Anti-Luck Epistemologies and Necessary Truths*

Jeffrey Roland and Jon Cogburn

Forthcoming in *Philosophia*

Abstract

That believing truly as a matter of luck does not generally constitute knowing has become epistemic commonplace. Accounts of knowledge incorporating this anti-luck idea frequently rely on one or another of a safety or sensitivity condition. Sensitivity-based accounts of knowledge have a well-known problem with necessary truths, to wit, that any believed necessary truth trivially counts as knowledge on such accounts. In this paper, we argue that safety-based accounts similarly trivialize knowledge of necessary truths and that two ways of responding to this problem for safety, issuing from work by Williamson and Pritchard, are of dubious success.

Introduction

Sam walks by Big Ben at three o'clock, looks up at its face, and thinks, "It is now three o'clock." However, unbeknownst to Sam, Big Ben malfunctioned and stopped precisely twelve hours ago. At least since Gettier (1963), such examples have been taken to highlight that true beliefs acquired by luck fail to be knowledge.[1] Sam believes truly that it's three o'clock, but that his belief is true is a matter of luck. In response, epistemologists have attempted to characterize knowledge so as to rule out lucky acquisition of true belief. Alvin Goldman's causal theory of knowledge and various versions of reliabilism are prime exhibits of these attempts. More recently, so-called *anti-luck* epistemologies have come into their own in the work of Fred Dretske, Robert Nozick, Ernest Sosa, Keith DeRose, Duncan Pritchard, and others.

Anti-luck approaches to knowledge typically divide into those that endorse a sensitivity-based account of knowledge and those that endorse a safety-based account of knowledge.[2] Safety-based

*We would like to thank James Rocha and an anonymous referee for helpful comments. An earlier version of this paper was presented at the 2009 meeting of the Alabama Philosophical Society. Thanks to the participants of our session there for helpful discussion.

[1] Russell arguably called attention to this issue (1912, XIII).

[2] Sensitivity-based accounts are defended in Dretske (1971) and Nozick (1981). Safety-based accounts are defended in DeRose (1995, 2004), Sosa (1996, 1999, 2000), Pritchard (2005, 2007, 2009), and Williamson (2000). For an extended treatment of the role of luck in knowing, see Pritchard (2005).

accounts currently appear to have the upper hand in this debate. Our aim in this paper is the modest one of showing that a problem for sensitivity-based accounts of knowledge concerning necessary truths is equally problematic for safety-based accounts and that two key attempts to address this problem for safety-based views at best face significant challenges.

1 Sensitivity

Intuitively, S's true belief $\langle p \rangle$ is sensitive just in case S wouldn't have believed $\langle p \rangle$ had it been false $\langle p \rangle$.[3] Sensitivity-based accounts of knowledge hold that knowledge is sensitive true belief. Consider Sam's case. Sam's belief that it's three o'clock fails to be sensitive: had it not been three o'clock when Sam walked by Big Ben (say at 2:55) and saw it showing three o'clock, he still would have believed that it was three o'clock. So even though Sam has a true belief that it's three o'clock, on a sensitivity-based account of knowledge he doesn't know that it's three o'clock.

According to sensitivity-based accounts, the counterfactual states of affairs encoded in the sensitivity condition explain why Sam does not know that it is three o'clock. Sam is lucky in that, though he acquired a true belief, he might easily have acquired a false belief, and he still would have held it. Compare the case of Ernie, who throws darts at a map to decide on which city block to look for a dollar bill. If Ernie finds a dollar using this process, we wouldn't say that he's good at finding dollars; we'd rather say that he was lucky to the find the dollar. It would have been easy for Ernie to go looking where the dart said he should and not find a dollar there. Ernie got lucky. Similarly, it's too easy for Sam to believe falsely that it's three o'clock using the same belief-forming process he actually used to form the true belief. Sam got lucky. Hence, he doesn't know.

Contrast this with the case where Big Ben is working properly. Now Sam's belief that it's three o'clock is sensitive: had it not been three when Sam looked at Big Ben, Big Ben wouldn't have read three o'clock, and Sam would have believed that it was whatever time Big Ben (correctly) read. So Sam has a true sensitive belief that it's three o'clock, and on a sensitivity-based account of knowledge he knows that it's three o'clock.

Sensitivity-based accounts hold that the counterfactual states of affairs encoded in the sensitivity condition explain why, in the working-clock scenario, Sam does know that it is three o'clock. The

[3]Here, 'p' is a variable ranging over declarative sentences and '$\langle p \rangle$' denotes the proposition expressed by p. The latter can be read 'that p'.

belief was not lucky because the process by which he formed the his time-belief (looking at a working, accurate clock) tracks truth with respect to time across a relevant set of possible worlds. Analogously, if Ernie finds a dollar bill using a dog that can sniff out the cocaine residue on paper currency, Ernie's finding the dollar isn't a matter of luck. This process reliably leads to finding dollar bills; it tracks dollar bills across a sufficiently robust set of possible worlds.

Recall the semantics of counterfactuals advanced by David Lewis:[4]

> *Standard Semantics for Counterfactuals* A conditional of the form, "If it were the case that p, then it would be the case that q," is true in a conversational context C and a world w just when every (C, w)-closest $\langle p \rangle$-world is a $\langle q \rangle$-world.

Here '(C, w)-closest' is read as 'most relevantly similar to w given the conversational context C.' We take these to be the standard semantics for counterfactuals. Using these semantics we can formulate a

> *Sensitivity-based Account of Knowledge* S knows $\langle p \rangle$ if and only if:
>
> (Sen1) S believes $\langle p \rangle$;
>
> (Sen2) $\langle p \rangle$ is true;
>
> (Sen3) Were $\langle p \rangle$ false, S would not believe $\langle p \rangle$.

Here (Sen3) is treated according to the standard semantics for counterfactuals.

That closeness of worlds has to be cashed out in terms of similarity according to some relevant metric tied to conversational context is clear from examples.[5] Consider:

(1) If Caesar had commanded in Iraq, then he would have used nuclear weapons.

(2) If Caesar had commanded in Iraq, his troops would have built gigantic walls around the cities, trapping his enemies within.

If the context makes salient Caesar's ruthlessness, (1) comes out true and (2) false. If the context makes salient the actual strategies he followed in pacifying Gaul, then (2) comes out true and (1) false.

[4] See (1973; 1986).

[5] Our remarks on the need for context relativity, both formulations of the standard semantics for counterfactuals, and the problem of counterpossibles are informed by the presentation in Brogaard and Salerno (forthcoming).

When assessing a belief for sensitivity, it's important not to alter the way in which the agent arrives at (or, if necessary, sustains) the belief. In neither of the above examples do we alter Sam's belief-forming circumstances by having him do anything other than look at Big Ben and credulously believe what it says. The idea is to maintain the epistemically relevant features of the agent's situation as much as possible as we move from world to world. If we don't do this, then sensitivity will have little claim to capturing the anti-luck intuitions it's intended to capture. We codify this idea in the

> *Constancy Principle* When considering changes in an agent's doxastic attitude with respect to a proposition across worlds, the process/method by which the agent acquires (or sustains) that attitude must remain constant.

This potentially raises as many questions as it settles. If there is any concern about what counts as a good metric on a space of possible worlds, individuating belief-forming methods is probably also a matter of concern. Indeed, individuating such methods involves us with the infamous and difficult *generality problem*, most closely associated with process relibilism. That said, since we are not defending these accounts the assumption that they can help themselves to a theory that individuates such procedures is being charitable.[6]

2 A Problem for Sensitivity

Suppose that Sam returns home later to do some work. Little does he know that his neighbor Bertie's experiments with plasma have damaged the circuit board of his calculator so that it always tells the user that whatever number entered is prime. Sam uses his calculator to randomly check whether or not 131,071 is prime, and it answers affirmatively. As a result, Sam forms the true belief that 131,071 is a prime number. This is analogous to the Big Ben example. And if one accepts that an agent can be justified in believing a necessary truth without proving the truth in question, then this is a case of an agent being justified in believing a necessary truth as a matter of luck. (The number Sam randomly enters just happens to be a prime.)

[6]Pritchard has defended this need to hold epistemic processes constant. See (2005, Chapter 6). See also Juan Comesaña's *same-basis safety* in (2005).

Outside of mathematics and logic, where the gap between justification and truth is greater, it is easy to come up with less controversial examples of this sort involving necessary truths. Assume that Sam is credulously looking through an American book chain's "metaphysics" section. As he glances at the various volumes about reincarnation, astrology, telepathic powers and whatnot, he has no idea that he is viewing knowledge façades. He randomly picks up a book on homeopathic treatment and reads that water is H_2O, thereby forming the true, necessary belief that water is H_2O. It's a matter of luck that Sam picked up a book with that accurate chemical information in it rather than one of the many books on the same shelf filled with falsehoods. (Compare the famous Ginet–Goldman fake barns case in Goldman (1976).)

In both of these cases, Sam lucks into a necessarily true belief $\langle p \rangle$. Sensitivity-based accounts of knowledge have difficulty accommodating these types of cases. When $\langle p \rangle$ is necessarily true, sensitivity-based accounts vacuously yield that a belief $\langle p \rangle$ is knowledge. There are no possible worlds where 131,071 is not a prime number or where water fails to be H_2O. One way to think about this problem is to note that on the standard semantics for counterfactuals "if it were not the case that p, then S would not believe that p" is trivially satisfied whenever $\langle p \rangle$ is a necessary truth. Thus, on the standard semantics, sensitivity-based accounts entail that all believed necessary truths are known. Sensitivity makes knowledge of necessary truths cheap, easy, and universal—a fact that has been used to argue against sensitivity-based accounts of knowledge.[7]

This problem with sensitivity-based accounts is a particularly problematic instance of what has become known as the *problem of counterpossibles*. A counterpossible is a counterfactual conditional with a necessarily false antecedent. The problem is that many counterpossibles seem to be false, contrary to the uniform prediction of the standard semantics for counterfactuals. For example:

> If Hobbes had managed to square the circle, then *Leviathan* would have been a more important book than *The Critique of Pure Reason*.

So the defender of sensitivity-based accounts needs to adopt a variant semantics for counterfactuals, or abandon application of his account to beliefs with necessarily true content. Nozick has taken the latter option.[8] But one who wants an account of knowledge which uniformly applies to contingent

[7] See, e.g., Sosa (1999).

[8] Sosa attributes this move to Nozick in (1999, p. 146).

and necessary propositions must choose the former. For this case, the following reformulation of the standard semantics for counterfactuals makes the task clear.

> *Reformulated Standard Semantics for Counterfactuals* A conditional of the form, "if it were the case that p, it would be the case that q," is true in context C and world w just in case either (i) there is no $\langle p \rangle$-world, or (ii) there is a $\langle p \wedge q \rangle$-world that is (C, w)-closer to the actual world than any $\langle p \wedge \neg q \rangle$-world.

Again, '(C, w)-closer' is read as 'most relevantly similar to w given conversational context C.' Now we can see what the sensitivity theorist needs to do to save her account from triviality when considering necessary truths. First, she needs to make sense of worlds where necessary falsehoods are true and, second, she needs to make sense of relevant similarity between worlds in a way that respects the Constancy Principle.

3 Safety

If the difficulty with sensitivity-based accounts of knowledge just outlined seems to receive little attention, it is almost certainly because there is a widespread conviction that anti-luck epistemic intuitions can be adequately captured by a theory that does not have to confront this or similar problems, by a so-called *safety-based* account of knowledge. Intuitively, S's true belief $\langle p \rangle$ is safe if and only if it's not the case that S's belief $\langle p \rangle$ could easily have been false; i.e., in worlds that don't radically differ from S's actual world, S believes $\langle p \rangle$ only if $\langle p \rangle$ is true. This yields the following

> *Safety-based Account of Knowledge* S knows $\langle p \rangle$ if and only if:
>
> (Saf1) S believes $\langle p \rangle$;
>
> (Saf2) $\langle p \rangle$ is true;
>
> (Saf3) were S to believe $\langle p \rangle$, $\langle p \rangle$ would be true.[9]

Consider Sam's case again. Sam's belief that it's three o'clock is unsafe. Let w be a world that differs from $@_{Sam}$[10] only in that Sam walks by Big Ben five minutes before three o'clock. In w, Sam

[9] This rendering, which typically serves as the jumping off point for contemporary discussions, is essentially that found in Sosa (1999). In §5, we consider accounts based on Williamson's and Pritchard's formulations of safety.

[10] Here '$@_{Sam}$' denotes Sam's actual world. In general, '$@_S$' denotes an agent S's actual world.

still believes that it is three o'clock. But in w the content of Sam's belief is false. Hence, according to the safety-based account of knowledge, Sam doesn't know that it is three o'clock. Sam's belief, formed using the way in question (looking at the broken clock and deferring to it), does not hold true across a robust enough set of worlds.

On the other hand, consider a case where Big Ben is working properly. Now if we consider any world w where Sam still believes that it's three o'clock and which is relevantly similar to $@_{Sam}$ according to the Constancy Principle, Sam's belief is true at w (since by the Constancy Principle the way Sam comes to believe is held constant). A world in which Sam still believes that it is three o'clock and it is not three o'clock would have to violate the Constancy Principle. According to the safety-based account of knowledge, the counterfactual states of affairs encoded in the safety condition (Saf3) explain why, in the working clock scenario, Sam knows that it is three o'clock. Again, Sam's belief was not lucky because the process by which he arrived at his time-belief (looking at a working, accurate clock) produces beliefs that hold true across a robust relevant set of possible worlds.

4 Problems for Safety

As we already noted, safety theorists have criticized sensitivity-based accounts of knowledge on the basis of the problem such accounts have with necessary truths. Sosa, for example, writes that "[S]ensitivity is doubtful as a condition for our being correctly said to have knowledge of any apodictically necessary truth A, given how hard it would be to make sense of the supposition that not-A" (Sosa, 1999, p. 6). This criticism constitutes an indirect argument for the superiority of safety-based accounts, provided that they don't have the problem with necessary truths that sensitivity-based accounts do. For the remainder of this paper, we focus on whether or not the superiority of safety-based accounts is defensible on these grounds.

Before considering how well safety-based accounts of knowledge handle knowledge of necessary truths, note that an objection to safety-based accounts that has nothing to do with necessity pushes safety-theorists towards a slight modification of the standard Lewisian semantics for counterfactuals. The objection we have in mind is the so-called *true–true objection*.[11] This objection turns on

[11] See DeRose (2004, §5).

the observation that, on Lewisian semantics for counterfactuals, subjunctive conditionals with (actually) true antecedents and consequents are automatically true. It follows that when (Saf1) and (Saf2) hold, (Saf3) is automatically true. Thus, the safety-based account collapses into a true belief account of knowledge.

The intuition underlying the modification of the semantics for counterfactuals which enables us to avoid this problem is that a conditional of the form "if it were the case that p, it would be the case that q" is true just in case *nearly all* relevantly similar $\langle p\rangle$-worlds are $\langle q\rangle$-worlds.[12] The idea is that to see whether S's true belief $\langle p\rangle$ is safe, we should consider the subclass of $\langle S$ believes that $p\rangle$-worlds that are relevantly similar to $@_S$—in particular, those worlds in which (respecting the Constancy Principle) S comes to believe $\langle p\rangle$ in the same way as in $@_S$—and see whether or not nearly all of those worlds are $\langle p\rangle$-worlds. So, for example, Sam's belief that it is three o'clock is unsafe since most worlds where he believes that it is three o'clock on the basis of seeing stopped-at-three Big Ben read three o'clock are not worlds in which he passes Big Ben at three o'clock. Modifying the standard Lewisian semantics in accordance with this intuition gives us:

> *Modified Standard Semantics for Counterfactuals* A conditional of the form, "if it were the case that p, it would be the case that q," is true in context C and world w just in case either (i) there is no $\langle p\rangle$-world or (ii) if there is a $\langle p\rangle$-world, then nearly all (C, w)-relevantly similar $\langle p\rangle$-worlds are $\langle q\rangle$-worlds rather than $\langle \neg q\rangle$-worlds.

This takes care of the true–true objection, but it's not too hard to see that even with this modified semantics safety-based accounts fare no better with respect to necessary truths than do sensitivity-based accounts. This since, even on the modified semantics, if $\langle p\rangle$ is necessarily true, then every relevantly similar $\langle S$ believes that $p\rangle$-world is a $\langle p\rangle$-world. *A fortiori*, if $\langle p\rangle$ is necessarily true, then nearly every relevantly similar $\langle S$ believes that $p\rangle$-world is a $\langle p\rangle$-world. So (Saf3) is trivially true given (Saf1) in the case that $\langle p\rangle$ is a necessary truth.

Suppose, for example, that Sam looks at a working calculator which tells him that 131,071 is a prime number, and on this basis Sam comes to believe (truly) that 131,071 is indeed prime. On the safety-based account of knowledge, this results in knowledge for Sam if and only if

[12] See, e.g., Pritchard (2005, 2007) for formulations of safety with this intuition built in.

(†) were Sam to believe that 131,071 is a prime number, it would be the case that 131,071 is a prime number.

According to the Modified Standard Semantics for Counterfactuals, (†) is true in a context of utterance C at a world w just when:

(i) there is no ⟨Sam believes that 131,071 is a prime number⟩-world; or

(ii) if there is a ⟨Sam believes that 131,071 is a prime number⟩-world, then nearly all (C, w)-relevantly similar such worlds are also ⟨131,071 is a prime number⟩-worlds.

Clearly (i) is false, since $@_{Sam}$ is a ⟨Sam believes that 131,071 is a prime number⟩-world. However, (ii) is trivially true.

Since every world is a world in which 131,071 is prime, every ⟨Sam believes that 131,071 is a prime number⟩-world is a ⟨131,071 is a prime number⟩-world. So nearly every relevantly similar ⟨Sam believes that 131,071 is a prime number⟩-world is a ⟨131,071 is a prime number⟩-world. And there is at least one ⟨Sam believes that 131,071 is a prime number⟩-world, by (Saf1). It follows that Sam knows that 131,071 is a prime number. The thing to notice is that this argument works just as well if Sam's calculator is broken. This shows that on safety-based accounts of knowledge, knowledge of necessary truths is as cheap, easy, and universal as it is on sensitivity-based accounts.

5 Rejoinders and Replies

This problem with safety-based accounts of knowledge, though it has not received much attention in the epistemology literature, isn't entirely unknown. For one example, Williamson (2000) acknowledges the problem but quickly dismisses it as a genuine challenge to safety on the grounds that Sam's belief that 131,071 is a prime number (and others like it) "fails to be knowledge because the method by which he reached it could just as easily have led to a false belief in a different proposition" (Williamson, 2000, p. 182).[13] For another, Pritchard (2009) endorses modifying his safety-based account of knowledge to handle necessary truths along similar lines. He writes, "[A]ll

[13]The example on which Williamson is commenting here isn't exactly analogous to ours in that, if Sam has no reason to believe his calculator is broken he would intuitively be justified in his belief. In Williamson's example (a coin tossing case), however, the agent's belief is intuitively unjustified. We won't comment on this further, but it is a difference in the cases.

we need to do is to talk of the doxastic result of the target belief-forming process, whatever that might be, and not focus solely on the belief in the target proposition" (Pritchard, 2009, p. 34).[14] In addition, and significantly, Pritchard adopts a central feature of virtue epistemology as practiced by Sosa (2007), Zagzebski (1999), and, most especially, Greco (1999, 2000). We take these rejoinders in turn.

5.1 Williamson

We can better see what's going on with Williamson's dismissal by more precisely rendering its grounds. Williamson understands an agent S's true belief $\langle p \rangle$ to be safe only if:

> (G) The method M which resulted in S's belief $\langle p \rangle$ is such that, for every proposition $\langle q \rangle$, M could not easily result in a false belief $\langle q \rangle$ for S.

In keeping with our counterfactual treatment of 'could not easily', (G) can be refined to:

> (G′) The method M which resulted in S's belief $\langle p \rangle$ is such that, for every proposition $\langle q \rangle$, nearly every $(C, @_S)$-relevantly similar world where S believes $\langle q \rangle$ as a result of M is a $\langle q \rangle$-world.

It's not hard to see that the condition on safety underlying Williamson's dismissal is a condition on methods or processes of belief formation. Safety, as we've characterized it and as it's typically characterized, is a condition on beliefs (or, more precisely, on agent–belief pairs). So Williamson's appeal to (G′) is a modification of the safety-based account of knowledge as set out earlier. We think there is no refuge for safety theorists here.

A slight modification of the calculator example is immune to Williamson's dismissal. Suppose that Sam's calculator is broken, that Sam is ignorant of this fact, and that a benevolent demon ensures that Sam's calculator shows only correct answers.[15] In this case Sam's belief that 131,071 is a prime number is still true at nearly every relevantly similar ⟨Sam believes that 131,071 is a prime number⟩-world, for the same reasons as above, and so is safe according to (Saf3). But now it's not the case that the method by which Sam reached his belief could easily have led him to have a false

[14] This constitutes an extension of Pritchard's view since that view only explicitly applies to what he calls *fully contingent* propositions—propositions that are not necessary in any sense (logically, metaphysically, physically, etc.).

[15] Cf. Greco's helpful demon counterexample to simple (process) reliabilism in Greco (1999, p. 286).

belief in a different proposition. Let $\langle q \rangle$ be a proposition appropriate to the method in question (i.e., a mathematical proposition amenable to answering by calculator). If $\langle q \rangle$ is true, then Sam's method (consulting the broken calculator being fed correct answers by the benevolent demon) will indicate that. And if it's false, Sam's method will indicate that. Failure on either of these counts would require violating the Constancy Principle. So there is no appropriate proposition distinct from $\langle p \rangle$ such that Sam might have easily falsely believed it on the basis of the method that produced his belief $\langle p \rangle$. It follows that Sam's belief is safe even taking (G′) into account.

One might respond on behalf of the safety-theorist that the demon calculator case doesn't really get at a problem for safety in the sense of (Saf3), but rather causes trouble for the additional condition (G′). And this is correct. That propositions appropriate to the method in question are necessary ensures satisfaction of (Saf3); that the method only gives correct answers, and does so in all worlds admissible by the Constancy Principle, ensures the satisfaction of (G′). Both of these components play a role in securing the judgment that Sam knows according to the modified account. Focusing exclusively on satisfying (Saf3) would be insufficient. But this isn't a strike against our reply; it's supposed to be a reply to Williamson's modified safety-based account. Given this, it's natural that the reply in large part target the modification to the original safety-based account.

Another response a defender of safety might offer takes issue with the belief-forming method used by Sam in the demon-calculator case including the demon feeding correct answers to the calculator. If we allowed the veracity of the answers provided by the demon to vary while still counting the method as the same Sam uses in $@_{Sam}$, then arguably the class of relevantly similar worlds would include many worlds where Sam formed false beliefs on the basis of this method. This would, of course, give us a violation of (G′). The problem with this response is that it's not at all clear why we should allow this variation in the veracity of the answers provided by the demon. If methods were being individuated internally, then we could make a case for this variation. But safety theorists seem to agree that, however the generality problem is to be solved (if it's soluble at all), belief-forming methods should be individuated externally. Pritchard, e.g., writes that "the 'way' in which the belief is actually formed needs to be individuated *externally* rather than *internally*," and of this condition, "I take it that this is relatively uncontroversial…" (Pritchard, 2005, p. 152, original emphases). Williamson (2000, ch. 7) expresses similar sentiments. At best, this response incurs a non-trivial challenge to spell out an externalist account of individuating methods of belief

formation which rules out strange-process cases like the demon calculator without also ruling out paradigm cases of knowing.

5.2 Pritchard

Pritchard has advanced slightly different versions of a safety condition on knowing in different places,[16] but the core of his understanding of safety is nicely captured in his (2007) formulation:

> (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. (p. 283)

It turns out that (SP*) is ambiguous, depending on how 'the target proposition' is understood.[17]

Suppose it's the safety of S's belief $\langle p \rangle$ which is at issue, and in the actual world S's belief $\langle p \rangle$ results from S's use of method M. One reading of (SP*) takes 'the target proposition' to denote the belief, say the belief $\langle q \rangle$, that results from S's use of M in whatever nearby world (not necessarily the actual world). On this reading, S's belief $\langle q \rangle$ needn't be the same belief whose safety is at issue, i.e., S's belief $\langle p \rangle$. Taken this way, (SP*) might appear to be immune to the problem we've raised concerning safety and necessary truths. Indeed, this way of understanding (SP*) seems well in line with Pritchard's suggestion for how to extend his account to cover knowledge of necessary truths[18] (though not obviously with his understanding of (SP*) on its own). Suppose, for example, that Dave forms the belief that $2 + 2 = 4$ on the basis of tossing a coin.[19] Since Dave's belief is necessarily true, there is no nearby possible world where it's false. But given the method Dave uses to form his belief, there are many nearby possible worlds were Dave forms a false belief distinct from his belief that $2 + 2 = 4$ (e.g., the belief that $2 + 2 = 5$) using the same method. So (SP*) rules Dave's belief that $2 + 2 = 4$ unsafe, even though it's a belief in a necessary truth.

The idea here is that a belief which results from a method that does not produce true beliefs in nearly all nearby worlds is unsafe. Notice that this relies on the requirement that the method M which results in the belief the safety of which is at issue is such that in nearly all nearby worlds beliefs resulting from M are true. But this is just a terminological variant of (G′). So Pritchard's

[16] See, e.g., (2005, p. 156) and (2009, p. 34) in addition to the version about to be stated.
[17] Thanks to an anonymous referee for encouraging us to expand our discussion of Pritchard's safety condition.
[18] See Pritchard (2009, p. 34).
[19] This example comes from Pritchard (2009, p. 34).

safety-based view, where 'the target proposition' in (SP*) is taken to denote whatever belief that results from S's use of M in whatever nearby world, relies on (G′). Hence, it has the same issues as a solution to the problem raised for safety in §4 as Williamson's modified account does.

Alternatively, (SP*) can be read so that 'the target proposition' denotes S's belief $\langle p \rangle$, i.e., the very belief the safety of which is at issue. On this reading, it's clear that Pritchard's safety-based view has the same problem as other safety-based views covered in §4. However, Pritchard does more in (2009) than just gesture at a way of extending his safety-based account of knowledge to cover knowledge of necessary truths, and one might think that combining this second reading of (SP*) with Pritchard's modification to his view yields an account of knowledge which is both safety-based and able to handle knowledge of necessary truths. Unfortunately for the safety theorist, things don't work out so nicely.

Pritchard augments his account of an agent's knowing $\langle p \rangle$ to incorporate "an ability condition of some sort—i.e., a condition to the effect that the true belief was gained via the employment of the agent's reliable cognitive abilities" (Pritchard, 2009, p. 41). In short, Pritchard modifies his view so that an agent S knows $\langle p \rangle$ just in case S has a true safe belief $\langle p \rangle$ which results from the exercise of her reliable cognitive abilities (roughly, character traits such as careful attention to evidence and cognitive faculties such as various modes of perception—abilities the use of which is advantageous in acquiring true beliefs).[20]

To help get a handle on the role cognitive abilities are supposed to play in Pritchard's account of knowledge, we consider a case discussed in (2009).[21] Temp is a person who forms beliefs about the temperature of the room he's in by looking at a thermometer on the room wall. Though he has no reason to think that the thermometer isn't working normally, in fact it's broken and giving random readings within a certain range of values. Despite this, looking at the thermometer is a highly reliable way of forming true beliefs about the temperature in the room, since there is someone observing from a hidden location who adjusts the temperature of the room to match whatever the thermometer reads whenever Temp checks it.

[20]Technically, the four conditions on knowing here given may require minor supplementation—"tweaking"—to get jointly sufficient conditions on knowing. However, Pritchard takes these four conditions to consititute the "core" of what knowing requires. (See Pritchard (2009, pp. 34–35, 41).) As such, we will treat them as jointly sufficient for present purposes.

[21]See pp. 40–41.

Temp's beliefs about the room temperature are reliably true. Indeed, they're safe: in nearly every relevantly similar world where Temp forms a given temperature belief about the room in the same way as in $@_{Temp}$, that belief will be true. Pritchard endorses all of this. But, of course, a satisfactory account of knowledge should not count Temp as knowing the temperature of the room when he looks at the thermometer.

Pritchard's diagnosis of the problem with this case is that the *direction of fit* between Temp's beliefs and the facts is wrong: the facts are changing to match Temp's beliefs, rather than the other way around. The remedy, according to Pritchard, is to introduce the cognitive ability condition given above. Temp doesn't know because his temperature beliefs are true (i.e., match the facts) not as a result of the exercise of his cognitive abilities, but rather through the (unknown to Temp) change of circumstances to fit those beliefs. We don't positively evaluate the skill of an archer by virtue of her hitting a target that was (unknown to her) moved into the path of her arrow because her hitting the target owed little to her ability as an archer. Similarly, we don't positively evaluate the belief of an agent by virtue of that belief's being true due to the facts being arranged to make it so because acquisition of a true belief in such a case owes little to the agent's ability as an epistemic agent.

The Temp case is similar to our demon calculator case, in that both are strange process-type cases. Moreover, we're willing to grant that modifying a safety-based account of knowledge by adding a cognitive ability condition satisfactorily handles the Temp case.[22] However, such modification won't handle the demon calculator case. The crucial difference between the cases is that in the Temp case the target proposition is contingent, while in the demon calculator case the target proposition is necessary. If $\langle p \rangle$ is a necessary truth, it's not possible to change the facts to fit an agent's belief. So any interference in a case where the target proposition is a necessary truth has to be in service of the agent acquiring a belief that accurately reflects how things already are, not in service of the agent acquiring a belief which is then made true. And this is just what happens in the demon calculator case. The case is specified so that Sam's beliefs accurately reflect what's already true. The direction of fit between belief and the facts is precisely what it should be—belief fit to facts.

[22] It's not clear to us why Pritchard's modified account doesn't collapse into a virtue reliabilist account, essentially Greco's agent reliabilism. If there is such a collapse, this could be used to argue against mounting a defense of *safety-based* epistemology in the way Pritchard (2009) does. That said, we won't worry over this point here.

A safety theorist might reply that Sam's true belief in the demon calculator case does not appropriately result from exercise of his cognitive abilities. After all, it's the demon's intervention that's responsible for Sam acquiring a true belief. Were the demon not intervening, Sam's acquiring true beliefs by consulting his calculator would be very unlikely. Notice, however, that the same can be said of the case where Sam consults a working calculator. In such a case, were the calculator not working properly, Sam's acquiring true beliefs by consulting the calculator would be very unlikely. But we don't thereby judge that Sam's calculator-based beliefs don't appropriately result from exercise of his cognitive abilities. So why should we so judge in the demon calculator case? In both cases, we consider what would be the case were some feature of the case to be different, a feature which is external to Sam and critical to the reliability of the method involved. This arguably violates the Constancy Principle. But at best this reply incurs the burden of explaining why we should count Sam's belief in the working calculator as satisfying the cognitive ability condition, but not similarly count his belief in the demon calculator case. The rub is that there's nothing Sam is doing in one case that he's not doing in the other. The only difference in the cases is in factors external to Sam and it's hard to see what Sam could reasonably be expected to do to distinguish them, especially since the methods involved are equally reliable.

Another way of seeing the problem here is by considering the way in which Greco understands cognitive abilities.[23] For Greco, following Sosa, cognitive abilities constitute an agent's intellectual character. As such, those abilities are stable dispositions to believe of the agent. So, for example, S's stable disposition to believe that things are such-and-so on the basis of things visually appearing to be such-and-so, in normal viewing conditions and when S is motivated to believe what's true (as opposed to, e.g., what might be comforting), is a cognitive ability. Thus, S's belief that there is a cat on the mat upon having the appropriate visual experience in normal viewing conditions when motivated to believe what's true is grounded in this stable disposition and, hence, results from the exercise of (one of) her cognitive abilities. But it's easy to see that the very same stable disposition to believe (viz., the disposition to believe what his calculator tells him when there's no reason to suspect it's malfunctioning and he's motivated to believe what's true) grounds Sam's belief in the demon calculator case as in the working calculator case. So in both cases Sam's belief results from the exercise of his cognitive abilities (indeed, the same cognitive ability). But then we can't affirm

[23] See, e.g., (Greco, 1999, pp. 286–291).

that his belief satisfies the cognitive ability condition in one case but not in the other. Hence, Sam either knows in both cases or he knows in neither case. Either way, the reply under consideration comes up short.

6 Concluding remarks

If the foregoing is correct, then safety-based accounts have no advantage over sensitivity-based accounts when it comes to necessary truths. Both either must adopt a radically variant semantics for counterfactuals or come up with an account of the existence of non-metaphysically possible worlds and how such worlds can be more or less similar to metaphysically possible worlds. If this has been overlooked, it is almost certainly because the literature has focused almost exclusively on the problem of counterpossibles, of how necessarily false antecedents make counterfactuals trivially true, and neglected the way in which necessarily true consequents also make counterfactuals trivially true. Safety theorists might find hope in the work of Williamson or Pritchard; however, that hope will require difficult work to be realized, if it can be realized at all.

References

Brogaard, B. and Salerno, J. (forthcoming). "Remarks on Counterpossibles". In Johan van Benthem, Vincent F. Hendricks, John Symons, and Stig Andur Pedersen, editors, *Between Logic and Intuition: David Lewis and the Future of Formal Methods in Philosophy*. Synthese Library, forthcoming.

Comesaña, J. (2005). "Unsafe Knowledge". *Synthese*, 146:395–404.

DeRose, K. (1995). "Solving the Skeptical Problem". *Philosophical Review*, 104:1–52.

DeRose, K. (2004). "Sosa, Safety, and Skeptical Hypotheses". In Greco, J., editor, *Sosa and His Critics*, pages 22–41. Blackwell Publishers.

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

Gettier, E. (1963). "Is Justified True Belief Knowledge?". *Analysis*, 23(6):121–123.

Goldman, A. (1976). "Discrimination and Perceptual Knowledge". In *Liasons: Philosophy Meets the Cognitive Sciences*, pages 85–103. MIT Press.

Greco, J. (1999). "Agent Reliabilism". *Philosophical Perspectives*, 13:273–296.

Greco, J. (2000). *Putting Skeptics in Their Place*. Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Blackwell Publishers.

Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell Publishers.

Nozick, R. (1981). *Philosophical Explanations*. Oxford University Press.

Pritchard, D. (2005). *Epistemic Luck*. Oxford University Press.

Pritchard, D. (2007). "Anti-Luck Epistemology". *Synthese*, 158:277–297.

Pritchard, D. (2009). "Safety-Based Epistemology: Whither Now?". *Journal of Philosophical Research*, 34:33–45.

Russell, B. (1912). *The Problems of Philosophy*. Oxford University Press.

Sosa, E. (1996). "Postscript to 'Proper Functionalism and Virtue Epistemology'". In Kvanvig, J., editor, *Warrant in Contemporary Epistemology*. Rowman & Littlefield.

Sosa, E. (1999). "How to Defeat Opposition to Moore". *Philosophical Perspectives*, 13:141–153.

Sosa, E. (2000). "Skepticism and Contextualism". *Philosophical Issues*, 10:1–18.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowlewdge*. Oxford University Press.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press.

Zagzebski, L. (1999). "What is Knowledge". In Greco, J. and Sosa, E., editors, *Epistemology*, pages 92–116. Blackwell Publishers.