sosa would have us believe that the safety principle delivers exactly this result when it comes to this case, since a possible world in which something keeps the bag from getting to the basement is very distant from the actual world, because the obstruction of the chute is a very rare, and hence extremely unlikely event.

So far so good. But it is time for me to present my first main concern with the safety principle. Perhaps a closer look at *Chute* will unveil a problem. Contrary to what Sosa suggests, it seems very plausible to say that a world in which something happens in such a way that the bag does not get to the basement as a result — say, a piece of the tin wall comes loose — *is still very close to the actual world*, given the similarity criterion. Few things would need to change in order to turn @ into w_1; w_1 is very similar to @, therefore it is a close world relative to @. Notice, however, that the description of the case tells us that the situation in which something obstructs the passage of the bag is a *very rare situation*. This observation surely means that it is *very unlikely* that something obstructs the chute — in other words, the probability that something obstructs the passage of the bag through the chute is very low. But

this does not undermine our claim that this is a very close situation, *modally speaking*, since there seem to be persuasive reasons offered in the literature for thinking that probability and modal closeness come apart — against a more popular and perhaps entrenched view that suggests they stick together.

We will briefly examine one argument supporting this apparently heterodox though highly intuitive view. By using the well-known scene of the lottery case, Duncan Pritchard helps us understand how low probability events can still be modally close to the actual world:

> What the lottery case reminds us is that an event can be modally close even when probabilistically unlikely. [...] The possible world in which one is leaping about with joy in one's room because one is a lottery winner is very alike to the possible world in which one is tearing one's ticket up in disgust — all that needs to change is that a few coloured balls fall in a slightly different configuration. [...] The moral is that modal closeness comes apart from probabilistic closeness. [...] Indeed, this is why people play lotteries, and yet do not place bets on modally far-fetched events with similarly massive odds. [...] An awful lot would need to change about the actual world to make it such that I am an Olympic sprint champion. (Pritchard 2016b, pp.552–3)

This seemingly plausible idea that probability and modal closeness do not stick together is a motivation to be skeptical of the success that, according to Sosa, safety allegedly has over sensitivity when it comes to difficult cases like *Chute*. Although it is very unlikely that something will prevent the bag from getting to the basement, it seems that few things would need to be different from the way they actually are in order to make that happen, which is to say that there may well be a close world where the bag is not in the basement and the subject falsely believes that it is.

Thinking about what features the relevant possible world should have in order for the target-belief in *Chute* to be unsafe, Pritchard writes:

> If significant change in the actual circumstances is required to ensure that the rubbish fails to reach the basement (significant change, moreover, which is undetectable to [the subject]) then his belief will be safe, since it could not easily have been false. (Pritchard 2012, p.255)

Now think about how much needs to change so that the bag fails to arrive in the basement. It suffices that one piece of the tin wall comes loose. Is this a significant change in the actual circumstances? Not at all. And what about a world where a fat cat enters into the chute and gets stuck in it minutes before S puts her bag there, so that the bag gets stuck at the point where the cat lies, quiet and unable to move? Would this sad event involve a significant change in the actual circumstances? Again, not at all. We just need a world where some neighbor has a clumsy, fat cat walking around the chute. A world in which Incredible Huck punches the chute at the moment

the bag is being put into it requires dramatic changes relative to the actual world, not a world in which a simple piece of the tin wall comes loose or a cat gets stuck in the chute and obstructs the bag.

It is time to ask this question: Are not we reading the safety principle in too strong a way? Is the safety of a given true belief inconsistent with the presence of a false belief (formed in the same way as in the actual world) in the class of the close possible worlds? Should we completely exclude the possibility of acquiring a false belief in a close possible world? Indeed, usually the safety principle is put in absolute terms, with no exception. However, there is a version of the principle that seems to accommodate that possibility, which is the topic of the next section.

4. Weak safety to the rescue?

As we have seen, the safety principle has a hard time with inductive knowledge, at least as far as its usual, strong construal is concerned. Safety advocates, faced with such difficulty, have proposed a refurbished, weakened version of the principle, commonly called *weak safety*, which allows one to form a false belief in a minor portion of close worlds while maintaining a safe belief in the actual world. Duncan Pritchard's version of weak safety is by far the most popular among similar proposals, so let us take a look at it:

> For all agents, p, if an agent knows a contingent proposition p, then, in nearly all (if not all) nearby possible worlds in which she forms her belief about p in the same way as she forms her belief in the actual world, that agent only believes that p when p is true. (Pritchard 2005, p.163)

The initial appeal of this strategy is quite clear. By allowing one to form a false belief in some close worlds — as long as one does so *only* in a minor portion of them! — weak safety manages to accommodate the characteristic fallibility of inductive knowledge. I may come to know that my trash bag will be soon in the basement even though such knowledge is fallible, since once in a while trash bags get stuck in the chute, and so there is a possibility of error nearby. What provides my inductive true belief with the modal stability needed for knowledge, the weak safety proponent argues, is that in most of the close worlds my trash bag *does* get to the basement.

Another important advantage promised by Pritchard when advertising weak safety is that it handles the lottery problem, namely the problem concerning the fact that I have massive evidential (probabilistic) support to believe that my lottery ticket is a loser — after all there are one million tickets out there, so that my chances are one in a million — yet it seems that I cannot know that it is a loser nonetheless. But wouldn't the lottery scenario be troublesome for weak safety?[9] For weak safety says that your belief is safe if in most close worlds where you believe as you did in the

actual world your belief is true, and the belief that my lottery ticket lost meets this requirement. It is the case that in most close worlds where I believe that my lottery ticket lost based on the probabilities of winning the lottery, I end up with a true belief, so that the belief that my ticket lost is safe and thus could be qualified as knowledge.

Notice that the challenge of accommodating inductive knowledge forces safety to be weaker, whereas the lottery problem forces safety to be strong. As we just saw, the belief that my ticket lost meets a weakened safety principle, as that one proposed by Pritchard, but many take this as a pretty undesired result, since there is a widely shared intuition according to which no one can know that her ticket lost based only on its probabilities of winning the lottery.[10] This shared intuition receives support from the assumption that there is a close world where I do win the lottery and so the belief that my ticket lost is false, and if there is such a world, then knowledge of this belief, formed as it was formed, is precluded. The only way to avoid my lottery belief being counted as safe is to require strong safety for knowledge, under which no false belief in close worlds is tolerated.

It is precisely to accommodate these two demands that Duncan Pritchard suggests what I will call the *risk-weighted construal of safety*. He wants us to look at the principle we called weak safety above under a different perspective, one that highlights the risk in play when assessing the safety of a belief. The main idea of his proposal appears in the following excerpt:

> The crux of the matter is that when it comes to evaluations of whether an event is lucky, not all the near-by possible worlds are on a par. Instead, those near-by possible worlds that are closer to the actual world — i.e., which are most similar to the actual world — carry more weight in our evaluations of whether an event is lucky than those near-by possible worlds which are less close. (Pritchard 2016a, p.32)

Pritchard is suggesting a continuum picture here in order for us to understand how safe a belief has to be so that it be counted as knowledge. According to this *continuum* picture, safety, other things being equal, is compatible with false beliefs in close possible worlds, as long as these worlds are not the closest ones. We call this risk-weighted safety because it assumes that modal stability tolerates risk within a certain range of close worlds, but does not tolerate it with respect to those worlds that are closest to the actual world. Thus, the risk of believing a falsehood must be weighed in regard to different possible worlds so that we can properly judge whether a belief is safe or not.

But problems lie ahead. Risk-weighted Safety exhibits a bipolar formulation, which sometimes puts the bar for modal stability too high, sometimes too low. The point is that we end up with different safety requirements for different classes of possible worlds. For the broader class of possible worlds that are close to the actual

world, weak safety is the proper requirement on knowledge, *except* for a subclass of these worlds, the one comprehending the closest possible worlds, for which strong safety is said to be the proper requirement on knowledge. That strikes me as a juxtaposition of two principles which carry with them very different assumptions about the nature of knowledge, as I will explain.[11]

So far as I can see, strong safety is what safety was originally meant to be, since Ernest Sosa (1999a; 1999b), at least. In the context of the skeptical challenge, safety arises as the substitute for sensitivity, allowing anti-skeptics to keep the closure principle and to get a Moorean resistance to skepticism off the ground. However, when it comes to modal stability, pretty much nothing changed. Both sensitivity and safety were concerned with robustness for knowledge. Fred Adams and Murray Clarke put this point in a pretty clear way with respect to tracking theories:

> ...[T]he tracking theories see knowledge as a real relation between a believer, the truth of the relevant belief, and the environmental conditions that nomically ensure the truth of the relevant belief (screening off mere luck that the belief is true). (Adams and Clarke 2005, p.208)

In this same spirit, safety was meant to capture this relationship of stability between belief and truth. Here is an expression of this *desideratum* by Ernest Sosa (1999b, p.374), describing a feature of his safety account: "My alternative to Nozickian tracking is "Cartesian" because it jibes with the combination of self-intimation and infallibility distinctive of Cartesian privileged access." And what could a Cartesian requirement of modal stability on knowledge require but infallibility in close worlds? The desire was to completely exclude the possibility of error in a close counterfactual situation

Remember that the dismissal of the sensibility condition in favor of the safety condition was due to the fact that people thought of sensitivity as a condition that makes us take into account far-off worlds, whereas what we should take into account was the truth-connection only in close worlds. And clearly the first safety advocates wanted not only *mere truth-connection in close worlds*, after all reliabilism had been around for a long time offering exactly that kind of robustness.[12] They wanted *infallibility in close worlds*, and only strong safety does justice to this desire.

Weak safety, in turn, mitigates the robustness of strong safety by being compatible with false beliefs in close worlds. As we have seen, such compatibility is a necessary concession if safety is to accommodate inductive knowledge. Yet, that room for error is exactly what the modal theories of knowledge have tried to avoid. Therefore, opting for weak safety frustrates the very motivation for opting for something like safety in the first place.

To sum up, this strategy of weakening the safety condition is not a good option for the safety advocate, since it frustrates the very idea of infallibility in close worlds

that safety was meant to capture. Nor will the risk-weighted construal of safety save it from bad results, the worst one being postulating distinct modal stability requirements for distinct classes of possible worlds. Now we turn to more acute safety failures.

5. More safety failures

Now we will turn to my second main concern with the safety principle. As it was said earlier, the safety principle was tailored to deal with the veritic epistemic luck phenomenon. Presumably when a belief is true only as a matter of luck, it is unsafe and, hence, not an instance of knowledge, and that is why safety is taken as being a necessary condition on knowledge. In this section, we challenge this result. We will consider cases of unsafe belief which, I will argue, we should count as cases of knowledge.

Let us begin with the following case, presented by Ram Neta and Guy Rohrbaugh as a counterexample to the safety principle:

> [WATER]: I am drinking a glass of water which I have just poured from the bottle. Standing next to me is a happy person who has just won the lottery. Had this person lost the lottery, she would have maliciously polluted my water with a tasteless, odorless, colorless toxin. But since she won the lottery, she does no such thing. Nonetheless, she almost lost the lottery. Now, I drink the pure, unadulterated water and judge, truly and knowingly, that I am drinking pure, unadulterated water. 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. . . . 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. (Neta and Rohrbaugh 2004, pp.399–400)

It is pretty clear why this is a case of unsafe belief. Although in the actual world I end up believing *truly* that I am drinking pure water, there is a very close possible world in which I falsely believe that to be the case. Look, for instance, at the world in which the crazy person standing next to me has lost the lottery and, as a result, has polluted my water. But now take a look again at what happens in the actual world. Her psychopathic tendencies notwithstanding, the crazy person, apprised of the lottery results, does not even get close to my glass of water in the actual circumstances, for her intentions have already disappeared. Should I be counted as ignorant of my target-belief that I drink pure water even though this belief stopped being threatened from the very moment the person came to know the lottery results? Some people think I should not.

Ram Neta and Guy Rohrbaugh (2004) believe this is case of knowledge, though it is, more precisely, a case of *unsafe knowledge*, in Juan Comesãna's words.[13] They suggest that the protagonist has a true belief that she is drinking pure water *because of the exercise of her cognitive abilities* — in this case, her competent perception. When a success is attributable to the exercise of one's abilities, it counts as an achievement, and knowledge itself is a kind of *achievement* — an intellectual achievement. Neta and Rohrbaugh also suggest that achievements are compatible with the risk of not being successful, so that achievements can be unsafe, according to safety theories. Therefore, there is no problem if in *Water* the protagonist does not met the safety requirement, since that does not prevent her from achieving knowledge. In short, knowledge is an achievement and achievements allow for risk, and since *Water* seems to be a case of knowledge, safety cannot be a necessary condition on knowledge, they argue.

Duncan Pritchard (2015) disagrees with them. Pritchard thinks that *Water* is a Gettier case, since the protagonist acquires a true belief only by a stroke of luck — her belief could have easily been false. And what is more, Pritchard offers an error theory that, he says, explains the intuitions of those who disagree with him. He argues that the kind of epistemic luck that is present in *Water* is compatible with the true belief's being a cognitive achievement (that is, a cognitive success that is due to the subject's cognitive abilities). According to Pritchard, those of us who have the intuition that the protagonist in this case has knowledge are swayed by the fact that she exhibits a cognitive achievement and, as a result, *her belief is much more epistemically qualified than the typical gettierized beliefs* we are familiar with — like S's belief that it is 5p.m. in *Stopped Clock*.

Most of the Gettier cases involve a kind of luck that cancels out the effect of the bad conditions for acquiring knowledge in which the gettierized subject finds herself; that kind of luck cancels out the hazard of the actual circumstances in which she forms her belief. Different gettierized subjects find themselves in different kinds of bad conditions, and their predicament may be exposed through various explanations. We usually say, for instance, that the original cases Gettier (1963) presented feature subjects that are under the influence of misleading evidence (testimonial and memory evidence in those cases) from which they derive a false belief, whereas some non-inferential Gettier cases' most distinctive feature seems to be that the protagonist's belief is somehow disconnected from what makes it true, as we see in *Stopped Clock*. Other cases still are best explained in terms of the unreliability of the belief-forming process which produced the target-belief, or in terms of the hostile environment that furnishes the protagonist with an inaccurate representation of the reality most of the time, as in *Fake Barns*. These are all bad luck elements. Gettierized subjects get to the truth despite the imminent danger in their epistemic conditions, for given those conditions, they would surely end up with a false belief, were it not for the

"good" luck that intervenes. Trying to capture this dynamic, Linda Zagzebski (1994) offered an account of Gettierization in line with the features of the typical Gettier cases we exposed above, which I will call the *double luck view of gettierization*. It says essentially that the structure of the Gettier cases involves some element of good luck counteracting some element of bad luck.

However, the purportedly knowledge-preventing epistemic luck Pritchard says that is present in *Water* is not the well-know veritic epistemic luck, but rather what he baptized *purely modal veritic epistemic luck* (Cf. Pritchard 2015, pp.104–5). This kind of luck has nothing to do with the actual circumstances the subject is in, but has to do only with what happens with her target-belief in the close possible worlds. Purely modal veritic epistemic luck does not cancel out the effect of misleading evidence, disconnection between belief and truth, or hostile environment, rather it indicates that the problem lies in the modal fragility of the true belief that was formed in the actual world. Despite the differences between this kind of luck and the typical veritic epistemic luck, their effect on the subject's target-belief is the same: they make it unsafe and, as a result, a case of ignorance.

We want to investigate why, for at least some of us, it is far from clear that *Water* is not a case of knowledge, though it is quite clear that it is a case of unsafe belief. Let us begin by noticing that in the typical Gettier cases the scenario is set up in such a way as to prevent the protagonist from forming a belief that is true. In the *Stopped Clock* case, for example, *S* is looking at a clock that is not functioning, so that in those conditions *S* would, almost certainly, form a false belief about the time. In *Fake Barns*, *S* is in a region that is full of fake barns she cannot distinguish from genuine barns, so that in those conditions she is not poised to form a true belief. But in Water, the subject *does* seem to be poised to form a true belief and, thus, to acquire knowledge. Nothing happens with her water, there are no bottles around with polluted water, nobody has the intention of polluting the subject's water any more.

The point here is that while in the typical Gettier cases the standard conditions are unfavorable for forming a true belief, in cases of purely modal veritic epistemic luck, like *Water*, the subject has not encountered any obstacle preventing her from forming a true belief. On the contrary, in *Water* she drinks pure water which nobody has touched in an environment with no polluted water nearby.

To make my point more appealing, I will ask the reader to consider a variation on the *Fake Barns* case, inspired by some remarks made by Pritchard (2015, p.105) and quite similar to a case Neta and Rohrbaugh (2004, p.401) first put forward in the literature. Let us call it *No Fake Barns*. Imagine that the carpenter responsible for assembling and placing the fake barns every morning in Phony Barn Country was about to begin his job — namely, lining up ten barn façades around only one genuine barn — when he received a call from his wife who then told him that the lottery ticket he bought yesterday was a winner. The carpenter, extremely euphoric,

got immediately into his car without even beginning his job, leaving the region with only the genuine barn and no barn façade. During that day, every tourist who drove through the region and formed a belief about the presence of a barn *did form a true belief*.

We just turned the original *Fake Barns* case into a case of purely modal veritic epistemic luck. In this version of that case there is nothing epistemically damaging in *S*'s actual circumstances, all the bad events obtaining only in close possible words. Now try to put yourself in the tourists' shoes for a moment and think about their situation and their position to acquire knowledge concerning the presence of a barn. Did not they know that they saw a barn? Bear in mind the fact that nothing has put them into an unfavorable situation for acquiring knowledge this time — no misleading evidence, no belief-truth disconnection, no hostile environment. During all that day, all barns the tourists had to look at were genuine barns, with no indistinguishable façades to worry about.

Perhaps we can weaken the shades a bit more and make the line between gettierization and knowledge clearer. Consider the differences between *Water* and the following variation of that case — *Water II*. The case is the same as Water except that this time the person standing next to me has lost the lottery and then has polluted my water with the toxin he had in his bag. However, by virtue of being improperly stored, it did not poison my water. While drinking the water, I formed a true belief that I was drinking pure water (that is, water without any active toxin).

What we just saw is a Gettier case much more in line with the double-luck view of gettierization than the original *Water*. In *Water II*, we find a subject who is very likely to form a false belief, for she is drinking from a glass that was manipulated by the crazy man next to her, but who forms a true belief, as it turns out, given the interference of a dosage of good luck. And the same is true concerning the *No Fake Barns* case. To sum up, in Water, the luck in play is the purely modal veritic epistemic luck, whereas in *Water II* the luck in play is just the old, well-known veritic epistemic luck. The latter kind of luck leads us to the truth despite the damaging conditions in the scenario. The former kind of luck simply puts us in a world where we are in a position to acquire knowledge and thrive as cognizers. The critical point is that even in those cases where knowledge is not threatened, the safety condition is not met, which leads us to think that safety is definitely not a necessary condition on knowledge. Thus, for all that has been said so far, I think it is reasonable to conclude that (i) safety is not a necessary condition of knowledge, since in cases of unsafe belief like *Water* and *No Fake Barns* the protagonists seem to be in conditions that are good enough to acquire knowledge, and the relevant intuitions vindicate the claim that (ii) the double luck structure suggested by Zagzebski — good luck counteracting bad luck — is a necessary feature of Gettierization.

6. Concluding remarks

I have tried to provide the reader with reasons for believing each of the following claims. Although the sensitivity theory appears to be vulnerable to the objection that it is too strong in incorrectly excluding cases of inductive knowledge, the safety theory, aimed at replacing it as the correct modal condition on knowledge, does not fare much better on this front. I have raised doubts about whether there was any principled safety condition capable of accommodating the possibility of inductive knowledge. More importantly, I argued that safety, in its best formulation, is not a necessary condition on knowledge. I considered two cases that, so far as I can see, support that claim. These cases seem to have also revealed that the double-luck structure of the typical Gettier cases is, in fact, an essential feature of gettierization. Thus, I answer negatively the question I have proposed as the title of this essay: safety failures do not necessarily preclude one from acquiring knowledge.

References

Adams, F.; Clarke, M. 2005. Resurrecting the tracking theories. *Australasian Journal of Philosophy*, **83**(2): 207–21.

Armstrong, D. 1973. *Belief, Truth and Knowledge*. London: Cambridge University Press.

Bogardus, T. 2014. Knowledge Under Threat. *Philosophy and Phenomenological Research* **88**(2): 289–313.

Comesaña, J. 2005. Unsafe knowledge. *Synthese* **146**(3): 395–404.

Dretske, F. 1971. Conclusive reasons. *Australasian Journal of Philosophy* **49**(1): 1–22.

Engel, M. 1992. Is epistemic luck compatible with knowledge? *Southern Journal of Philosophy* **30**(2): 59–75.

Feldman, R. 2003. *Epistemology*. Upper Saddle River, NJ: Prentice Hall.

Gettier, E. 1963. Is justified true belief knowledge? *Analysis* **23**: 121–3.

Goldman, A. 1976. Discrimination and perceptual knowledge. *Journal of Philosophy* **73**: 771–91.

———. 1979. What is Justified Belief? In: G. Pappas (ed.) *Justification and Knowledge*, pp.1–25. Boston: D. Reidel.

Greco, J. 2007. Worries about Pritchard's safety. *Synthese* **158**(3): 299–302.

———. 2010. *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge: Cambridge University Press.

———. 2012. Better safe than sensitive. In: K. Becker; T. Black (eds.) *The Sensitivity Principle in Epistemology*, pp.193–206. Cambridge: Cambridge University Press.

Klein, P. 2012. What Makes Knowledge the Most Highly Prized Form of True Belief. In: K. Becker; T. Black (eds.) *The Sensitivity Principle in Epistemology*, pp. 152–69. Cambridge: Cambridge University Press.

Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.

Luper, S. 1984. The epistemic predicament. *Australasian Journal of Philosophy* **62**: 26–50.

———. 2003. Indiscernibility skepticism. In: S. Luper (ed.) *The Skeptics: Contemporary Essays*, pp.183–202. Aldershot: Ashgate.
Neta, R.; Rohrbaugh, G. 2004. Luminosity and the safety of knowledge. *Pacific Philosophical Quarterly* **85**(4): 396–406.
Nozick, R. 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press.
Pritchard, D. 2005. *Epistemic Luck*. Oxford: Oxford University Press.
———. 2007. Anti-luck epistemology. *Synthese* **158**(3): 277–97.
———. 2012. Anti-Luck Virtue Epistemology. *Journal of Philosophy* **109**(3): 247–79.
———. 2014. Knowledge Cannot Be Lucky. In: M. Steup; J. Turri; E. Sosa (eds.) *Contemporary Debates in Epistemology*. 2nd ed. Oxford: Blackwell.
———. 2015. Anti-luck epistemology and the Gettier problem. *Philosophical Studies* **172**(1): 93–111.
———. 2016a. *Epistemology*. 2nd edition. New York: Palgrave Macmillan.
———. 2016b. Epistemic Risk. *Journal of Philosophy* **113**(11): 550–71.
Russell, B. 1948. *Human Knowledge: Its Scope and Limits*. New York: Allen and Unwin.
Shope, R. 1983. *The Analysis of Knowing: A Decade of Research*. Princeton, New Jersey: Princeton University Press.
Sosa, E. 1991. *Knowledge in Perspective*. Cambridge: Cambridge University Press.
———. 1999a. How to defeat opposition to Moore. *Philosophical Perspectives* **13**: 137–49.
———. 1999b. How Must Knowledge Be Modally Related to What Is Known? *Philosophical Topics* **26**(1/2): 373–84.
———. 2007. *A Virtue Epistemology: Volume I: Apt Belief and Reflective Knowledge*. Oxford: Oxford University Press.
Stalnaker, R. 1968. A Theory of Conditionals. In: N. Rescher (ed.) *Studies in Logical Theory*, pp.98–112. Oxford: Oxford University Press.
Unger, P. 1968. An analysis of factual knowledge. *Journal of Philosophy* **65**(6): 157–70.
Vogel, J. 1987. Tracking, closure, and inductive knowledge. In: S. Luper (ed.) *The Possibility of Knowledge: Nozick and His Critics*, pp.197–215. Rowman & Littlefield.
Wedgwood, R. 2014. *In defense of adherence*. http://certaindoubts.com/in-defence-of-adherence/. Access: 18/06/2018.
Zagzebski, L. 1994. The inescapability of Gettier problems. *Philosophical Quarterly* **44**(174): 65–73.
———. 1996. *Virtues of the Mind: An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press.
———. 2009. *On Epistemology*. Belmont, CA: Wadsworth.

Notes

[1]See Shope (1983) for historical remarks in this respect.

[2]This is Peter Klein's version of the original *Fake Barns* case presented by Carl Ginet to Alvin Goldman, first published in Goldman (1976).

[3]See Pritchard (2005), (2007), and (2012), for his full-fledged anti-luck epistemology.

[4]The label is found in Pritchard's work mentioned in footnote 3, but it dates back to Mylan Engel (1992).

[5]These remarks were inspired by Richard Feldman's comments on the tracking theory in Feldman (2003, p.86).

[6]The reader will find a strong objection to the adherence condition in Feldman (2003, pp.86–7), and a defense of it in Wedgwood (2014).

[7]These are highly controversial issues and the reader is not required to accept the metaphysical stances I am adopting here. Just take my claims as heuristic tools to better grasp the idea of counterfactual scenarios and some relations among them.

[8]The need for this requirement is made clear in Pritchard (2007).

[9]The difficulty of weak safety in dealing with the lottery problem was already reported by John Greco (2007).

[10]The hidden assumption here seems to be that safety is a *sufficient* condition for knowledge. Yet this is a highly contentious assumption. The safety advocate could get rid of this predicament, in which she allegedly finds herself, by saying that she does not think of safety as a sufficient condition for knowledge, but think of it only as a *necessary* condition for knowledge, and that other conditions could be found out as equally necessary for the acquisition of knowledge. The epistemologist who sympathizes with some virtue-theoretic account of knowledge, for instance — advanced by Ernest Sosa (1991, 2007), John Greco (2010), and Linda Zagzebski (1996; 2009), among others — could try to handle the lottery problem by requiring a certain virtue/ability condition on knowledge. Thanks to Gregory Gaboardi for calling my attention to this point.

[11]An anonymous referee has made the following comment on my objection to the Risk-Weighted Construal of Safety: "...the point of this formulation of safetiness is to make the notion gradable, that is, we do not have a clear-cut distinction anymore, but different degrees in which a proposition can be safe (or risky). The authors seem to be assuming that there has to be some sort of threshold of closeness between possible worlds, below which strong safety is in force, and above which weak safety is in force. But why is this position forced upon those who maintain the view?" I see no problem with a gradable notion of safety. It does make sense to speak of a belief being safer than another belief — as it makes sense to speak of a belief being better justified than another one. However, surely there must be a threshold, even if we cannot tell whether or not a certain belief has crossed it. A true belief might be safe below the bar, safer than that, much safer than that, etc. But what is the level of safety required on knowledge? "It depends on the range of possible worlds you are taking into account!", the advocate of the risk-weighted construal of safety would say. Again, that is a too convenient condition on knowledge and we should be at least suspicious of it.

[12]Alvin Goldman (1979, p.11) writes: "I have characterized justification-conferring processes as ones that have a 'tendency' to produce beliefs that are true rather than false." And he elaborates: "The term 'tendency' could refer either to actual long-run frequency, or to a 'propensity', i.e., outcomes that would occur in merely possible realizations of the process. Which of these is intended? Unfortunately, I think our ordinary conception of justifiedness is vague on this dimension too. For the most part, we simply assume that the 'observed' frequency of truth versus error would be approximately replicated in the actual long-run, and also in relevant counterfactual situations, i.e., ones that are highly 'realistic', or conform closely to the circumstances of the actual world. Since we ordinarily assume these frequencies to be roughly the same, we make no concerted effort to distinguish them." (Goldman 1979, p.11). Although Goldman tries to preserve some vagueness in his account of reliability, we

can see that the tendency of a belief-forming process to produce true beliefs, if understood in the propensity sense, issues a condition requiring *a greater number of true beliefs (over false beliefs) in close worlds*, since these are the counterfactual situations 'highly realistic' that are 'closer to the circumstances of the actual world'.

[13]Juan Comesaña (2005) put forward a strong objection to safety theories from an alleged case of knowledge in which the safety condition was not met, hence the label 'unsafe knowledge'. For a reply to Comesaña's attack on safety and for a new objection to the safety principle, see Bogardus (2012).

Acknowledgments

I want to thank Eduardo Alves, Felipe Medeiros, Gregory Gaboardi, Rodrigo Borges, and two anonymous referees for their helpful comments on an earlier version of this essay. I especially want to thank Claudio de Almeida for his discussions with me about the main topics addressed here. I am also grateful for the partial support my research received from CAPES.