# Knowledge and Safety

Christoph Kelp

University of Stirling

99/6 Montgomery Street
Edinburgh
EH7 5EY
United Kingdom

cffk1@stir.ac.uk

*Abstract.* This paper raises a problem for so-called safety-based conceptions of knowledge: It is argued that none of the versions of the safety condition that can be found in the literature succeeds in identifying a necessary condition on knowledge. Furthermore, reason is provided to believe that the argument generalises at least in the sense that there can be no version of the safety condition that does justice to the considerations motivating a safety condition whilst, at the same time, being requisite for knowledge.

## I. INTRODUCTION

A view that has enjoyed a significant degree of attention in recent epistemology is the view that knowledgeor, to be more precise, knowledge of fully contingent propositionsrequires safe belief. Safety here is a modal condition. Duncan Pritchard has suggested the following rough formulation of the safety condition:

> (SP) S's belief is safe *iff* in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world the belief continues to be true.(Pritchard 2007: 6??)

Pritchard has argued that the safety condition can be motivated by the intuitively very plausible idea that knowledge is non-lucky true belief. The crucial idea here is that the safety condition captures the sense in which knowledge excludes luck and is thus the core condition of any anti-luck epistemology. Alternatively, Ernest Sosa (1999) has motivated the safety condition by its ability to give a better account of inductive and anti-sceptical knowledge than other modal conditions. It is not hard to see that if it turned out that safety is not even necessary for knowledge that would be a major setback for those who try to put safety to use in an anti-luck epistemology and to account for inductive and anti-sceptical knowledge. In this paper, I will argue that none of the statements of the safety condition that can be found in the literature succeeds in identifying a necessary condition for knowledge. I will also adduce some considerations that suggest that my argument poses a genuine problem for any safety condition on knowledge. To be more precise, I will provide reason to believe that no amended version of the safety condition can succeed in doing proper justice to the motivations for the safety condition just outlined, whilst, at the same time, being a condition necessary for knowledge. In this way my argument not only challenges defenders of safety to state the safety condition in such a way that it is a genuinely necessary condition for knowledge but also to provide reason

for accepting it in the first place. It is my suspicion that defenders of the safety condition cannot meet this challenge.

## II. REFINEMENTS OF THE SAFETY PRINCIPLE

It may be argued that if the safety condition is to be motivated by the idea that it captures the sense in which knowledge excludes luck, then the safety condition as stated in SP will be too weak. To see why this is so, notice that if the safety condition captures the sense in which knowledge excludes luck, then we must expect it to explain our intuition that we do not know in advance that a given ticket in a fair lottery will not win—no matter how high the odds against winning are. After all, a belief that a given ticket in fair lottery will lose, even if true, is too luckily true to qualify as knowledge. Accordingly, given that safety captures the sense in which knowledge excludes luck, it had better be the case that a belief that a given ticket in a fair lottery won't win turns out to be unsafe. However, if safety is construed along the lines of SP, a belief that a given ticket in a fair lottery won't win will not always turn out to be unsafe. Suppose one believes that some ticket won't win the lottery on the basis of the fact that the odds against winning are massive. Since the odds against winning are massive, the number of nearby possible worlds at which the ticket wins will be very small. So, at the majority of those nearby possible worlds at which one believes that the ticket will lose on the basis of probabilistic evidence against winning, one's belief will be true. So, according to SP, one's belief that the ticket will lose is safe. Safety, on this construal, will not explain why we don't know that a given ticket in a fair lottery won't win. Since provided the safety condition does capture the sense in which knowledge excludes luck, we must expect it to explain this, however, SP is too weak a version of the safety principle to be plausible.

In view of this problem, Pritchard considers two ways of strengthening the safety

principle. Here is the first one:

> (SP*) S's belief is safe *iff* in nearly all (if not all) near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world the belief continues to be true. (Pritchard 2007: 6??)

We may presume that one's belief that the ticket at issue won't win the lottery when based on the massive odds against winning will not be true in nearly all (or at the very least not in all) nearby possible worlds. For there is a number of nearby possible worlds at which the ticket wins the lottery. Accordingly, the safety principle, construed along the lines of SP*, in conjunction with the claim that knowledge entails safe belief will serve to explain one's ignorance of the proposition that the ticket in question will lose. The defect of the safety principle construed along the lines of SP is remedied.

It is noteworthy that this way of construing the safety principle is not unprecedented in the literature. For instance, Ernest Sosa can be reconstructed as construing safety along similar lines:

> [A] belief by S is "safe" iff: … not easily would S believe that p without it being the case that p. (Sosa 1999: 142)

Given a standard possible worlds semantics of the relevant modal notions, Sosa's claim is tantamount to the claim that S's belief is safe iff S avoids false beliefs at nearby possible worlds.

Another defender of safety who goes down this path (or at least something very similar to it) is Timothy Williamson. Williamson characterises the safety condition in the following way:

> Reliability and unreliability, stability and instability, safety and danger, robustness and fragility are modal states. They concern what could easily have happened. (Williamson

2000: 123)

> For present purposes [i.e. for the purposes of spelling out the notion of reliability that, according to Williamson, is necessary for knowledge], we are interested in a notion of reliability on which, in given circumstances, something happens reliably if and only if it is not in danger of not happening… In particular, one avoids false belief reliably in [a case] if and only if one avoids false belief in every case similar to . (Williamson 2000: 124)

Williamson claims that states such as safety and reliability concern what could easily have happened. At the same time, he also maintains that in order to believe safely (or "reliably" in Williamson's terms), one must avoid false belief in similar cases. Given that this is so, it might now seem that there are two characterisations of safety in Williamson. However, given a standard possible worlds semantics of the relevant modal notions (of easy possibility) and given a standard understanding of distance between possible worlds, Williamson can be interpreted as giving a single characterisation of safety ("reliability")—one that is very much in line with the ones offered by Sosa and Pritchard. To see how this works, notice, first, that, according to a standard possible worlds semantics of the notion of easy possibility, something could easily have happened just in case it happens at a nearby possible world. Moreover, according to the standard understanding of distance between possible worlds, distance between possible worlds is a function of similarity between worlds. The more similar a possible world is to another possible world, the closer (more nearby) it is. If we take the range of nearby possible worlds to be worlds at which those cases that are similar to the actual world obtain, we can reconstruct Williamson's remarks here as effectively claiming that one's belief is safe ("reliable") if and only if one avoids false belief at nearby possible worlds.

Before I move on I would like to highlight one important aspect of the safety principle, viz. that there is reason to believe that it must at the very least feature an index to ways of belief-formation. Pritchard is explicit about this: He requires the nearby possible worlds at which S has to continue to believe truly in order to believe safely to be worlds at which S continues to

form her belief in the same way as in the actual world. However, a similar move can also be found in Sosa (2002: 275-6) who in a later paper proposes to index the safety principle to what he calls indications of truth and in Williamson (2000: 123) who maintains that the initial conditions need to be held fixed or almost fixed. We must suppose, I take it, that the way of belief-formation is part of the initial conditions that need to be held fixed. Indexing to ways of belief-formation is necessary in order to secure correct predictions in cases in which one uses one such way in the actual world and on that basis forms a true belief that *p*, while at (at least some of) the nearest possible worlds at which *p* is false, which may be very nearby, one forms ones belief in a different way and so ends up with a false belief that *p*. The most prominent case of this sort is Robert Nozick's (1981: 179) grandmother case: Suppose granny is visited by her grandson and comes to believe by looking at him that he is well. Granny is good at telling these things by looking. Intuitively, she knows that her grandson is well. At some nearby possible worlds, however, her grandson is ill. In order to save granny from distress, at (some of) these worlds her family tells her that her grandson is well but had something important to do and for that reason couldn't come and visit. Granny forms a false belief at these possible worlds. So, in the absence of the index to ways of belief-formation, granny's true belief, acquired in the actual world by looking, that her grandson is well will be unsafe. A theory that makes safety, so construed, necessary for knowledge will predict, counterintuitively, that granny does not know that her grandson is well when she looks at him and on that basis forms a true belief to that effect. Indexing to ways of belief-formation will remedy this defect. After all, at those nearby possible worlds at which granny forms a false belief she comes by her belief in a different way than in the actual world. She relies on testimony rather than on looking. Accordingly, it is important to be aware that the safety principle will need to witness at least an index to ways of belief-formation.

## III. COMESAÑA'S ARGUMENT AGAINST SAFETY

Juan Comesaña has recently argued that a safety condition of the kind defended by Pritchard, Sosa and Williamson is not a necessary condition for knowledge. He adduces the following example to bring the point home:

> There is a Halloween party at Andy's house, and I am invited. Andy's house is very difficult to find, so he hires Judy to stand at a crossroads and direct people towards the house (Judy's job is to tell people that the party is at the house down the left road). Unbeknownst to me, Andy doesn't want Michael to go to the party, so he also tells Judy that if she sees Michael she should tell him the same thing she tells everybody else (that the party is at the house down the left road), but she should immediately phone Andy so that the party can be moved to Adam's house, which is down the right road. I seriously consider disguising myself as Michael, but at the last moment I don't. When I get to the crossroads, I ask Judy where the party is, and she tells me that it is down the left road. (Comesaña 2005: 398)

Comesaña points out that, intuitively, he knows that the party is down the left road. At the same time, his belief is not safe. Since he almost decided to disguise himself as Michael, at some nearby possible worlds, he does disguise himself as Michael in which case the party will be moved just after Judy tells him that it is down the left road. At such worlds, Comesaña ends up with a false belief. Since at those worlds, Comesaña forms his belief in the same way as in the actual world—viz. by testimony from Judy—his belief is unsafe. Hence, knowledge does not require safety. Or so argues Comesaña.

However, Comesaña's argument strikes me as unconvincing. To see why this is so, recall, first, that distance of possible worlds is a function of similarity between possible worlds —the more similar a possible world is to another one, the closer it is. Now, let's ask how similar a world at which Comesaña acquires a false belief that the party is down the left road is to the actual world at which he comes to know the very same proposition. It would seem that there are some significant differences between those worlds: At the very least, Comesaña must

have decided to disguise himself as Michael, he must have successfully done so, Judy must have formed a false belief that she is talking to Michael, she must have made a phone call, the party must have been moved. Given that this is so, however, the defender of safety may now venture argue that the worlds at which Comesaña acquires a false belief are not similar enough to the actual world to still be counted as nearby. (The point here is, of course, that a situation can *almost* obtain, while, at the same time, the worlds at which it obtains are quite *dissimilar* from and hence *not close* to the actual world.) If the worlds at which Comesaña acquires a false belief are not nearby, however, then his belief that the party is down the left road will still be safe. The defender of safety would then be off the hook.

Now Comesaña may retort that even if there are a significant number of things that have to be different if, in the example, he is to end up with a false belief, the worlds at which he ends up with a false belief are still similar enough to the actual world to be counted as nearby. Even so, however, the concession that there are a significant number of things that have to be different may be all the defender of safety needs to rescue the safety condition. To bring out exactly why this is so, it will be helpful to contrast Comesaña's case with the sort of case in which a defender of safety would want the safety condition to explain the subject's ignorance. The most significant class of cases comprises, of course, Gettier cases. (If one wants to motivate safety by its ability to capture the sense in which knowledge excludes luck, then another significant class of cases will comprise lottery cases.) Consider, for instance, the case of Henry who drives through the country, looks at the only real barn in a field otherwise full of barn façades and comes to believe, by looking, that he is facing a barn. Intuitively, Henry's belief does not qualify as knowledge. At the same time, there is excellent reason to believe that belief turns out to be unsafe. After all, at some nearby possible world Henry looks at a barn façade and acquires a false belief. Thus, the safety condition will be able to explain our intuition

that Henry lacks knowledge.

But now notice just how similar a situation in which Henry acquires a false belief is to the situation that actually obtains: All that has to happen is that Henry looks out of the window a few moments earlier or later. Importantly, it is plausible that a situation in which Henry looks out of the window a few moments earlier or later is much more similar to the situation that actually obtains than the situation in which Comesaña acquires false belief concerning the whereabouts of the party is to the situation in which he acquires a true belief. (Recall all the things that have to be different for Comesaña to end up with a false belief.) What these considerations show is, of course, that there is a relevant difference between the cases against in which the defender of safety wants the safety condition to predict ignorance and Comesaña's case. So, even if the defender of safety has to concede that worlds at which Comesaña ends up with a false belief are similar enough to count as nearby, he may now place further restrictions on the safety principle that will allow him to continue to use safety to predict ignorance in, for instance, Gettier and lottery cases, whilst, at the same time, also allowing him to analyse Comesaña's belief as safe.

One promising way of so restricting the safety principle has recently been proposed by Pritchard (although in a slightly different context). This version of the safety principle weights nearby possible worlds depending on how close they are to the actual world. The crucial idea is that continuing to believe the truth at very close nearby possible worlds is more important than it is at nearby possible worlds that are not so close. Here, then, is Pritchard alternative version of the safety principle:

(SP**) S's belief is safe *iff* in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, the belief continues to be true. (Pritchard 2007: 20)

There is reason to believe that Comesaña's problematic belief satisfies SP**. After all, since there is a significant number of things need to be changed for Comesaña to end up with a false belief, it is also plausible, first, that there is no very close nearby possible world at which all of these things change and, second, that at *most* nearby possible worlds some such fact remains unchanged. If so, however, Comesaña's belief satisfies SP**. (At the same time, a defender of SP** can make a concession to Comesaña and allow that there are *some* nearby possible worlds at which all the things that need to be changed for him to end up with a false belief do change.) The defender of safety is, once again, off the hook.

## IV. A NEW ARGUMENT AGAINST SAFETY

In this section I will present a new argument to the effect that safety is not a necessary condition for knowledge. Like Comesaña, I will present a case in which the subject intuitively knows the proposition believed, while, at the same time, her belief is unsafe. Unlike Comesaña's argument, the subject's belief is unsafe no matter whether the safety condition is construed along the lines of SP, SP*, or SP**. The case is a variation of what, presumably, was the first Gettier case (due, somewhat surprisingly, to Bertrand Russell). Russell (1948: 170-1) imagined a situation in which he wakes up in the morning, comes to down the stairs, has a look at the grandfather clock, sees that it reads 8.22 and on that basis forms a belief that it's 8.22. Russell's belief is well justified: He knows the clock to be highly reliable, has no reason to believe that it is not working properly etc. Moreover, his belief is true. It is in fact 8.22. However, here comes the catch, the clock has stopped working exactly twelve hours earlier. Notice that, in the present version of the case, Russell's belief that it's 8.22 is not safe. After all, it is plausible that there is a wide range of close and very close nearby possible worlds at which

Russell comes down a minute earlier or later. (All that has to happen for him to come down a minute later, for instance, is that he stays in bed for a minute longer—and notice just how easily that can happen.) If at such a world he acquires his belief by reading the stopped clock, he will form a false belief. At the same time, in the present version of the case, intuitively, Russell does not know that it is 8.22 when he forms a true belief by reading the grandfather clock. Thus, in the present version case, the safety condition manages to accommodate our intuitions rather neatly.

But now consider the following variation of the case. Suppose Russell's arch-nemesis has an interest that Russell forms a belief (no matter whether true or not) that it's 8.22 by looking at the grandfather clock when he comes down the stairs. Russell's arch-nemesis is prepared to do whatever it may take in order to ensure that Russell acquires a belief that it's 8.22 by looking at the grandfather clock when he comes down the stairs. (Since we are concerned with a conceptual claim here Russell's arch-nemesis may have means available to do so that we can imagine only in our wildest dreams. For instance, Russell's arch-nemesis may be an evil-demon who can set the clock to 8.22 with his invisible hand a second before Russell looks at it.) However, Russell's arch-nemesis is also lazy. He will act only if Russell does not come down the stairs at 8.22 of his own accord. Suppose, as it so happens, Russell does come down the stairs at 8.22. Russell's arch-nemesis remains inactive. Russell forms a belief that it's 8.22. 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. After all, he looks at a perfectly working clock, he has the ability to read the clock, exercises his ability and hits upon the truth through the exercise of this ability. However, Russell's belief that it's 8.22 is not safe—never mind whether safety is construed along the lines of SP, SP* or SP**. At all nearby possible worlds at which he comes down a minute earlier or later his arch-nemesis steps on the scene and sets the clock to 8.22

anyway. At those worlds, Russell forms a false belief that it's 8.22. At the same time, he forms his belief in the same way as in the actual world—by reading the clock. If at all nearby possible worlds at which Russell comes down a minute earlier or later he still forms a belief that it's 8.22 in the same way, it is not the case that at most—never mind nearly all or all—nearby possible worlds at which he forms a belief in that way he avoids forming a false belief. So, Russell's belief is unsafe if safety is construed along the lines of SP or SP*. Furthermore, since some of the possible worlds at which Russell comes down a minute earlier or later are among the very close nearby possible worlds (again, notice just how easily Russell may have stayed in bed a minute longer), it is not the case that at all very close nearby possible worlds at which he forms his belief in the same way he avoids false beliefs. So, Russell's belief is unsafe if safety is construed along the lines of SP**, too.

## V. CONCLUDING REMARKS: THE PROSPECTS FOR SAFETY

If the arguments I have presented are sound, then safety—at least in the versions found in the literature—will not serve as a necessary condition for knowledge. The question remains, however, whether there are other versions of the safety principle that will be more successful than the ones discussed. One may wonder, for instance, whether the safety principle could be restricted in such a way that Russell's problematic belief turns out to be safe. Now, I do not doubt that there are some ways of so restricting the safety principle. There is reason to believe, however, that any such safety principle will fail to do justice to the considerations that motivated the safety principle in the first place—that is, the idea that safety captures the sense in which knowledge excludes luck or that it gives a better account of inductive and anti-sceptical knowledge than other modal conditions on knowledge.

To see exactly why this is so, notice, first, that the structure of the present case is very

similar indeed to the structure of a core Gettier case—*viz.* the case of Henry in barn façade county. In my case, there are plenty of possible situations in which Russell ends up looking at a stopped clock and forms a false belief. Similarly, in Henry's case, there are plenty of possible situations in which Henry ends up looking at a barn façade and also forms a false belief. Moreover, in each case, these possible situations might, it would seem, equally easily obtain. All that has to happen in the variation of the grandfather clock case is that Russell stays in bed a minute longer, for instance, while all that has to happen in Henry's case is that Henry looks out of the window a minute later. Unlike in Comesaña's case, in my case there is no significant number of things that have to change for him to end up with a false belief and, accordingly, no relevant difference between my case and Henry's that a defender of safety may venture to exploit by placing suitable restrictions on the safety principle. On the contrary, given the similarities in structure between my case and the case of Henry, it would seem that, on any version of the safety principle, the subject's belief will turn out safe in the one case just in case it will turn out safe in the other. In consequence any set of restrictions on the safety principle that will render Russell's belief safe will also render Henry's belief safe.

The second part of my argument aims to show that on any version of the safety principle that does justice to the considerations that have been adduced to motivate the safety condition on knowledge Henry's belief must turn out to be unsafe. Since we have just seen that there is reason to believe that any version of the safety principle that renders Henry's belief unsafe will also have to render Russell's belief unsafe, if my argument is successful, there is no version of the safety principle that does justice to the considerations motivating it, whilst, also analysing Russell's belief as safe.

Let us first turn to the first motivation for safety—that the safety condition captures the sense in which knowledge excludes luck. An intuitively plausible explanation of why subjects

in Gettier cases lack knowledge is that their beliefs are just too lucky to qualify as knowledge. Accordingly, it will not be surprising that an intuitively plausible explanation of why Henry does not know that he is looking at a barn is that his belief is just too lucky to qualify as knowledge. That means, however, that any version of the safety condition that does justice to the present motivation will have to analyse the beliefs of subjects in Gettier cases (among them Henry's belief) as unsafe. So, the first way of motivating the safety principle will not mesh with a version of the safety principle that analyses Russell's belief as safe.

Recall that, according to the second motivation for safety, the safety condition on knowledge does better in explaining inductive and anti-sceptical knowledge than any other modal condition on knowledge. Now it is obvious that this way of motivating the safety condition will be successful only if the idea there must be some modal condition on knowledge is itself suitably motivated. However, modal conditions on knowledge are again typically motivated by their ability to explain our intuitions in Gettier cases. For instance, both Fred Drestke (1971) and Robert Nozick (1981), who were, presumably, the first to introduce modal conditions on knowledge, are very clear about this. They both use Gettier-style cases in order to establish that their preferred modal condition has an edge over so-called causal conditions on knowledge. But if modal conditions on knowledge are motivated by their ability to predict ignorance in Gettier cases, then in order to be able to motivate the safety condition in the way envisaged, we must spell out the safety condition in such a way that Gettiered subjects' beliefs turn out to be unsafe. Since the case of Henry in barn façade county is a Gettier case—and quite a paradigmatic one at that—that means that we must spell out the safety condition in such a way that Henry's belief turns out to be unsafe. So the second way of motivating the safety condition does not mesh with a version of the safety principle that analyses Russell's belief as unsafe.

Given that this is so, defenders of the safety condition on knowledge owe us not only a

statement of the safety principle on which Russell's belief comes out safe but also a reason to believe why the intended version of the safety condition is required for knowledge in the first place. Since I cannot see how any such reason could be provided without adverting to the safety condition's ability explain our intuitions in Gettier cases, I suspect that defenders of the safety condition will be unable to meet this challenge.

## BIBLIOGRAPHY


Comesaña, Juan. 2005. "Unsafe Knowledge." *Synthese* 146: 395-404.

Dancy, Jonathan. 1985. *Introduction to Contemporary Epistemology*. Oxford: Blackwell.

Dretske, Fred. 2000. "Conclusive Reasons." In Sven Bernecker and Fred Dretske (eds.), *Knowledge. Readings in Contemporary Epistemology*. Oxford: OUP, 42-62.

Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 829-39.

Greco, John. 2003. "Virtue and Luck, Epistemic and Otherwise." *Metaphilosophy* 43: 353-66.

——. 2007. "Worries about Pritchard's Safety." *Synthese* 158: 299-302.

Nozick, Robert. 1981. *Philosophical Explanations*. Oxford: OUP.

Pritchard, Duncan. 2005. *Epistemic Luck*. Oxford: OUP.

——. 2007. "Anti-Luck Epistemology." *Synthese* 158: 277-297.

Russell, Bertrand. 1948. *Human Knowledge: Its Scope and its Limits*. London: Allen & Unwin.

Sosa, Ernest. 1999. "How to Defeat Opposition to Moore." In James Tomberlin (ed.), *Philosophical Perspectives 13: Epistemology*. Oxford: Blackwell, 141-54.

——. 2002. "Tracking, Competence, and Knowledge." In Paul Moser (ed.), *The Oxford Handbook of Epistemology*. Oxford: OUP, 264-286.

# NOTES