SAVING SOSA'S SAFETY

Mark McBRIDE

ABSTRACT: My purpose in this paper is to (begin to) defend safety as a necessary condition on knowledge. First, I introduce Ernest Sosa's (1999) safety condition. Second, I set up and grapple with Juan Comesaña's recent putative counterexample to safety as a necessary condition on knowledge; Comesaña's case forces us to consider Sosa's updated (2002) safety condition. From such grappling a principled modification to Sosa's (2002) safety condition emerges. Safety is safe from this, and like, attacks.

KEYWORDS: Ernest Sosa, safety, knowledge

0.1. My purpose in this paper is to (begin to) defend *safety* as a necessary condition on knowledge. First, I introduce Ernest Sosa's (1999) safety condition.[1] Second, I set up and grapple with Juan Comesaña's recent putative counterexample to safety as a necessary condition on knowledge;[2] Comesaña's case forces us to consider Sosa's updated (2002) safety condition.[3] From such grappling a principled modification to Sosa's (2002) safety condition emerges. Safety is safe from this, and like, attacks.

1. Safety Introduced

1.1. Sosa offered the following first pass at a safety condition on knowledge (time designations suppressed throughout):

> 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

[1] Ernest Sosa, "How Must Knowledge be Modally Related to What is Known?" *Philosophical Topics* 26 (1999): 373-384. Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2000), ch.5, also operates with a (distinct) safety condition on knowledge.

[2] Juan Comesaña, "Unsafe Knowledge," *Synthese* 146 (2005): 395-404.

[3] Ernest Sosa, "Tracking, Competence, and Knowledge," in *The Oxford Handbook to Epistemology*, ed. Paul Moser (New York: Oxford University Press, 2002), 264-286.

© LOGOS & EPISTEME, III, 4 (2012): 637-652

> perhaps not as a matter of strict necessity, S would not believe that p without it being so that p.)[4]

By such a condition's lights, on its supporters' views, we're not prevented from having quotidian knowledge, and nor are we prevented from having knowledge of the falsity of an array of sceptical hypotheses. So far, so good, one might think, for safety.

1.2. Note that safety – a reliability notion – is a squarely *externalist* condition on knowledge. It does not inquire one jot, for example, into a putative knower's recognition of certain epistemically salient facts about the basis, or bases, on which he adopts a particular belief. Rather, for a belief to be safe is simply for a particular modal relation to hold between a subject's belief that p and the fact that p.

1.3. Before coming to Comesaña's putative counterexample to safety, note one (familiar) modification to safety:

> A belief that p by S is safe iff S would not believe that p *on the same basis* without it being so that p.[5]
>
> (Or: S B(p) on basis e → p.)

This modification – as Comesaña points out[6] – is incorporated by Sosa's (2002) updated condition on knowledge, a condition which we'll be focusing on, and modifying, in the remainder of this paper.[7]

---

[4] Sosa, "Modally," 378. Or: S B(p) → p (and not the stronger: □ [S B(p) → p]). Read 'S B(p)' as: "A subject, S, believes that p." Following Sosa, we can read '→' as '*subjunctively implies*': if p → q, "its being so that p offers some guarantee, even if not an absolute guarantee, that it is also the case that q" ("Tracking," 284, n.4). If one formulates safety in terms of a subjunctive conditional one will operate with an account of the semantics of subjunctive conditionals not rendering true-true subjunctives trivially true – cf. Robert Nozick, *Philosophical Explanations* (Cambridge: Belknap, 1981), 680-681, n.8. And, assuming the truth of the relevant subjunctive conditional, any plausible semantics therefor will have the relevant material conditional not coming out false at the actual world.

[5] Comesaña, "Unsafe," 397 (my emphasis).

[6] Comesaña, "Unsafe," 403, n.4.

[7] I assume that all cases considered herein involve beliefs formed on the same basis in the actual and relevant counterfactual circumstances (cf. Timothy Williamson, "Replies to Critics," in *Williamson on Knowledge*, eds. Patrick Greenough and Duncan Pritchard (Oxford: Oxford University Press, 2009), 307.

2. Comesaña's Putative Counterexample: HALLOWEEN PARTY

2.1 Comesaña asks us to consider the following case:

> 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.

And Comesaña's gloss thereon:

> 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 is true. That is, in this case I know that p but my belief that p is not safe – I have unsafe knowledge.[8]

2.2. Ahead of grappling with HALLOWEEN PARTY let's note Sosa's updated safety(-related)[9] principle – a principle motivated in response to cases demonstrating that *outright tracking*[10] isn't necessary for knowledge. Let's introduce it first – laden with heretofore unexplained Sosa-terminology – and explain the terminology by way of applying it to HALLOWEEN PARTY, the case in hand. Here's the updated principle:

> 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

[8] Comesaña, "Unsafe," 397.

[9] I say 'safety(-*related*)' as it does not, unlike the first pass set out at 1.1 *supra*, take the explicit form of a *definition* of safety. It rather takes the explicit form of a (disjunctive) necessary condition on knowledge, though it *should also* be taken as stating a (disjunctive) necessary and sufficient condition for *safe acceptance* – cf. n.21 *infra*. (The same goes, *mutatis mutandis*, for my modification of this principle to come.) Note, moreover, that it employs the more general notion of acceptance rather than belief. (This will not matter, however, for present purposes, as throughout we assume the form of acceptance in question is belief.)

[10] "One tracks the truth, outright, in believing that p IFF one would believe that p iff it were so that p: i.e., would believe that p if it were so that p, and only if it were so." (Sosa, "Tracking," 267)

> indication as such not outright but *guided* by C (so that S accepts the indication as such *on the basis* of C).[11]

The indication, I(p), in HALLOWEEN PARTY, is Judy's testimony to me that the party is down the left road.[12] Disjunct (a) doesn't hold: Judy's testimony doesn't indicate the truth outright as Judy's testimony indicates the truth dependently on the fact that-I-do-not-appear-to-Judy-Michael'ly (C).[13] What is it for an indication to 'indicate[] the truth outright'? This happens iff I(p) → p. That leads us to disjunct (b). The first conjunct of the conjunctive condition contained in disjunct (b) is true. What is it for an indication to 'indicate[] the truth dependently on C'? This happens iff C obtains and [C&I(p)] → p, but ~[I(p) → p]. The second conjunct of the conjunctive condition contained in disjunct (b), however, is false: ex hypothesi I don't accept Judy's testimony as true conditional on the fact that-I-do-not-appear-to-Judy-Michael'ly. As the case is set up, I'll accept Judy's testimony whether or not I appear to her Michael'ly. So I don't accept the indication '*guided* by,' or '*on the basis* of,' C. (If I *do* accept Judy's testimony as true conditional on the fact that-I-do-not-appear-to-Judy-Michael'ly, HALLOWEEN PARTY becomes a straightforward case of safe knowledge.)

And so Sosa's updated (2002) safety principle – as Comesaña notes[14] – cuts no ice against HALLOWEEN PARTY. By its lights we still have unsafe knowledge.

---

[11] Sosa, "Tracking," 275-276. Sosa adds the following disjunct to (b) in his most recent condition on (animal) knowledge based on an indication: "*…or else…C is constitutive of the appropriate normalcy of the conditions for the competence exercised by S in accepting I(p).*" (footnote omitted) (Ernest Sosa, *A Virtue Epistemology, Vol.1* (Oxford: Clarendon Press, 2007), 105.) The candidate C in HALLOWEEN PARTY (to come) does not satisfy this disjunct. Sosa defends this (2007) condition on (animal) knowledge yet now disavows that safety is a necessary condition on (animal) knowledge: the addition of this disjunct must disqualify the principle in question from counting as a *safety* principle (*Virtue*, 92-93). Finally, for more on the basing relation – which features in both the antecedent and consequent of Sosa's (2002) principle –, see Keith Korcz, "The Epistemic Basing Relation," in *The Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, 2006.

[12] Sosa wavers on this ("Tracking"). Transposing things to HALLOWEEN PARTY: at times Sosa takes the indication (or: safe *deliverance*) to be what Judy's testimony *causes* in me, but at other times he takes it to be *Judy's testimony to me itself*. Comesaña – and I will follow suit – goes with the latter interpretation (though I do not think this is crucial). Also, note it is, crucially, "Judy's testimony *to me*." For ease of prose I omit the 'to me' hereinafter, but please read it in.

[13] Because we are interested in Judy's testimony *to a particular subject*, and that subject is me, the C on which I focus is as stated, and not the more general: that-Judy-is-not-appeared-to-Michael'ly.

[14] Comesaña, "Unsafe," 399.

2.3. Let's now grapple with HALLOWEEN PARTY. We need to modify Sosa's (2002) safety principle.[15] My modified principle aims to capture a pre-theoretic notion of beliefs which are safe from danger of being false, just as other objects can be safe from myriad dangers. Moreover, the modification is motivated by scrupulous attention to the externalism which underpinned safety's first formulation. Earlier, we noted the squarely externalist nature of safety's initial formulation: it's merely the positing of a modal relation between a subject's belief and a fact. So we might be suspicious of the internalist flavour to 2.2's updated safety principle: 2.2's principle requires – *modulo* no outright indication of truth – that the putative knower accept the indication in question '*guided* by' and '*on the basis* of C.' I take it that, in order to do this, the putative knower *in HALLOWEEN PARTY* (viz. me) must, at the very least, *recognise* (or: *be aware of*) the condition under which the indication in question indicates the truth.[16] (And so the internalism in question here is *access internalism about grounds*: one must have access to the conditions for a ground (or indication) counting as justified (or safe).)[17] Elseways there is no principled reason for the putative knower in HALLOWEEN PARTY (viz. me) to accept Judy's testimony guided by, or on the basis of, the fact that-I-do-not-appear-to-Judy-Michael'ly.[18]

But why – *modulo* no outright indication of truth – require this for knowledge? It seems that this further requirement, as to the putative knower's recognition-based-acceptance in HALLOWEEN PARTY, is more aptly viewed as a requirement – *modulo* no outright indication of truth –, not on knowledge *simpliciter*, but on knowing that one knows.

---

[15] As a closely-related alternative to my proposal (to come) one might develop an explicitly time-sensitive notion of safety (cf. Mark Sainsbury, "Easy Possibilities," *Philosophy and Phenomenological Research* 57 (1997): 907-919, Christopher Peacocke, *Being Known* (Oxford: Clarendon Press, 1999), 310-328, and Williamson, *Limits*, 124)).

[16] Sosa, "Tracking," 271.

[17] Jim Pryor, "Highlights of Recent Epistemology," *British Journal for the Philosophy of Science* 52 (2001): 106-108.

[18] I restrict my claim here *to HALLOWEEN PARTY* (and like cases). (Compare: If disjunct (a) *were* to hold, even though S *would* know p *on the basis* of an indication I(p), I would not take any internalism to be implicated thereby. This is because I take there to be a fundamental difference (in this regard) between accepting an indication outright (disjunct (a)) and not outright (disjunct (b)) – cf. disjunct (b)(ii) to come.) The form of acceptance required by 2.2's safety principle – *modulo* no outright indication of truth – could, in some cases, be cashed out simply in terms of a modal relation between a subject's acceptance that p and the fact that p (Sosa, "Tracking," 272). But, insofar as my restricted claim is right, we've departed from a purely externalist safety condition. Cf. Sosa's later comments on 'guidance' ("Tracking," 282).

2.4. So the challenge is to set out a modified – more externalist – version of Sosa's updated (2002) safety principle. We need two novel pieces of terminology. First, a schema introducing the notion of a *safe condition*:

> (CSAFE) A condition, C, is safe iff C obtains,[19] and if C were the case in the way described in the thought-experiment under consideration, then C would hold in all[20] close possible worlds.

We'll refer to a safe condition as (a) *CSAFE*. Just as we can talk of the safety of a subject's belief that p – where that is cashed out as a modal relation between a subject's belief that p and the fact that p –, so we can talk of the safety of a condition, C – where that is cashed out in terms of how far into modal space C holds, conditional on C being the case in the way described in the thought-experiment under consideration. That is, in the thought-experiments to come, we suppose the candidate C is the case in the way described in the thought-experiment under consideration, and then, given an intuitive ordering of worlds, check whether that condition, C, holds in all close possible worlds. In what follows, I want to suggest that it's intuitive to add a disjunct to Sosa's safety principle (thereby weakening it) making reference to the safety of candidate *conditions*.

Now here's our modified safety principle:

> 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

---

[19] This functions, in part, to prevent necessarily false conditions from being trivially safe conditions. On standard semantics for subjunctive conditionals necessarily false antecedents make (vacuously) true subjunctive conditionals. Sosa's account of dependent indication (see 2.2 *infra*) itself requires that C obtains. But I prefer an independent obtention requirement on CSAFEs themselves. Finally, by 'obtain' I take it that Sosa means 'obtain *in the actual world*.' That is, in engaging with these thought-experiments (which may, though need not of course, be actual cases), we *suppose* C obtains *in the actual world*, and not in some (remote) possible world which may have *bizarre metaphysics*. This is one reason why the following, admittedly cleaner, safe condition schema will not do: A condition C is safe at a world w iff C holds in all close possible worlds to w.

[20] One might explore alternative formulations, for example replacing 'all' with 'all or nearly all' – cf. Duncan Pritchard's safety account in *Epistemic Luck* (Oxford: Clarendon Press, 2005). It has, though, been noted by John Greco that Pritchard's account may have especial difficulties with the *lottery problem* (cf. n.40 *infra*) ("Worries about Pritchard's Safety," *Synthese* 158 (2007): 299-302). It should be noted that Pritchard has since attempted to amend his account of safety in an effort to respond to Greco's (and others') objections ("Safety-Based Epistemology: Whither Now?" *Journal of Philosophical Research* 34 (2009): 33-45). I don't attempt to adjudicate on this debate here.

> outright, OR (b) either (i) 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),[21] or (ii) for some non-trivial condition CSAFE, I(p) indicates the truth dependently on CSAFE, and S accepts that indication not-as-such outright.[22]

And now our second piece of terminology. Call a CSAFE meeting the requirements of disjunct (b)(ii) (viz. it is non-trivial and I(p) indicates the truth dependently on it) *relevantly-safe* – (a) *CR-SAFE*.[23] And a condition is trivial iff it is, or entails, the putatively known proposition; non-trivial otherwise.[24] In HALLOWEEN PARTY this non-triviality requirement thus rules out conditions such as: that-the-party-is-down-the-(bumpy-)left-road. Note that the (putatively relevantly-safe) condition that-I-do-not-appear-to-Judy-Michael'ly *does not entail* that Judy's testimony that the party is down the left road is true: Judy's testimony could still have been false for any number of reasons (albeit such reasons obtain, ex hypothesi, only in distant possible worlds).

But consider the condition: [p v ~I(p)]. The disjunction as a whole neither is, nor entails, p; and I(p) indicates the truth dependently on the disjunction (by disjunctive syllogism). Objection:[25] To allow this as a CR-SAFE would be to trivialise the notion of CR-SAFEs: *for any p* one could construct a CR-SAFE consisting of the

---

[21] What is now called 'disjunct (b)(i)' must be retained. Though – *modulo* no outright indication of truth – such 'guidance' is no longer *necessary* for *safety*, it is still (stand-alone) *sufficient* therefor (cf. n.9 *supra*): If I *do* accept Judy's testimony as true guided by the condition that-I-do-not-appear-to-Judy-Michael'ly, HALLOWEEN PARTY, we've seen, becomes a straightforward case of safe knowledge (and, arguably, second-order knowledge), even if that *condition* does not, suppose, obtain safely.

[22] That is, S *does* accept the indication *outright*, but *not-as-such* outright, as the indication in question, if disjunct (b)(ii) is to be satisfied, is *not*, ex hypothesi, an outright indication of truth. Finally, it is worth noting that disjunct (b)(i) *can* (though of course need not) be satisfied by a CSAFE. Mutual exclusivity would still be maintained between (b)(i) and (b)(ii) due to the different forms of acceptance involved in satisfaction of the two disjuncts. (For the mutual exclusivity of disjuncts (a) and (b) ((b(ii) in particular), see 2.5 *infra*.)

[23] This second novel piece of terminology is necessary. For example, that-2+2=4 is, and that-my-washing-machine-is-functioning can be, a CSAFE. Without more, these conditions aren't relevant to our inquiry. We need to isolate a proper subset of CSAFEs – CR-SAFEs – in which we're particularly interested.

[24] Comesaña, "Unsafe," 403, n.7.

[25] I close the paper, in 2.7-2.10 *infra*, with four numbered objections to my *fully interpreted* safety principle. This objection, as with the subsequent objection in 2.5, bears on the antecedent matter of *correctly interpreting* my safety principle.

disjunction of p and the negation of an indication that p. One should combat this by making it sufficient for triviality that a *disjunct* of the condition is, or entails, p. Reply: While this objection draws attention to an interesting class of condition, it ignores the fact that being *C*$_{SAFE}$ is a prerequisite for being C$_{R\text{-}SAFE}$. However, it will follow that on any occasion in which [p v ~I(p)] *is* C$_{SAFE}$ it will also be C$_{R\text{-}SAFE}$. To the extent that this is a problematic result – something on which I do not here commit –, we will need to modify our definition of triviality in line with this objection.

2.5. This modified safety principle is a move towards the externalism which motivated initial (1999) formulations of safety, and dispenses with the internalist flavour of subsequent (2002) formulations. (Recall: 2.2's principle requires – *modulo* no outright indication of truth – that the putative knower accept the indication in question '*guided* by' and '*on the basis* of C.' My modified safety principle rejects this requirement.) My claim here is only this: Insofar as one is interested in defending safety as a necessary condition on knowledge, why not see how far one can get with a more externalist account thereof? After all, as noted, initial formulations of safety were (purely) externalist.

Does this modified safety principle, however, handle HALLOWEEN PARTY? Do we get the result that I gain knowledge of the whereabouts of the party from Judy's testimony – chiming with our intuitions – with the belief on which such knowledge is based rendered safe by dint of fulfilment of disjunct (b)(ii)? To answer these questions we first, obviously, assess this (more externalist) safety condition's success in handling HALLOWEEN PARTY. But our enquiry should not rest there. We'll then move on to consider its plausibility (in general) by considering some objections thereto.

And so to HALLOWEEN PARTY itself and the candidate condition that-I-do-not-appear-to-Judy-Michael'ly. It's plausible that if this C were the case in the way described in HALLOWEEN PARTY – at which point in time, ex hypothesi, crucially my *decision has been made* not to disguise myself as Michael – Judy won't be appeared to Micheal'ly by me in any close possible worlds. I take it we should read such a decision into HALLOWEEN PARTY; elseways how do we explain my move from 'seriously considering disguising myself as Michael' to – 'at the last moment' – not doing so?[26] To be sure, there are remote worlds in which, even after the decision has been made not to disguise myself as Michael, I end up

[26] Cf. Joseph Raz, *Practical Reason and Norms*, 2nd ed. (Oxford: Oxford University Press, 1999), 65. Comesaña ("Unsafe," 399) reads such a decision in. This suggests a candidate (complementary) C$_{R\text{-}SAFE}$: that-I-*decide*-not-to-appear-to-Judy-Michael'ly (see 2.9 *infra*).

disguising myself as Michael.[27] But, provided we stick to the case as set up, these 'disguising myself as Michael'-worlds will not be close enough to threaten the safety of my true belief that the party is down the left road.

It is the element of *prior decision* – reached, I take it, as a result of deliberation on the reasons for or against the action in question; with decisions themselves terminating that deliberation and being reasons[28] – which distinguishes HALLOWEEN PARTY from ensuing cases we'll consider. At a more general level, a condition will be CSAFE if[29] there is some (*non-luck-infected*)[30] factor – whether a mental act, as in HALLOWEEN PARTY, or not – which predates the putatively safe condition, and serves to *secure* that condition's holding in all close possible worlds. So this candidate condition is CSAFE. Moreover, we saw in 2.2 that Judy's testimony indicates the truth dependently on this (non-trivial) CSAFE. So it's a relevantly-safe condition: it's CR-SAFE. We thus have disjunct (b)(ii) of 2.4's modified safety principle being met. We, untroublingly, have safe knowledge in HALLOWEEN PARTY.

There is, however, a complication here relating to how an indication can indicate the truth dependently on a CSAFE. Or, put differently: how a CSAFE can be a CR-SAFE. Objection: For Sosa, we've seen, an indication indicates the truth dependently on a condition iff C obtains and [C&I(p)] → p, but ~[I(p) → p]. But if C is a CSAFE, (*a fortiori*) obtains, and [C&I(p)] → p, that seems to entail that: [I(p) → p]. Reply: However this is not so. Though there is, at root, one question to be determined in HALLOWEEN PARTY – viz. do I possess knowledge? –, two 'contexts of thought or discussion' are 'relevant' at different stages of enquiry into that question.[31] At the first stage of enquiry – determining whether the condition in question is CSAFE – schema (CSAFE) *makes salient* the way in which the condition

[27] There are also worlds – I take it remote, if it were the case that-I-do-not-appear-to-Judy-Michael'ly in the way described in HALLOWEEN PARTY – in which I *don't decide* not to dress as Michael. I am not, by diktat, holding *that decision* fixed *across all worlds*.

[28] These remarks are taken from Raz ("Reasons for Action, Decisions, and Norms," in *Practical Reasoning*, ed. Joseph Raz (Oxford: Oxford University Press, 1978), 135, *PRN,* 65-72). For Raz, "a decision is always, for the agent, a reason for performing the act he has decided to perform and for disregarding further reasons and arguments. It is always both a first-order and an exclusionary reason" (*PRN*, 66). Consistently with this, "in most cases the refusal to reopen the case is not absolute" (*PRN*, 67). Cf. also Joseph Raz, *Engaging Reason* (Oxford: Oxford University Press, 2002), ch.1.

[29] I am not committed to the 'only if' claim.

[30] I leave this notion intuitive, but for an extended analysis of epistemic luck, see Pritchard, *Luck*. It is omitted in what follows, as only non-luck-infected factors can secure the holding of conditions in all close possible worlds.

[31] Sosa, "Tracking," 271.

came about in the thought-experiment under consideration. At the second stage of enquiry – determining whether the condition in question satisfies disjunct (b)(ii) – the foregoing feature of the condition is *not* rendered salient: Sosa's formulation of when an indication indicates the truth dependently on a condition, of course, makes no reference to C_{SAFEs}. It is only by recognising these two different contexts within a single project of enquiry that we can pay due deference to the initial intuitive pull towards thinking of HALLOWEEN PARTY as a case of unsafety – recognising, that is, that I *could very easily* (in some context of thought or discussion) have disguised myself as Michael. And this will be a general feature of applying my safety principle.[32] Thus, in HALLOWEEN PARTY we assess whether the condition that-I-do-not-appear-to-Judy-Michael'ly satisfies disjunct (b)(ii) *not* building in information about precisely how that condition came about in the thought-experiment (i.e. via a prior decision). Given this, the foregoing entailment does not hold and, plausibly: $\sim[I(p) \rightarrow p]$. To fail to adopt this approach, Judy's testimony would end up indicating the truth outright (*modulo* my reading of HALLOWEEN PARTY). (And, more generally, to fail to adopt this approach, condition (b)(ii) of my safety principle would be unsatisfiable, with my proposal boiling down to Sosa's updated principle.) While the result would still be safe knowledge, by my reckoning something important would be lost in describing the case this way. Overall, this complication demonstrates the fine line between outright and dependent indications of truth.[33]

2.6. Now, as a preamble to considering objections, let's distinguish two epistemological projects one might undertake. First, one might attempt to defend safety as a necessary condition on knowledge. This is my project in this paper. Second, and more ambitiously, one might attempt to give a *reductive analysis* of knowledge, with safety as a component part – perhaps: all and only safe true beliefs count as knowledge. For familiar reasons, any such reductive analysis fails to have the resources to account for knowledge of necessary truths.[34] More prosaically, insofar as Kelly Becker's case,[35] in which a person believes that the earth revolves around the sun solely on the basis of his adherence to a religion in which the sun is worshipped, is non-knowledge such a reductive analysis would fail on this score too. But note, such an analysis is not vulnerable to Sherrilyn

[32] To recognise the foregoing is not, I take it, to perforce become an *epistemic contextualist* – see Patrick Rysiew, "Epistemic Contextualism," in *The Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, 2007.

[33] Sosa, "Tracking," 270-271.

[34] And for less familiar problems with such a reductive analysis, see David Manley, "Safety, Content, Apriority, Self-Knowledge," *The Journal of Philosophy* 104 (2007): 408.

[35] Kelly Becker, "Reliabilism and Safety," *Metaphilosophy* 37 (2006): 691-704.

Roush's FAIRY GODMOTHER case[36] of putative safe non-knowledge, in which a fairy godmother – let's say, of nomological necessity – renders true, for any p, S's belief that p, however faulty S's mode of reasoning in coming to believe that p. Recall (from n.4), our formulation of safety using a subjunctive conditional was: S B(p) → p. It wasn't the stronger: □ [S B(p) → p]. As such, we can – without complication – rely on the non-obtaining of fairy godmothers in close possible worlds.

The more ambitious project of reductive analysis, however, is not my project here. Insofar, then, as other safety accounts *can* successfully undertake this more ambitious project, my project might seem unduly *unambitious*. But my project would only be *mistaken* should my safety condition not feature as a necessary component of the reductive analysis. (For other accounts which *might* be thought to provide the basis for a reductive analysis – accounts which are not in competition with, and indeed may need to be supplemented by, my account – cf. *method safety/process reliabilism* and *virtue reliabilism*. Each of these alternative accounts is, however, vulnerable to objections – most notably, perhaps, the *generality problem*.) Still, insofar as we follow Sosa[37] in considering safety an advance on *sensitivity*,[38] and insofar as the sensitivity condition allowed for progress on the Gettier problem,[39] it would be troubling for my proposed safety condition if one could readily cook up Gettier-style cases of safe (true beliefs which are) non-knowledge. Any putative Gettier-style cases – see objections 1 and 2 (to come) – of safe non-knowledge should be accommodated by my project.[40]

2.7. Objection 1 and Reply 1: Suppose Judy flips a coin in a situation like HALLOWEEN PARTY but absent the 'that-I-do-not-appear-to-Judy-Michael'ly' condition. Instead, if the coin comes up tails, she'll direct me down the left road to the party at Andy's; if it comes up heads, she'll direct me down the left road to Andy's, but will immediately phone Andy so the party can be moved to Adam's. Call this JUDY COIN-FLIP. Suppose the coin lands tails. Do I know that the party

[36] Sherrilyn Roush, *Tracking Truth* (Oxford: Oxford University Press, 2005), 122-123.

[37] Sosa, "Modally."

[38] Viz.: If p weren't true, S wouldn't believe that p via M. (Or: ~p → ~[S B(p) via M].) This is a (Nozick-inspired) refinement on Nozick's 'condition (3)' (*Explanations*, 172).

[39] Edmund Gettier, "Is Justified True Belief Knowledge?" *Analysis* 23 (1963): 121-123.

[40] I am content to classify the classic *lottery case* – in which one truly believes one's single ticket in, say, a million-ticket lottery loses – as unsafe non-knowledge: although the odds of winning the lottery are minuscule, there are close possible worlds in which one wins. Space prevents detailed defense of this classification.

is at the house down the left road?[41] It seems that I don't know this. Is my safety condition met? Suppose the candidate C$_{\text{R-SAFE}}$ here is: that-Judy-is-not-appeared-to-heads'ly. Is this C *indeed* safe? If this C were the case in the way described in JUDY COIN-FLIP, would C hold in all close possible worlds? No. That the flipped coin lands tails in our case has no (strong) bearing on what way the coin lands in close possible worlds; in particular, that the flipped coin lands tails in our case does not make it the case that the coin lands tails in all close possible worlds. And so we don't have a case of safe non-knowledge. Rather, it's, untroublingly, unsafe non-knowledge.[42]

HALLOWEEN PARTY – as with nearly all thought-experiments – is, of course, under-described. Clearly I am making mileage out of a prior decision in HALLOWEEN PARTY securing C's (viz. that-I-do-not-appear-to-Judy-Michael'ly) holding in all close possible worlds. But suppose – as Comesaña does[43] – that the decision not to disguise myself as Michael was formed – as is, concededly, left open by HALLOWEEN PARTY – on the basis of a coin-flip landing tails (or conditional on my one ticket winning a million-ticket lottery). Call this PARTYGOER COIN-FLIP. Suppose the coin lands tails (or I win said lottery). Now, it's not so that if C were the case in the way described in PARTYGOER COIN-FLIP, C would hold in all close possible worlds. Result (*pace* Comesaña):[44] more of an intuitive pull to withhold knowledge. We have unsafe non-knowledge (as in JUDY COIN-FLIP).

Summary diagnosis: In all the cases we've considered so far there's *some* (however weak) initial intuitive appeal to ascribe knowledge – after all, all the cases have a source of knowledge (whether testimony or perception) operating successfully. As we fill in the cases it becomes clear that the relevant source only operates successfully dependently on some or other (non-trivial) condition being

[41] This case is found in Comesaña ("Unsafe," 402). And one could construct a similar case in which Judy tells the truth conditional on it being the case that-Judy's-one-ticket-wins-a-million-ticket-lottery, and her ticket in fact wins said lottery.

[42] I give a like diagnosis, *mutatis mutandis*, of Alvin Goldman's FAKE BARNS ("Discrimination and Perceptual Knowledge," *The Journal of Philosophy* 73 (1976): 771-791), and Ram Neta and Guy Rohrbaugh's two cases ("Luminosity and the Safety of Knowledge," *Pacific Philosophical Quarterly* 85 (2004): 396-406). (To the extent that denying Neta and Rohrbaugh's cases involve knowledge is a bullet, I am prepared to bite it – cf. n.45 *infra*.) Though note the following putative difference between FAKE BARNS and Neta and Rohrbaugh's cases: the threat to knowledge in FAKE BARNS is *actual* – there really are fake barns around – whereas the threat in Neta and Rohrbaugh's (as in HALLOWEEN PARTY) is *purely counterfactual*.

[43] Comesaña, "Unsafe," 402.

[44] Comesaña, "Unsafe," 402.

the case. And the relevant condition, in each thought-experiment, might – it seems – very well not have been the case. Now we have an intuitive pull to withhold knowledge. As we fill in the cases further we discover – my contention – that our willingness to ascribe knowledge in this or that case is a function of whether or not the relevant condition, if it were the case in the way described in the thought-experiment under consideration, holds in all close possible worlds. In other words, it's a function of whether the relevant condition, C, is *safe*.

Indeed, on the back of this summary diagnosis, I'm open to persuasion – *contra* my initial diagnosis of HALLOWEEN PARTY at 2.5 *supra* – that, in HALLOWEEN PARTY, the condition that-I-do-not-appear-to-Judy-Micheal'ly is *not* safe. More descriptive information about the case pointing in this direction could come to light. Moreover, orderings of modal space are contentious. If this condition is not after all safe, discovery that it is not safe will, I suggest, be matched by – will generate – an intuitive pull to withhold knowledge.[45] We'd, untroublingly, have unsafe non-knowledge.

Throughout, I – following most leading proponents of safety – rely on an intuitive ordering of possible worlds and do not commit on any substantive account of orderings of possible worlds (such as David Lewis's).[46] Clearly, this leaves room for disagreement over whether a condition is safe (e.g. on account of context dependence and/or vagueness infecting the relevant subjunctive conditional which is being given a possible worlds analysis). But perhaps this is exactly what we should expect in hard cases.[47] It must be conceded, however, that it is the very fact that modal orderings are contentious which leads some philosophers to give accounts of knowledge which do not use modal conditions at all.

2.8. Objection 2: My proposal trivialises the safety condition, for almost every true belief will, on this objection, turn out to be safe. Consider, for instance, PARTYGOER COIN-FLIP, and grant that the condition that-I-do-not-appear-to-Judy-Michael'ly is not C$_{R\text{-}SAFE}$. That doesn't by itself show that the belief in question isn't safe, for there may be other C$_{R\text{-}SAFE}$s relative to which the belief is safe. In this case, let the candidate condition be: that-the-party-is-at-Andy's-

[45] For a contrasting strategy to that adopted in this paper, see Williamson ("Critics," 305): "One may have to decide whether safety obtains by first deciding whether knowledge obtains, rather than vice-versa." Sloganistically, Williamson's is a 'knowledge first' strategy; mine (at least in hard cases) a 'safety first' strategy.

[46] David Lewis, "Counterfactual Dependence and Time's Arrow," *Nous* 13 (1979): 455-476.

[47] Cf. Tamar Gendler and John Hawthorne on the putative instability of knowledge-intuitions in hard cases ("The Real Guide to Fake Barns: A Catalogue of Gifts for your Epistemic Enemies," *Philosophical Studies* 124 (2005): 331-352).

house. This condition is, on this objection, CR-SAFE. That-the-party-is-at-Andy's-house doesn't *entail* that the party is at the house down the left road, and thereby counts as non-trivial. And Judy's testimony does indicate the truth dependently on this condition. But we've classed PARTYGOER COIN-FLIP as a case of intuitive non-knowledge.[48]

Reply 2: But the condition that-the-party-is-at-Andy's-house is not (relevantly-)CSAFE. It's not the case that, if this C were the case *in the way described in PARTYGOER COIN-FLIP*, the party would be at Andy's house in all close possible worlds. The party is only at Andy's house in PARTYGOER COIN-FLIP thanks to a coin-flip landing tails (or my winning said lottery). And if, by contrast, this C is *stipulated to be* (relevantly-)CSAFE, the case is changed beyond all recognition and I don't see that the resultant case would be a genuine Gettier-case. That is, suppose, for contrast, the party *is* at Andy's house in all close possible worlds. *Now* is my belief that the party is at the house down the left road a clear case of non-knowledge? I don't think so.[49,50]

2.9. Objection 3: My proposal does not tell us *how to find* CR-SAFEs. Perhaps we're better off with Sosa's original proposal that – *modulo* no outright indication of truth – the putative knower must accept the indication '*guided* by,' or '*on the basis* of,' C. (Sosa's original proposal, though, is, of course, vulnerable to Comesaña's HALLOWEEN PARTY counterexample.)

Reply 3: I agree that no algorithm for finding CR-SAFEs is on offer. But: so what? I take it 2.4's safety principle states a (disjunctive) necessary condition on knowledge. It doesn't have epistemic pretensions to furthermore help us *identify* CR-SAFEs. Identifying such conditions is for (common-sense, philosophical) judgment to do (though this is not to say such identification will always be easy).

---

[48] This objection would putatively generalise to Gettier-cases like Keith Lehrer's NOGOT AND HAVIT ("Knowledge, Truth and Evidence," *Analysis* 25 (1965): 168-75), in which the subject's belief that *someone* in his office owns a Ford is safe dependently on the putative CR-SAFE: that-*Havit*-owns-a-Ford. Again: that-Havit-owns-a-Ford doesn't entail that someone in the subject's office owns a Ford; only that-Havit-*who-is-in-the-subject's-office*-owns-a-Ford entails that.

[49] And in NOGOT AND HAVIT, the condition that-Havit-owns-a-Ford is, for all we're told in that case, not (relevantly-)CSAFE. If it's *stipulated to be* CSAFE, it's less clear we have a genuine Gettier-case of non-knowledge – cf., *inter alia*, Peter Klein, "Useful False Beliefs," in *Epistemology: New Essays*, ed. Quentin Smith (Oxford: Oxford University Press, 2008), 25-61, for the possibility of knowledge inferred from falsehoods. As noted in 2.6, though, I don't claim to have set out a 'Gettier-proof' safety condition.

[50] Is the condition that-the-party-is-at-Andy's-house a candidate (complementary) CR-SAFE *in HALLOWEEN PARTY*? To answer this, we need more information about the likelihood of Michael himself (and any other potential 'Michael-disguiser,' such that there be) talking to Judy at the crossroads.

2.4's safety principle is none the worse for leaving this epistemic task to judgment. Try plugging some non-trivial conditions into the relevant subjunctive conditional and then evaluate it. We might be pleasantly surprised – I conjecture – by the paucity of conditions – none? one? *just* more than one? – which turn out to be CR-SAFEs in this or that case.[51]

2.10. Objection 4: Whether a condition counts as (relevantly-)safe depends on how the condition and the facts that pre-date the condition are described. In JUDY COIN-FLIP, for example, the condition that-Judy-is-not-appeared-to heads'ly does not seem to be safe, and (as a result) it is a case of unsafe non-knowledge. But what prevents us from describing the relevant condition as the condition that-Judy-is-not-appeared-to-heads'ly-given-the-fact-that-the-coin-lands-tails? This fact pre-dates the condition and, on this objection, guarantees that the condition holds in all close possible worlds. Using such a description, JUDY COIN-FLIP would come out as a case of either safe non-knowledge (which is troubling for my project) or safe knowledge (which is counterintuitive).

Reply 4: Objection 4 describes, not two different descriptions of one condition, but rather two different conditions – two ways of picking out different features of the world. Given a way close worlds are, we can fully expect two different conditions – two ways of picking out different features of the world – to differ in whether or not they're (relevantly-)safe.[52] As it happens, here, on a correct construal of the new condition, it shares the property of the condition in JUDY COIN-FLIP of failing to be safe (and so failing to be relevantly-safe), and thus the difficulties which would have arisen had we had a case of safety do not arise. (On a mistaken construal, we'll see, the new condition has different properties.)

Let me explain. The logical form of the new condition is, abbreviating, the following conditional: T ⊃ ~APP H.[53] According to our safe condition schema (of 2.4 *supra*), to be safe a condition must 'obtain,' and 'hold' in all close possible worlds. To do this, a conditional must be non-vacuously true throughout these

[51] Some CR-SAFEs – in the event of there being more than one in a particular case – will, however, be *explanatorily superior* to others.

[52] Beyond the claim that if one has two ways of picking out different features of the world one has two different conditions, I don't commit on more substantive individuation criteria for conditions – that is, criteria for telling one numerically distinct condition from another. More specifically, I don't commit on whether Leibniz's law – the principle of the Indiscernibility of Identicals – holds for the *modal* property of (relevant-)safety (or the *logico-linguistic* property of logical form considered in the next paragraph). (Even more plainly, I don't need to commit on the status of the principle of the Identity of Indiscernibles.)

[53] T = the-coin-lands-tails; ~APP H = Judy-is-not-appeared-to heads'ly.

worlds. And, while this conditional will not be false in any close possible worlds, it will go vacuously true – the coin will land heads – in some. We cannot, by diktat, stipulate that the coin lands tails in all close possible worlds: we are beholden to modal space. This condition, thus, is not safe. (If, mistakenly, one took non-falsity in all close possible worlds to be sufficient for a conditional to be a safe condition, this conditional, while safe, will not be *relevantly*-safe – consider the close worlds in which it goes vacuously true.)

Having said all this, let me concede that it *may be* that whether a condition counts as (relevantly-)safe *can depend* on how the condition is described. Return, for example, to HALLOWEEN PARTY. And suppose, with me, that the condition that-I-do-not-appear-to-Judy-Micheal'ly is relevantly-safe. But now also suppose that Judy happens to be the tallest person invited to Andy's party. On one plausible way of individuating conditions, the condition that-I-do-not-appear-to-the-tallest-person-invited-to-Andy's-party-Micheal'ly is the *same condition* as the one we've classed as relevantly-safe – it picks out the same features of the world – *just newly described*. But, equally plausibly, the newly described condition may fail to be (relevantly-)safe (cf. n.51 *supra*). But, even if all this is so: so what? A given belief will count as *safe* if there is *some description* of a condition under which the condition in question counts as relevantly-safe.

3. Conclusion

3.1. I haven't conclusively demonstrated that (2.4's) safety is a necessary condition on knowledge. I have, though, dismissed some cogent objections thereto.[54]

[54] Thanks to Lee Walters for stimulating discussion.