JUAN COMESAÑA

UNSAFE KNOWLEDGE

ABSTRACT. Ernest Sosa has argued that if someone knows that p, then his belief that p is "safe", and Timothy Williamson has agreed. In this paper I argue that safety, as defined by Sosa, is not a necessary condition on knowledge – that we can have unsafe knowledge. I present Sosa's definition of safety and a counterexample to it as a necessary condition on knowledge. I also argue that Sosa's most recent refinements to the notion of safety don't help him to avoid the counterexample. I consider three replies on behalf of the defender of safety, and find them all wanting. Finally, I offer a tentative diagnosis of my counterexample.

1. INTRODUCTION

Ernest Sosa has argued that if someone knows that p, then his belief that p is "safe", and Timothy Williamson has agreed.[1] In this paper I argue that safety, as defined by Sosa, is not a necessary condition on knowledge – that we can have unsafe knowledge. In the next section I present Sosa's definition of safety, and in Section 3 I present a counterexample to it as a necessary condition on knowledge. In that same section I also argue that Sosa's most recent refinements to the notion of safety don't help him to avoid the counterexample. In Section 4 I consider three replies on behalf of the defender of safety, and find them all wanting. Finally, in Section 5 I offer a tentative diagnosis of my counterexample.

2. SENSITIVITY AND SAFETY

What else, besides truth and belief, is required for knowledge? Robert Nozick famously proposed as a further requirement that the belief in question be sensitive to the truth of the matter, where a belief that p by a subject S is *sensitive* if and only if S would not believe that p if p were false.[2] In the terminology of possible worlds, a belief is sensitive for a subject just in case the subject does not believe it in any of the close possible worlds where it is false.

Soon after Nozick proposed it, the sensitivity requirement was subject to counterexamples like the following:

Synthese (2005) 146: 395–404
DOI 10.1007/s11229-004-6213-7

GARBAGE CHUTE: I throw a trash bag down the garbage chute of my condo. Some moments later I believe, and know, that the trash bag is in the basement. However, the closest possible world where my belief is false is plausibly one where, unbeknownst to me, the bag is stuck somewhere in the chute, and I still believe that it is in the basement.[3]

In GARBAGE CHUTE, my belief that the trash bag is in the building's basement is not sensitive, yet it counts as knowledge.

The sensitivity requirement has also strong skeptical consequences. If our beliefs have to be safe to count as knowledge, then our beliefs that skeptical scenarios do not obtain (that we are not brains in a vat, that we are not dreaming right now, that we are not in The Matrix) do not amount to knowledge. If it were false that I am not a brain in a vat, then I would be a brain in a vat who believes that he is not a brain in a vat. Nozick thought that this consequence of the requirement of sensitivity was a virtue of the notion, but many epistemologists think that we do know that skeptical scenarios do not obtain, and so they have accordingly taken this consequence as a further counterexample to sensitivity.

The sensitivity requirement is a counterfactual and, as such, is not equivalent to its contrapositive. Having noticed this, Ernest Sosa proposed that we replace the requirement of sensitivity with its contrapositive, which he calls *safety*. Sosa has several formulations of the safety requirement. Here is his first approximation:

Call a belief by S that p "safe" iff: S would not believe that p without it being so that p. (Alternatively, a belief by S that p is "safe" iff: as a matter of fact, though perhaps not as a matter of strict necessity, S would not believe that p without it being so that p.) (Sosa 1999, 378)

In terms of possible worlds, a belief that p by S is safe if and only if there is no close possible world where S believes that p and p is false.

That characterization of safety is not entirely satisfactory, though. One problem with it is that a good basis for my belief can preempt a bad one.

SICK PATIENT: I am seriously ill and I ask my doctor whether I will live one more week. The doctor performs a test on me, and answers affirmatively. I now base my belief that I will live one more week on the doctor's testimony. But suppose that my condition is caused by a rapidly spreading virus, whose rate of infection is somewhat erratic, and which could have acted in my body more quickly than it actually did. Had it acted more quickly, the test would have indicated this and the doctor would have told me that I wasn't going to live one more week. But suppose also that, had that happened, I would still have believed, out of wishful thinking, that I was going to live one more week.

In SICK PATIENT, I still know that I will live one more week (in the actual scenario, my belief is based on the reliable testimony of the doctor, and wishful thinking doesn't play a part), even though there are close possible worlds where I have that same belief and it is false.

If the problem is that a belief can have different bases in different (but close) possible worlds, then the solution is to require same-basis safety:

A belief that p by S is safe if and only if S would not believe that p on the same basis without it being so that p.[4]

My belief in SICK PATIENT *does* satisfy this revised safety requirement, and the requirement itself is initially intuitively plausible. Moreover, unlike sensitivity, it does not entail that I do not know that the trash bag is in the basement in GARBAGE CHUTE, and it also does not entail that we do not know that skeptical scenarios do not obtain.[5] Despite it having these virtues, I will argue that the safety requirement is incorrect. I will present a case where a subject knows that p and yet the subject would easily have believed that p on the same basis on which he actually believes it without it being so that p.

3. A COUNTEREXAMPLE TO SAFETY

The case is the following:

HALLOWEEN PARTY: There is a Halloween party at Andy's house, and I am invited. Andy's house is very difficult to find, so he hires Judy to stand at a crossroads and direct people towards the house (Judy's job is to tell people that the party is at the house down the left road). Unbeknownst to me, Andy doesn't want Michael to go to the 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.

In this case, after I talk to Judy I know that the party is at the house down the left road, and yet it could very easily have happened that I had the same belief on the same basis (Judy's testimony) without it being so that the belief was true. That is, in this case I know that p but my belief that p is not safe – I have unsafe knowledge.

Sosa (2002) presents a modification of the safety condition that might be thought to help with HALLOWEEN PARTY. Sosa says that a basis can be safely related to a certain fact p not directly but dependently on a certain condition. In Sosa's terminology, belief sources issue "indications" that certain facts obtain. An indication that p, I(p), "indicates the truth outright" if and only if I(p) would be so only if p were so; and I(p) indicates the truth *dependently on a condition C* if and only if (i) I(p) doesn't indicate the truth outright, (ii) C obtains, and (iii) C and I(p) would jointly be so only if p

were so. The safety-related necessary condition for knowledge is, then, the following:

S knows that p on the basis of an indication I(p) only if either (a) I(p) indicates the truth outright and S accepts that indication as such outright, or (b) for some condition C I(p) indicates the truth dependently on C and S accepts that indication as such not outright but guided by C (so that S accepts the indication as such on the basis of C).[6]

In HALLOWEEN PARTY, the indication is the fact that Judy tells me that the party is at the house down the left road. Does that testimony indicate the truth either outright or dependently on some condition that guides my belief? It is clear that Judy's testimony doesn't indicate the truth outright. Again, it could easily have happened that Judy said that the party is at the house down the left road without it being so that the party was at the house down the left road. But what about dependent indication? Isn't there a condition C such that Judy's testimony indicates the truth *dependently on C*? There obviously is: that condition is the fact that the subject that Judy is talking to doesn't look like Michael to her. It *could not* easily have happened that Judy said that the party is at the house down the left road to someone that doesn't look like Michael to her without it being so that the party is at the house down the left road. Judy's testimony, then, indicates the truth dependently on the condition that the subject she is talking to doesn't look like Michael to her.

But this fact doesn't save the safety condition from being refuted by HALLOWEEN PARTY, for clause (b) of the necessary condition is not just that there be a condition C such that the basis of the belief indicates the truth dependently on C,[7] but in addition my belief has to be *guided by* the presence of C. And in HALLOWEEN PARTY I am unaware of the relevance of the respective condition to the truth of Judy's testimony: I would have believed that p whether or not I looked like Michael to Judy. Therefore, HALLOWEEN PARTY is a counterexample to the safety condition even taking into account dependent indication.

4. OBJECTIONS AND REPLIES

In this section I will consider three ways in which a defender of safety can reply to the challenge posed by HALLOWEEN PARTY. First, she may say that it is not a counterexample to safety because, contrary to what I said, my belief that the party is at the house down the left road is safe. Second, she may say that it is not a counterexample to safety because, contrary to what I said, I do not know that the party is at the house down the left road. And third, she may say that it is not a counterexample to safety because,

although I do have knowledge, I don't have the kind of knowledge that requires my belief to be safe. Let's take those possible objetions in order.

First, then, a defender of safety could say that my belief that the party is at the house down the left road is safe after all. This reply seems to me the most plausible of the three that I will consider, but I think that it is ultimately unsatisfactory.

It is clear, to begin with, that in HALLOWEEN PARTY my belief does not satisfy Sosa's definition of safety: it *could* easily have happened that I had the same belief on the same basis and yet the belief was false. And, given that safety is a technical notion introduced by Sosa as a necessary condition on knowledge, we cannot retreat to a pre-theoretic notion of "safety" and claim that the example doesn't touch *it*. It is of course true that my belief has something epistemically good going for it – it is, after all, a piece of knowledge. We can, if we want, call that something epistemically good that it has going for it "safety". But what we do not have if we do this is a theory of what safety amounts to. Sosa tries to provide such a theory, and claims that for a belief that p to be safe is for it to have a specific modal relation to the fact that p. Maybe someone else will provide a different theory of what safety amounts to.[8] HALLOWEEN PARTY is advanced as a counterexample to Sosa's proposed definition, of course, not to the platitude that a proposition that amounts to knowledge has something good going for it, nor to an alternative definition that doesn't yet exist.

A more specific defense of the safety condition appeals to the fact that safety is time-sensitive. Some years ago, it was a close possibility that Brian would become a lawyer; now, however, it no longer is a close possibility: Brian is safely not a lawyer. Similarly, the defender of the safety condition for knowledge might say, before I decided not to dress up as Michael, it was a close possibility that I would falsely believe that the party is at the house down the left road; now, however, it no longer is a close possibility: I safely believe the truth of the matter.[9]

Again, it is clear that my belief in HALLOWEEN PARTY does not satisfy Sosa's definition of "safety", and so, if it is indeed the case that my belief in that case *is* safe according to a time-sensitive notion of safety, then Sosa's notion of safety is not time-sensitive. But even leaving that aside, it seems to me simply false that, in HALLOWEEN PARTY, after I decide not to dress up as Michael it is no longer a close possibility that I have a false belief. When considering whether the proposition that p obtains safely at t in the actual world, we consider whether it obtains in possible worlds that differ from the actual world just slightly *right before* t.[10] And, in HALLOWEEN PARTY, I seriously consider dressing myself up as Michael just before driving to the intersection where Judy is standing.

Moreover, we can change the case so that the time when I decide not to dress up as Michael is even closer to the time when I believe that the party is at the house down the left road. We can suppose, if we want, that I *was* dressed up as Michael, and that I decided to take the disguise off at the last minute, just before arriving at the intersection where Judy is. We can also make more radical changes to the case, by imagining that I am dressed up as Michael, but that I'm going to the party with Alex, and that we decide at the last moment that he will ask Judy for directions, not me. In any of those cases, there are possible worlds that differ from the actual world just in what happens right before I believe that the party is at the house down the left road, and that are such that my belief is false.[11] I conclude, then, that the appeal to the time-sensitivity of safety doesn't save the condition from being refuted by HALLOWEEN PARTY.

Second, the defender of safety might say that I do not know that the party is at the house down the left road. If the reply is left at that, however, it is overwhelmingly implausible. Of course, any theory can be made immune to counterexamples if one is willing to bite big enough bullets, but the price to pay is that it is no longer clear what the resulting theory is a theory of. The defender of safety is better off claiming that there is an ambiguity in "knowledge" and its cognates, and that I know that the party is at the house down the left road in a sense of "know" that is different from that in which safety is a requirement for knowledge. That is the kind of answer that I will consider next.

In the third place, then, I want to consider the reply that says that although I do have knowledge, it is not the kind of knowledge that requires my belief to be safe. Perhaps I know that the party is at the house down the left road in a sense of "know" related to knowing-how (for instance, I know how to get to the party), whereas safety is a requirement only for kinds of propositional knowledge that do not have implications regarding the subject's abilities.[12]

I think that, in the end, this reply is also implausible. I know that the party is at the house down the left road in the same sense that I know that I have hands. This seems to me intuitively true, but it is also supported by ambiguity tests used by linguists. Suppose that Fred deposited a check this morning, and that Jenny had a picnic by the river. The following sentence is then infelicitous: "Fred went to the bank, and so did Jenny". But there is nothing infelicitous in the sentence "Mary knows that she has hands, and Juan, that the party is at the house down the left road".[13] Moreover, while it is true that on the most natural reading of HALLOWEEN PARTY it is implied that I have certain abilities (like the ability of getting to the party), this is not essential to the example. The case could be modified so that I do

not have that ability, and we would still say that I have the knowledge in question.

5. DIAGNOSIS

What features of HALLOWEEN PARTY make it the case that I have unsafe knowledge? In other words, why is it that safety is not a necessary condition for knowledge? That is a very interesting question, and I do not know what the full answer is. I do have some suspicions regarding the diagnosis of the case, though, and I would like to conclude by laying them down.

It might be thought that by attacking safety as a necessary condition on knowledge I am arguing against epistemic externalism – the thesis that the epistemic status of a belief can depend on factors that are "external" to the subject, either in the sense that they are not part of the subject's mind or in the sense that they are not accessible in a privileged way by the subject. The safety conditional can, after all, be seen as an attempt to capture what it is for a belief to be reliable, and the requirement of reliability is a paradigmatically externalist requirement.

That is, I grant, a possible diagnosis of my counterexample, but it is not implied by the case, and it is not my diagnosis. Sosa's definition of safety can indeed be seen as an attempt to analyze what reliability amounts to, as can Nozick's definition of sensitivity, but neither safety's nor sensitivity's demise need be reliability's demise. Dissatisfaction with requirements on knowledge that take the form of modal conditionals will have adverse consequences for the requirement of reliability only if reliability *has* to be accounted for in terms of modal conditionals. But this need not be the case. Two other possibilities are salient. First, maybe reliability can be accounted for in terms of relevant alternatives, where what it is for an alternative to be relevant is not, in turn, accounted for in terms of modal conditionals. Second, maybe reliability is primitive, in the sense that it cannot be accounted for in terms of notions that do not presuppose the notion of reliability itself. However that may be, far from implying a rejection of reliability, my diagnosis of HALLOWEEN PARTY presupposes that reliability is a necessary condition for knowledge.[14]

It will be useful to start by considering the abstract structure of the case. In HALLOWEEN PARTY there is a source of information (Judy's testimony) whose reliability depends on whether some external circumstance obtains (whether I look like Michael to her). The external circumstance *almost* obtains, but doesn't. As a consequence, the source is reliable and (given that everything else in the case is epistemically propitious) the sub-

ject gains knowledge. What the case illustrates, I think, is that knowledge tolerates near-unreliability, whereas safety doesn't.

Suppose that the way I decide whether to dress up as Michael is by flipping a coin: heads I do, tails I do not. I flip a coin and it comes up tails. Contrast now HALLOWEEN PARTY so described with a case where *Judy* decides whether to tell the truth by flipping a coin: heads she lies, tails she tells the truth. She flips the coin and it comes up tails. In this second case, I do *not* know that the party is at house down the left road. In both cases, what decides whether I believe something true is the flip of a coin, but in the first case I know whereas in the second I don't. What accounts for this difference?

It seems to me that what accounts for this difference is that, in the second case, Judy's testimony is unreliable, whereas in the first it is reliable. The fact that she could have easily lied to me has a different explanation in both cases. In the second case, she could have easily lied to me because her testimony was unreliable, whereas in the first case she could have easily lied to me because, although her testimony was reliable, it was not *reliably* reliable. Had my flip of the coin come up heads, Judy's testimony would have been unreliable.

Going back to the original case, in HALLOWEEN PARTY Judy's testimony is reliable, although it is not reliably reliable: had I looked like Michael to her, she would have told me something that turns out to be false. And whereas reliability is a plausible necessary condition on knowledge (and that is why I do not have knowledge when Judy decides whether to lie to me by flipping a coin), reliable reliability is not. In other words, a belief need not be formed in a way that is reliably reliable in order to count as knowledge, as long as it is formed in a way that is reliable. But if a belief of mine is not reliably reliable, then there are close worlds where I believe it and yet it is false. Therefore, safety is incompatible with unreliably reliable beliefs.[15] And, given that knowledge is compatible with unreliably reliable beliefs whereas safety is not, there can be unsafe knowledge.

ACKNOWLEDGEMENTS

The idea for this paper arouse in conversation with Michael Pace. Many thanks also to Manuel Comesaña, Carolina Sartorio, Ernest Sosa, Brian Weatherson, the members of the metaphysics and epistemology reading group at UW-Madison (especially Alan Sidelle), and two anonymous referees for *Synthese*.

NOTES

[1] Sosa (1996, 1999, 2000, 2002), Williamson (2000).

[2] Cf. Nozick (1981, 172).

[3] This counterexample is due to Sosa (2000), who credits Vogel (1987) with the initial counterexamples to sensitivity.

[4] Cf. Williamson (2000, 149). This is equivalent to Sosa's requirement that S's belief must be based on a reliable indication – i.e., an indication that would not have been present without it being so that p (cf. Sosa 2002) – and it mirrors Nozick's modification of his sensitivity requirement by appeal to "methods".

[5] Because the closest worlds where I believe that the trash bag is in the basement are worlds where the trash bag indeed is in the basement, and the closest worlds where I believe that I am not a brain in a vat are worlds where I indeed am not a brain in a vat.

[6] Sosa (2002, 275–276).

[7] That would be an obviously trivial condition: every indication that p indicates the truth dependently on the condition that p, or any other condition that entails that p.

[8] In Williamson (2000, 149), Timothy Williamson cites (apparently approvingly) Sosa's modal definition of safety, but also expresses some willingness to go along with a topological characterization instead. Unfortunately, Williamson doesn't give enough details of the topological characterization to judge whether HALLOWEEN PARTY is a counterexample to it or not. It is, however, as far as I can see, a counterexample to premise (9) in his anti-luminosity argument, which premise, he says, is an assumption of "a connection between knowledge and safety from error". Cf. Williamson (2000, 128).

[9] Peacocke (1999, 310–328), advances a notion of "close possibility" which is time-sensitive in this sense, and Williamson (2000, 124), cites this treatment approvingly.

[10] Cf. Williamson (2000, 124), and Peacocke (1999, 322).

[11] One could insist that, even in those cases, there is a time t before I come to believe the proposition such that, after t, it no longer is a close possibility that I will have a false belief. But if this is true in these cases, then I cannot see why it is not true of any case where I have a true belief, and so this strategy threatens to collapse the notion of safety with that of truth simpliciter.

[12] Ernest Sosa suggested this kind of reply in private conversation.

[13] There might be something odd about that sentence, but it is the oddness of attributing knowledge that one has hands, rather than the oddness that would be generated if "knowledge" were ambiguous.

[14] I have defended reliability as a necessary condition for knowledge in Comesaña (2002).

[15] Williamson (2002, 124–128), argues that a topological characterization of safety has the consequence that safety itself doesn't iterate. But see my note 8.

REFERENCES

Comesaña, Juan (2002), "The Diagonal and the Demon", *Philosophical Studies* **110**, 249–266.

Nozick, Robert: 1981, *Philosophical Explanations*, Oxford University Press, Oxford.

Peacocke, Christopher: 1999, *Being Known*, Oxford University Press, Oxford.

Sosa, Ernest: 1996, 'Postscript to "Proper Functionalism and Virtue Epistemology" ', in J. L. Kvanvig (ed.), *Warrant in Contemporary Epistemology*, Rowman & Littlefield.
Sosa, Ernest: 1999, 'How Must Knowledge Be Modally Related to What Is Known?', *Philosophical Topics* **26** (1/2), 373–384.
Sosa, Ernest: 2000, 'Skepticism and Contextualism', in J. Tomberlin (ed.), *Philosophical Issues*, **10**, 1–18.
Sosa, Ernest: 2002, 'Tracking, Competence, and Knowledge', in P. Moser (ed.), *The Oxford Handbook of Epistemology*, Oxford University Press, Oxford.
Vogel, Jonathan: 1987, 'Tracking, Closure, and Inductive Knowledge', in S. Luper-Foy (ed.), *The Possibility of Knowledge*, Rowman and Littlefield.
Williamson, Timothy: 2000, *Knowledge and its Limits*, Oxford University Press, Oxford.

Department of Philosophy
University of Wisconsin-Madison
5113 Helen C. White Hall
Madison, WI 53706
U.S.A.
E-mail: jmcomesana@wisc.edu