Indexical Reliabilism and the New Evil Demon[*]

Brian Ball and Michael Blome-Tillmann

Abstract

Stewart Cohen's (1984) *New Evil Demon* argument raises familiar and widely discussed concerns for reliabilist accounts of epistemic justification. A now standard response to this argument, initiated by Alvin Goldman (1988) and Ernest Sosa (1993; 2001), involves distinguishing different notions of justification. Juan Comesaña (2002; 2010) has recently and prominently claimed that his *Indexical Reliabilism* (IR) offers a novel solution in this tradition. We argue, however, that Comesaña's proposal, suffers serious difficulties from the perspective of the philosophy of language. More specifically, we show that the two readings of sentences involving the word 'justified' which are required for Comesaña's solution to the problem are not recoverable within the two-dimensional framework of Robert Stalnaker (1999) to which he appeals. We then consider, and reject, an attempt to overcome this difficulty by appeal to a complication of the theory involving counterfactuals, and conclude the paper by sketching our own preferred solution to Cohen's *New Evil Demon*.

1. Indexical Reliabilism

Stewart Cohen's (1984) *New Evil Demon* argument raises familiar and widely discussed concerns for reliabilist accounts of epistemic justification. Here is the argument (we let 'NED' denote the *New Evil Demon Thesis*, 'SR' the thesis of *Standard Reliabilism*, and 'Biv' the brain in a vat in the closest world to actuality in which there is a brain in a vat[1]):

[*] Thanks to …. The paper is fully collaborative; authors are listed alphabetically.

[1] This way of introducing the name 'Biv' commits us to the *Limit Assumption* (Lewis 1973, pp. 19-21) that there is exactly one closest world to actuality in which there is a brain in a vat. This assumption could be avoided at some

The New Evil Demon Argument:

(NED) Biv's beliefs are as justified as our own beliefs. A

(1) Our beliefs are justified. A

(2) Biv's beliefs are justified. From NED, 1

(SR) *x*'s belief that *p* is justified iff it was produced by a reliable process. A

(3) Biv's beliefs were produced by a reliable process.[2] From 2, SR

The argument is valid, but its conclusion—(3)—is clearly false. Thus, we have to reject at least one of its premises—that is, either (NED), (1) or (SR). Typically, epistemologists opt to reject (SR) and consider the example a *reductio* of *Standard Reliabilism*.[3]

However, following Sosa (1993, p. 61; 2001, p. 387), Comesaña (2002, pp. 256-258; 2010) suggests that we distinguish more carefully between two different versions of reliabilism before rejecting the view entirely. In particular, using '@' to refer to our actual world—the possible world we inhabit—we can distinguish between 'Reliabilism 1' and 'Reliabilism 2':

> *Reliabilism 1* (R1):
> *x*'s belief that *p* in *w* is justified iff it was produced by a process that is reliable in @.
>
> *Reliabilism 2* (R2):
> *x*'s belief that *p* in *w* is justified iff it was produced by a process that is reliable in *w*.

Note, however, that (R2) just makes explicit what (SR) left implicit—so (R2) is clearly refuted by the above argument if (SR) was. To see this in detail, note that from (2), which makes implicit mention of Biv's beliefs in w_{BIV} (as we shall call the possible world Biv inhabits), and (R2) we can infer (3-2):

technical cost. Equally we assume there is just one brain in a vat at the nearest world with a brain in a vat. These assumptions make no difference to the substance of our argument.

[2] Comesaña (2002, p. 255) calls this claim "Demonic Reliability".

[3] Of course, one need not accept the conclusion of the *New Evil Demon*. Sceptics will reject premise (1), and ardent reliabilists will reject (NED).

(3-2) Biv's beliefs in w_{BIV} were produced by a process that is reliable in w_{BIV}.[4]

Surely, (3-2) is false, so (R2) does not help resolve the *New Evil Demon*. In fact, most epistemologists will recognize (R2) as the bad guy in the background of the *New Evil Demon*.[5]

How about (R1)? Note that (3) and (3-2) cannot be derived from (2) and (R1). The only conclusion we can derive from the conjunction of (2) and (R1) is (3-1):

(3-1) Biv's beliefs in w_{BIV} were produced by a process that is reliable in @.

Surely, (3-1) is true (making suitable assumptions about the individuation of methods). As a consequence, if we accept (R1), the *New Evil Demon* has been defused. However, note that (R1) is subject to other counterexamples: there are worlds in which people are justified on the basis of reliable clairvoyance, for instance, even though clairvoyance is not reliable in @.[6] Thus, (R1) does not really capture the spirit of reliabilism and avoids the *New Evil Demon* by pain of being subject to other counterexamples. *Prima facie* then, both (R1) and (R2) face difficulties.

Comesaña, however, thinks that he has a solution to the problem. In particular, he thinks that if we, firstly, adopt Stalnaker's (1999) two-dimensional semantics and, secondly, follow Lewis (1970) in recognizing (i) that 'actually' is an indexical expression (referring to the world in which it was uttered), and (ii) that it has another sense as well, in which it is effectively redundant, then each of the two positions above—(R1) and (R2)—can be expressed succinctly in one neat statement:

[4] Comesaña talks about brains-in-vats (or victims of evil demons, to be precise) in the plural and does not refer to a particular brain-in-a-vat by means of a proper name, as we do. However, changing our version of the *New Evil Demon* to match his would merely complicate the formulation of the principles at issue slightly: instead of referring to Biv's beliefs in w_{BIV} we would, on his formulation, need to refer to the beliefs of brains-in-vats in their respective worlds $w_{\mathrm{BIV1}}, \ldots, w_{\mathrm{BIVn}}$.

[5] Those who don't will still have to blame (NED) or embrace scepticism.

[6] See (BonJour 1980) for the clairvoyance cases. Sosa (2001, p. 390) recognizes that "we countenance possible higher beings who gain knowledge by properly and reliably forming ('warranted') beliefs, despite the fact that their epistemic ways, while successful in their world, would be miserably inadequate in ours." He does not appear to acknowledge that this poses a threat to (R1).

Indexical Reliabilism (IR):
x's belief that *p* is justified iff it was produced by a process that is actually reliable.[7]

Comesaña endorses this account of epistemic justification, and claims that it helps us to solve the *New Evil Demon*. More specifically, he points out that from (2) and (IR) we can no longer derive (3) but only the following indexical version of (3):

(3I) Biv's beliefs were produced by a process that is actually reliable.

But (3I), Comesaña claims, just like (IR), admits of two readings within Stalnaker's two-dimensional framework—namely, (3-1) and (3-2). Accordingly, the *New Evil Demon* suffers the fallacy of equivocation: the claim (3-1) which follows from the true readings of the premises of the argument is not absurd; yet (3-2), which is clearly false, follows only given a false reading of (NED). Similarly, Comesaña can claim about possible reliable clairvoyants that they are justified in the sense captured by (R2), but not in the sense captured by (R1): once we disambiguate between the two different readings of 'justified' the problem evaporates.

Comesaña claims that his solution to the *New Evil Demon* is new and offers an improvement to accounts already in the literature. He says:

> One of the most popular strategies [for dealing with the *New Evil Demon*] is the postulation of different senses of 'justification', one in which the victims are justified and the other one (the reliabilist one) in which they are not. (Comesaña 2010, p. 579)

But he maintains that his (IR) allows for a more satisfactory solution. In particular, "we don't need to say that there are two senses of 'justified,' just as we wouldn't say that there are two senses of 'here'" (2010, p. 580) simply because it can be used to refer to two different places on two different occasions. Rather, the semantics of 'actually' allows the two readings of (IR) and (3I). Thus, Comesaña (ibid.) claims that his indexical account is "better than solutions to the new

[7] The statement of (IR) is complicated slightly at (Comesaña 2010, p. 579); however, the additional clause added there is irrelevant to our concerns. One might also wonder whether (IR) ought to be formulated meta-linguistically, but we shall ignore these issues here.

evil demon problem that […] postulate ambiguity", presumably because it respects Grice's Razor—a methodological principle according to which one ought not to multiply senses beyond necessity.[8]

2. Diagonalization and the New Evil Demon

How are we to recover the two readings of (3I) that Comesaña postulates? Comesaña (2002, p. 251ff.; 2010, pp. 579-80) claims that Stalnaker's two-dimensional framework plays a crucial explanatory role in achieving this. In particular, he (2002, pp. 258-9) suggests that (3-1) expresses what Stalnaker (1978) calls the 'horizontal proposition' of (2) and (3I), while (3-2) expresses those sentence's 'diagonal proposition'. In accordance with these claims, Comesaña says that if (3-1) is true Biv is "horizontally justified", and he is "diagonally justified"[9] if (3-2) is true. Thus, on Comesaña's view there are two different types of justification that can be expressed by our predicate 'is justified'—without this account being an ambiguity theory.

To see in more detail how this account is meant to resolve the *New Evil Demon*, consider what Comesaña says about more everyday and obviously true ascriptions of 'justification' such as (4) (where 'Ernie', we take it, denotes Ernest Sosa):

(4) Ernie's beliefs are justified.

The standard reliabilist regards (4) as equivalent to (5):

(5) Ernie's beliefs are produced by a reliable process.

But Comesaña suggests that we treat (4) as semantically equivalent to (5I) instead:

(5I) Ernie's beliefs were formed by a process that is actually reliable.

[8] The principle is due to Grice (1989, pp. 47-9), who calls it "Modified Occam's Razor".

[9] The terms of 'horizontal' and 'diagonal justification' are employed in (Comesaña 2002, p. 256, fn. 17; 2010, p. 579).

And here is the propositional concept[10] for our actual utterances of (5I):

[5I] – Propositional Concept for (5I)[11]

	@	w_{BIV}
@	T	T
w_{BIV}	F	F

Comesaña (2002, pp. 258-9) now says that we can use (5I) to express two different propositions—its horizontal or its diagonal. And he says that when we express the horizontal we claim that Ernie is ‘horizontally justified’, whereas if we express the diagonal we say that he is ‘diagonally justified’. Further, Comesaña claims that the horizontal amounts to (5-1) whereas the diagonal amounts to (5-2):

(5-1) Ernie’s beliefs were produced by a process that is reliable in @.
(5-2) Ernie’s beliefs were produced by a process that is reliable in Ernie’s world.

Thus, the horizontal expresses the reading according to R1 and the diagonal the reading according to R2. This view is evidently both elegant and attractive, for it assigns exactly the two readings according to R1 and R2 to (5I)—and thus to (4)—without forcing us to postulate semantic ambiguity. Comesaña has shown that with respect to (5I) and (4) we can have our cake and eat it.[12]

Comesaña (2002, p. 258) suggests further that (2) and (3I) can be treated exactly analogously to (4) and (5I).[13] In fact, his comments suggest that (2) and (3I), just like (4) and (5I), are associated with a horizontal and diagonal reading, and that the horizontal reading corresponds to

[10] See (Stalnaker 1978).

[11] We assume here, as does Comesaña (in effect), that the context is one in which the conversational participants do not presuppose that there are no brains in vats (see below). We also follow Comesaña in restricting our attention to just two worlds.

[12] One complication is worth mentioning: on Stalnaker’s account it is not the case that both ‘readings’ are available in one and the same context. In this particular case, the horizontal cannot be asserted as it is true in every world in the context set. We discuss this issue in more detail below.

[13] Comesaña (2002, p. 258) says “‘Ernie’s beliefs are justified’, then, is ambiguous; and so is [(3)], and for the same reasons.” And he qualifies his use of ‘ambiguous’ as meaning ‘expresses distinct diagonal and horizontal propositions’.

(3-1) while the diagonal reading corresponds to (3-2):

(3-1) Biv's beliefs in w_{BIV} were produced by a process that is reliable in @.
(3-2) Biv's beliefs in w_{BIV} were produced by a process that is reliable in w_{BIV}.

If this is right, then the horizontal must be true, while the diagonal is false; and, of course, what matters here is whether the relevant claims are, as Comesaña puts it, "true *simpliciter*, that is, true in the actual world."[14]

There are a number of problems with this position. To begin with, consider the Stalnakerian propositional concept[15] for (3I):

[3I] – Propositional Concept for (3I)[16]

	@	w_{BIV}
@	T	T
w_{BIV}	F	F

With [3I] at hand, we can see that an actual utterance of (3I) is associated with two different propositions: the horizontal and the diagonal.[17] Thus, it seems Comesaña is right that, assuming Stalnaker's two-dimensional framework, (3I) has two different readings. However, note that Comesaña claims that while the relevant horizontal proposition is true, the diagonal proposition is false— that is, "false in the actual world, false *simpliciter*" (2002, p. 258).[18] This is rather puz-

[14] (Comesaña 2002, p. 256, emphasis in original).

[15] See (Stalnaker 1978).

[16] The same assumptions are made here as were made in connection with [5I] (see above).

[17] By 'the horizontal' we mean, of course, the top-most horizontal; for we are considering an utterance of (3I) in @.

[18] Ironically, Comesaña introduces this kind of talk of truth and falsity *simpliciter* in a context in which it is *in*appropriate to assess a claim for truth simpliciter. He writes,

> "In the demoners' world, taking experience at face value is highly *un*reliable: it usually yields false beliefs, and it would yield mostly false beliefs in counterfactual applications—*in the demoners' world*. Here, in this world, taking experience at face value is highly reliable: it usually yields true beliefs, and this is not just a statistical correlation, for it *would* yield true beliefs if used in appropriate circumstances. So the crucial question is what we are referring to when we say that a justified belief must have been produced by a process *most of whose outputs would be true*. True where? The immediate answer is: true *simpliciter*, that is, true in the actual world." (Comesaña 2002, p. 256)

zling: as is obvious from the propositional concept [3I], the diagonal of (3I) is true at @. Thus, if Comesaña wants our actual utterance of (3I) to have two differing readings with differing truth-values at our world, then those readings cannot be the horizontal and the diagonal propositions respectively: both are, after all, true at @. Moreover, note that Comesaña's claim that (3-2) expresses the diagonal proposition of (3I) cannot be correct either: the diagonal of (3I) is, as we have just seen, true at @, whereas (3-2) expresses a proposition that is false at @. As a consequence, we cannot account for the two alleged readings of (3I)—that is, (3-1) and (3-2)—by means of Stalnaker's two-dimensional semantics.

Here is another problem with Comesaña's claim that (3I) has two different readings that can be recovered within the framework of two-dimensional semantics. According to Stalnaker's two-dimensionalism, it is usually not the case that both readings are available in one and the same conversational context. But this is required for Comesaña's solution to the *New Evil Demon*. Remember that, according to Comesaña, the conclusion of the *New Evil Demon* is absurd only on the unsound reading of the argument, on which the conclusion amounts to (3-2), but not so on the sound one, on which the conclusion amounts to (3-1): according to Comesaña's diagnosis, the *New Evil Demon* suffers from the fallacy of equivocation. This very diagnosis, however, can only be correct if both readings of the premises and the conclusion are available in the context of the epistemological discussion conducted by Comesaña and also in this paper. If both readings were not to be available at the same time, Comesaña's diagnosis that the *New Evil Demon* suffers from the fallacy of equivocation would fail.

Semantic orthodoxy has it that the proposition that is asserted when one uses a sentence is the horizontal proposition semantically associated with the sentence in the context in which it is

However, it is clear that the beliefs in question must be true, not in the actual world, but in their respective nearby variants of the actual world. Obviously Comesaña means that the belief forming process must be reliable in @, and not that the beliefs formed by this process must be true in @.

used. Stalnaker argues that it is only under special circumstances that one will end up asserting the diagonal proposition instead. In particular, one cannot assert the necessary truth—the proposition that is true at every world—for this is uninformative. Nor can one assert the necessary falsehood. Stalnaker claims that when we enter into a conversation we make certain presuppositions—as do the other conversational participants. When these presuppositions are shared, we may take them to determine a set of possible worlds—those which are compatible with the presuppositions made—which Stalnaker (1978) calls 'the context set'. The aim of assertion, according to Stalnaker, is to narrow the range of possibilities that are compatible with the shared presuppositions in the context by ruling out those possibilities in which the proposition asserted is false. Accordingly, if one were to assert the necessary proposition one would not rule out any possibilities and one's assertion would be pointless. By contrast, if one were to assert the necessary falsehood one would rule out every possibility—but since one aims to assert the truth, this would be self-defeating. In circumstances such as these, Stalnaker claims, Gricean considerations lead to the conclusion that it is the diagonal proposition which is asserted.

In a conversational context in which the participants presuppose nothing false, yet do not presuppose that there is no brain in a vat, both @ and w_{BIV} will be compatible with their presuppositions.[19] It is a context of this kind which is represented above by our propositional concept [3I]. In such a context, however, the horizontal proposition associated with (3I), and indeed (2)—which, given (IR), is its definitional equivalent—is true in every world in the context set. It therefore cannot be asserted, according to Stalnaker; the diagonal proposition which is true at @ and false at w_{BIV} is asserted instead.

Thus, in Stalnaker's framework (3I) does not have two readings in a context in which @

[19] If it were presupposed that there is no brain in a vat then w_{BIV} would not be compatible with what was presupposed, and so would not belong to the context set. Similarly, if something false were presupposed, @ would not belong to the context set.

and w_{BIV} are the live options; and neither does (2). The reason is that one cannot assert the necessary truth on this approach. Thus, not only do the two readings that we can generate by means of Stalnaker's apparatus not correspond to (3-1) and (3-2) respectively, they also cannot—as Comesaña has it—be generated within one and the same context. Thus, Comesaña's contention that the *New Evil Demon* equivocates, in the context of Cohen's epistemological discussion, between two different readings cannot be substantiated by means of Stalnaker's two-dimensional approach.[20]

Comesaña might defend his view by rejecting certain aspects of Stalnaker's view. In particular, he might claim that Stalnaker is wrong to prohibit the assertion of necessary truths and necessary falsehoods, thereby rejecting Stalnaker's account of the mechanism whereby the assertion of the diagonal is triggered. Crudely, Comesaña might claim that we simply *choose* which proposition—the horizontal or the diagonal—to assert.

There are, however, two problems with this strategy. First, the result would be a view that is highly unorthodox in the philosophy of language. To the best of our knowledge, no one has seriously maintained that which proposition we may assert using a given sentence is entirely up to us.[21] But there is a dialectical downside to arguing, as Comesaña does, for an ecumenical position in epistemology—one which allows us to accept both reliabilism and (NED)—on the basis of a polarizing view in the philosophy of language. Why assume that controversy in epistemology must be due to miscommunication, while in the philosophy of language it involves genuine disagreement in which one side is ultimately mistaken? Indeed, it should be noted that Stalnaker's two-dimensional framework is not entirely uncontroversial either;[22] so that even if Comesaña were to stick within this framework his argument would face this challenge.

[20] Note that we are not disputing Comesaña's contention that (3I) *could have* expressed a truth; only that it in fact does so in some actual context with [3I] as its propositional concept.

[21] Of course, speakers can choose which of two meanings of an ambiguous word or phrase to convey; but Comesaña does not regard the two readings of the relevant sentences as disambiguations.

[22] See, for instance, (Soames 2005) and (Hawthorne and Magidor 2009).

A second and perhaps more important problem is that the proposed solution simply would not work. As we have seen, the horizontal and diagonal propositions in the propositional concept [3I] have the same truth value at @—that is, the same truth value *simpliciter*—as one another; so this move will not provide a true and a false reading of (3I). In fact, our argument here does not even rely on the claim that [3I] gives the correct propositional concept for (3I). (We think it does, but that is beside the current point.) Comesaña's claim that (3-1) expresses (3I)'s horizontal and (3-2) its diagonal must be mistaken—whatever propositional concept we may wish to assign to (3I). To see this, remember that Stalnaker (1978, p. 81) defines the diagonal proposition associated with an utterance as the proposition that is true at a world *w* iff what is semantically expressed by the utterance at *w* (its horizontal at *w*) is true at *w*. Thus, it follows straightaway from the definition of the diagonal that the diagonal proposition associated with (3I) is true at @ iff what (3I) semantically expresses at @ (its horizontal at @) is true at @. Comesaña's solution to the *New Evil Demon*, however, requires that our actual utterances of (3I) (utterances in @) express a horizontal and a diagonal whose truth-values *differ* relative to @ as circumstance of evaluation—namely, (3-1) and (3-2). Thus, Comesaña's solution to the *New Evil Demon* cannot be implemented as outlined above.

Let us reformulate this point in a different way. Note that (3-1) expresses a necessary truth, while (3-2) expresses a necessary falsehood: Biv's beliefs in w_{BIV} are produced by the process of relying on one's experiences, which is reliable in @ but not so in w_{BIV}. As a consequence, it also follows that while (3-1) is true relative to @ as circumstance of evaluation, (3-2) is false relative to @ as circumstance of evaluation. However, if (3-1) and (3-2) have different truth-values relative to @ as both context of utterance and circumstance of evaluation, then there cannot be a unique propositional concept such that (3-1) is its horizontal relative to @ while (3-2) is its diagonal. But if there cannot be such a unique propositional concept, then (2) and (3I) cannot be associated with such a propositional concept and Stalnaker's framework in conjunction with the indexicality of 'actually' cannot deliver the solution to the *New Evil Demon* sketched in Comesaña (2002). Again, this is so because the diagonal associated with a sentence is true at a world @ iff its horizontal at @ is true at @. The solution to the *New Evil Demon* sought by Comesaña, however, demands that (3I)'s diagonal be false at @ while its horizontal at @ be true at @. Given the definition of the diagonal, this view is contradictory—no matter what propositional concept we assign to (3I).

Summing up, Comesaña claims that (3I) has a true and a false reading and that those readings are given by (3-1) and (3-2). But those two readings cannot be accounted for by Stalnaker's framework; nor can they be accommodated in a more controversial (less ecumenical) version of two-dimensionalism. Indeed, the claim that the horizontal and diagonal propositions on the propositional concept for (3I), whatever it might be, differ in truth value is a contradiction.

3. Complications and Counterfactuals

The above difficulties for a simple resolution of the *New Evil Demon* by appeal to (IR), together with two-dimensional semantics are insurmountable. To be fair to Comesaña, however, we should note that so far we have only invoked the first of the two Lewisian theses concerning 'actually' which he endorses. Yet it is only after he mentions the second of the two theses that Comesaña writes:

> We are now in a position to see that there is a sense in which [(3I)] is false (false in the actual world, false *simpliciter*). (Comesaña 2002, p. 258)

What is Lewis' second claim about 'actually'? It is the claim that "we can distinguish primary and secondary senses of 'actual'"[23]; as noted above, the primary sense is one in which it acts as an indexical, referring to the world in which it is uttered, while on the secondary sense occurrences of the word are effectively redundant.

It is true that if we recognize these two senses of 'actual' and 'actually' we can generate the two readings of (3I) that Comesaña was looking for. Take, for example:

(3I) Biv's beliefs were produced by a process that is actually reliable.

Given how the name 'Biv' was introduced—namely, as a name of the brain in a vat inhabiting the closest possible world in which there is a brain in a vat, this claim is equivalent to:

(3IC) If there were a brain in a vat, her beliefs would be produced by a process that is actually reliable.

If Lewis is right that 'actually' has a sense on which it refers back to the world of utterance—that is, in our context, to @—and a sense in which it redundantly refers to the world under consideration—in this case, w_{BIV}—then we get the two readings of (3IC),[24] and hence of (3I), which Comesaña sought, namely:

(3-1) Biv's beliefs in w_{BIV} were produced by a process that is reliable in @.
(3-2) Biv's beliefs in w_{BIV} were produced by a process that is reliable in w_{BIV}.

So (3IC), and therefore (we may grant[25]) (3I), can be used to say something true or to say something false, provided that we can distinguish two senses of 'actually'. But distinguishing senses is simply postulating ambiguity. So Comesaña can get his two readings in this way only at the cost

[23] (Lewis 1970, p. 185; Comesaña 2002, p. 257).

[24] In fact, we only get the second reading for (3IC) if we replace 'is' with 'would be'—perhaps with the word 'actually' occurring between 'would' and 'be'. Let this claim be (3IC*). Then we can grant for the sake of argument—something that Comesaña needs—that (3IC*), like (3IC), is equivalent to (3I), under the hypothesis that 'actually' has the two senses Lewis postulates.

[25] See previous note.

of abandoning the advantage he claimed for his view over other proposed solutions to the *New Evil Demon*.

Perhaps it will be said that we have read Lewis's second thesis regarding 'actually' too flat-footedly; although he claimed that the word has two senses, this is more than was intended. Indeed, all that Lewis needed to say, and that Comesaña must follow him in saying, is that on some occasions of its use, 'actually' refers to counterfactual worlds; and on this point he was surely right. For consider two examples from Lewis (also quoted by Comesaña):

(6) If Max ate less, he would be thinner than he actually is.

(7) If Max ate less, he would actually enjoy himself more.

As Lewis emphasizes, in (6) 'actually' refers to the world of the context of utterance, whereas in (7) it refers to the counterfactual world in which Max eats less.

However, this is not all that Comesaña needs if he is to solve the *New Evil Demon* problem in his distinctively novel manner. In particular, Comesaña must hold that sentences involving 'actually' such as (3IC) have two readings *without this involving any ambiguity*. As we have seen, a flat-footed appeal to two senses of 'actually' does not serve Comesaña's purposes; but then some account must be given of how it is that (3IC) acquires its two readings.

One possible explanation is that (3IC) has two readings because 'actually' behaves semantically like a variable, or pronoun. Consider the sentence (8):

(8) If Copernicus hadn't existed, someone else would have proposed the theory he proposed.[26]

On its most natural reading, this sentence can be used to express the proposition that if Copernicus had not existed, someone else would have proposed the theory Copernicus proposed. But

[26] The example is due to Stalnaker (1999, 156).

we can imagine (8) being uttered, by a speaker pointing at Kepler, with a stress on 'he'. In such a context it expresses the proposition that if Copernicus had not existed, someone else would have proposed the theory Kepler proposed. The explanation of the fact that (8) has these two readings, it seems, is as follows. When (8) receives its standard reading, the pronoun 'he' occurring in it is bound by its antecedent 'Copernicus', whereas when it receives the reading on which it concerns Kepler 'he' is not bound by any earlier expression—rather, it occurs freely. Similarly, then, perhaps (3IC) has two readings because on some occasions of its use 'actually' occurs freely, and refers to the world in which it is uttered, while on the other occasions it is bound by the antecedent of the conditional.

This account of how the two readings are generated, however, will not serve Comesaña's purposes; for the difference between free and bound occurrences of variables is one of syntactic structure. The two readings of (8), for instance, arise from the fact that (8) is *structurally ambiguous*; it has two distinct syntactic parsings.[27] If (3IC) has two readings for similar reasons, then it too is structurally ambiguous; and if (2) gains its two readings through equivalence with (3IC), it is ambiguous as well.[28]

Another thing to notice about the above suggestion is that it does not make any use of two-dimensional semantics; yet Comesaña (2002, p. 249; 2010, p. 9) suggests that two-dimensional semantics is essential for his view. Might the two readings sought by Comesaña be generated somewhat differently within Stalnaker's framework than was suggested above? Consider one final attempt. As we have seen, (3I) is plausibly taken to be equivalent to the counterfactual conditional:

[27] See (Heim and Kratzer 1998, pp. 239-259).

[28] Compare, in this connection, the following proposed definition of 'good son': for all *x*, *x* is a good son iff *x* loves his mother. If it is then claimed that sentences involving 'good son' have multiple readings on the grounds that 'his' can refer to *x* or to some contextually salient individual, it would not, we think, be plausible to maintain in addition that 'good son' is not ambiguous.

(3IC) If there were a brain in a vat, her beliefs would be produced by a process that is actually reliable.

It might then be claimed that the *consequent* of (3IC) can indeed be represented by the propositional concept [3I], repeated here for convenience, while the whole conditional cannot:

[3I] – Construed here as the Propositional Concept for the Consequent of (3IC)

	@	w_{BIV}
@	T	T
w_{BIV}	F	F

How do we determine the truth-value of the whole counterfactual conditional (3IC)? The same way we determine the truth-value of any counterfactual conditional: the conditional has the same truth-value as the consequent does in the closest possible world where the antecedent is true. Assuming that w_{BIV} is the closest possible world where there is a brain in a vat, then whether (3IC) is true or false (true or false in the actual world, i.e., true or false 'simpliciter') depends on whether the proposition expressed by the consequent is the horizontal or the diagonal proposition depicted in [3I]. In other words, (3IC)—and thus (3I)—is a counterfactual conditional with a two-dimensional consequent, a consequent with a true horizontal proposition at w_{BIV} and a false diagonal proposition at w_{BIV}. This seems to give Comesaña exactly the two desired readings (3-1) and (3-2) and thus appears to resolve the *New Evil Demon*.

It would be nice, however, to know what the mechanism is whereby the consequent of a conditional such as (3IC) is associated with both the horizontal and the diagonal on [3I]. This may seem a strange request. After all, don't we know how (3I), when uttered in a world in which it is not a disguised counterfactual, is associated with these propositions? And isn't the mechanism in the case of the consequent of (3IC) just the same?

The answer to the second of these questions is 'no'. Following Comesaña, we have been assuming that utterances made using unembedded sentences such as (3I) are associated with propositional concepts in the manner described by Stalnaker (1999). But this mechanism, as we have seen, is pragmatic; the presuppositions of conversational participants determine a context set, and then diagonalization occurs, if it does, for Gricean reasons. Gricean pragmatic mechanisms, however, operate post-semantically; they do not enter into the process of semantic composition. Thus, it would seem, the diagonal proposition on [3I] cannot be associated with the consequent of (3IC); as a pragmatic content it cannot be semantically embedded.

Indeed, consider how Stalnaker himself deals with the interaction of conditionals and contexts (1999, p. 156); he claims that when a conditional, whether indicative or counterfactual, is asserted, we must evaluate the consequent relative not to the "basic context", but instead relative to a "derived context". Suppose then, that one utters (3IC) in a context in which @ and w_{BIV} are the live options.[29] Which worlds will be in the derived context against which we may interpret the consequent? According to Stalnaker, only worlds in which the antecedent is true will be included in the relevant set. Since the antecedent of (3IC) is not true at @ (where there are no brains in vats), @ will not occur in the derived context. Accordingly, [3I] (which is defined on @ and w_{BIV}) cannot provide the propositional concept for the consequent of (3IC), and that consequent cannot therefore be associated with both the horizontal and diagonal on [3I] by Stalnakerian means.[30] Comesaña might, of course, diverge from Stalnaker at this point; but again, in doing so he would be adopting a more controversial position in the philosophy of language in order to achieve ecumenicism in epistemology. The motivation for doing so is unclear; and, of

[29] In fact, in such a context it would be appropriate to utter the corresponding indicative conditional. But on Stalnaker's view, indicative and subjunctive (counterfactual) conditionals express the same kind of conditional proposition; so this does not affect our point.

[30] The effect is that the consequent of (3IC), on the reading we are now considering, where 'actually' picks up its reference from the derived context, is (3-2); accordingly, (3IC) itself is false (at both @ and w_{BIV}). Of course, there remains the reading where 'actually' is simply interpreted as referring to @; the consequent of (3IC) is then equivalent to (3-1), and the conditional itself is true (at both @ and w_{BIV}). (Notice, finally, that this vindicates our association of [3I] with (3I) in the previous section.)

course, the details of the proposal have yet to be seen.

Thus, as we have seen, we can generate the required readings of the key claims if, following Lewis, we recognize two senses of the word 'actually'—one on which it is an indexical, and one on which it is effectively redundant. But this suggestion requires the postulation of an ambiguity after all—and so if it is relied upon, Comesaña's view does not have the advantage it is alleged to have over more traditional ambiguity responses to the *New Evil Demon*. Moreover, it is not clear that any other appeal to the fact that 'actually' can be used to designate counterfactual worlds will serve Comesaña's purposes. Treating 'actually' as a variable which is sometimes free and sometimes bound involves the postulation of structural ambiguity; and Stalnaker's approach to two-dimensional semantics cannot be straightforwardly invoked to generate the two required readings of the consequent of (3IC). If he is to maintain that (2) and (3I) have the two readings he claims for them, while avoiding the postulation of ambiguity, Comesaña must appeal to a semantic mechanism other than those canvassed here.

But we can, in any case, set these semantic worries aside. The general proposal of this section suffers from a crucial defect. To see this note that the strategy being pursued relies essentially on the fact that (3I) is a counterfactual conditional in disguise. But, surely, we can construe versions of the *New Evil Demon* puzzle in which the premise corresponding to (2) does not refer to mere possibilia (such as Biv) and is therefore not a hidden counterfactual. Consider a scenario familiar from the film *The Matrix*: in the film Mr Anderson leaves his vat and enters the real world. But once he has done so he can engage in reasoning equivalent to the *New Evil Demon*, with the crucial difference that instead of referring to merely possible brains in vats he can refer to a world-mate, his former self:

Mr Anderson's New Evil Demon Argument:

(NED*) My former beliefs were as justified as my present beliefs. A

(1*) My present beliefs are justified. A

(2*) My former beliefs were justified. From NED*, 1*

(SR) *x*'s belief that *p* is justified iff it was produced by a reliable process. A

(3*) My former beliefs were produced by a reliable process. From 2*, SR*

Mr Anderson is puzzled, we may assume, for (3*) is clearly false. Assume further that the Oracle tells him about *Indexical Reliabilism*; he accordingly replaces (SR) with (IR). Mr Anderson realizes that he can now no longer derive (3*) but rather only (3I*):

(3I*) My former beliefs were produced by a process that is actually reliable. From 2*, IR

For Comesaña's solution to work he now needs to recover two readings of (3I*); one reading on which (3I*) is true at Mr Anderson's world—that is, the world at which Mr Anderson is puzzled by the above argument—and one on which (3I*) is false at that world. We have seen above that, due to Stalnaker's definition of the diagonal, (3I*) as uttered by Mr Anderson cannot be associated with a horizontal and a diagonal that differ in truth-value relative to Mr Anderson's world. Such a position would be contradictory. However, the strategy advocated in this section—namely, to point out that two different readings can be recovered once we acknowledge that (3I) is a counterfactual conditional in disguise is not available in this case either. For, surely, (3I*) in Mr Anderson's mouth is not a counterfactual conditional in disguise. Mr Anderson refers, after all, to his former self when uttering (3I*), not to a being that is merely possible from his point of view. Thus, we cannot make the counterfactual move sketched above. Yet we do not see any reasons to reject (NED*), (1*), or (2*) that would not also be reasons to reject (NED), (1), or (2).

To try to deal with this case (and others like it, in which a speaker passes judgment on a worldmate's justificatory status), Comesaña might complicate his formulation of (IR)—as he does in his (2002)—by adding temporal and spatial indexicals: *x*'s belief that *p* is justified iff her belief has been formed by a process that is actually reliable *here* and *now*. From this claim Mr Anderson would no longer be able to derive (3I*) but only:

(3I**) My old beliefs were produced by a process that is actually reliable here and now.

This claim, (3I**), has a true reading relative to the circumstances described, as desired. But 'here' and 'now', unlike 'actually', do not appear to have a redundant (as opposed to an indexical) interpretation;[31] in any case, it is obvious that the escaped Mr Anderson could not use 'now' to refer to the time of his past envatted self in uttering (3I**). The result, however, is that no false reading of (3I**) is available in the case at hand, and Comesaña's solution to the New Evil Demon that was based on the diagnosis of an equivocation is not viable, even with the proposed modifications to the thesis (IR). Moreover, and finally, suppose that Mr Anderson were inquisitive, and interested in epistemology prior to his escape from his vat; and suppose further that, upon reflection, he were to utter, prior to his escape, 'My beliefs are justified'. Surely, if our intuitions in the other cases under consideration are that the various brains in vats are correctly judged to have justified beliefs, the intuition in this case will also be that Mr Anderson's utterance in this case is correct too.[32] Yet the move currently being considered would not give Comesaña the desired truth-value (T) for this utterance—while Mr Anderson was in the Matrix, his be-

[31] If interpretations on which the referents of these terms are displaced away from the parametric value determined by the context of utterance are available, they are hard to come by. For instance, one can use 'here' to refer to a place one does not occupy if one is standing in front of a map and pointing at a representation of a place one is not in; but the circumstances which make this possible are quite particular. It does not seem that the relevant interpretations are systematically available in the way required by Comesaña's strategy.

[32] In effect, we are appealing here to the thought that, whatever lies behind Cohen's original intuition that brains in vats have justified but unreliable beliefs, the intuition would be equally legitimate were we to consider as open (i.e. consider as actual) the possibility that we ourselves are brains in vats.

liefs were not produced by a process that was reliable then and there. And, of course, appeal to diagonalization will at this point not be available for the reasons outlined in Section 2: no matter what we take the points of evaluation to be—whether they are worlds, or time-place-world triples, or something more baroque and complicated yet—the horizontal and diagonal propositions defined on those points will agree in truth value at the point occupied by the speaker. The envatted Mr Anderson, like all speakers, is at the very point of evaluation that he occupies, and Comesaña's two-dimensionalist strategy offers no way of generating two propositions associated with a given utterance which differ in truth value at the point occupied by the speaker.

Summing up, the fact that our initial (3I) is a counterfactual in disguise is a purely accidental feature of the original *New Evil Demon* puzzle, and any attempted resolution of the puzzle that relies on that fact is explanatorily inadequate.

4. Concluding Remarks

We have shown that Comesaña's attempts to account for the different alleged readings of (3I)—or (3I*)—by means of Stalnaker's two-dimensionalism fail. Thus, Stalnaker's framework in conjunction with the indexicality of 'actually' cannot deliver the desired solution to the *New Evil Demon*. This is so because the diagonal associated with a sentence is true at a world @ iff its horizontal at @ is true at @. The solution to the *New Evil Demon* sought by Comesaña, however, demands that (3I)'s diagonal be false at @ while its horizontal at @ be true at @. Given the definition of the diagonal, this view is contradictory. Responses to this concern that rely on the assumption that (3I) is a counterfactual conditional in disguise have also been found wanting: they involve the postulation of ambiguity, thereby surrendering any alleged advantage over rival views; and in any case they cannot handle obviously non-counterfactual versions of the *New Evil*

Demon, such as Mr Anderson's case from the Matrix. Baroque variants of (IR) are semantically suspect, and generate further epicycles without addressing the central problem.

Finally, we should like to note that even if Comesaña were to find a way to model the semantics of (2) and (3I) or (2*) and (3I*) in a way that avoids postulating an ambiguity, his account would still suffer from a crucial defect: the notion of epistemic justification, if it is to be of any interest, must presumably have *some* connection with epistemic normativity.[33] In particular, whether a belief is justified should have some bearing on whether the subject ought to have that belief so that, at the very least, if a belief is unjustified then the subject ought not to hold it, and if it is justified then the belief is permissible.[34] But on Comesaña's view, just as on ambiguity accounts such as Sosa's, justification risks being divorced from normativity. These views allow us to say that a subject—say a biv—both is, and is not, justified, depending on what we mean by 'justified'. And someone who complains that s/he wants to know whether the subject's belief is justified *tout court*—and so whether it is epistemically permissible—is thought to have missed the point. Given indexical reliabilism, the mistake this unfortunate dupe is supposed to have made is not unlike that made by someone who, upon hearing a first person say, 'I'm hungry', and a second say, 'I'm not hungry', breaks down, exclaiming, 'But I want to know whether or not I'm hungry!' It is surely uncharitable to epistemologists disputing the soundness of the *New Evil Demon* to regard them as simply talking past each other in this way; yet this is what Comesaña's account licenses, and indeed requires us to do.

That leaves us with Cohen's problem of the *New Evil Demon*. Here is the solution that we favour: contrary to Cohen's assumption and, perhaps, appearances, Biv is not justified in believing that she has a hand, and (NED) is false. Granted, Biv is *blameless* in having the beliefs that

[33] Cohen (1984, p. 282) insists on this point, too, but draws different conclusions from it than we do.
[34] Thanks to … on this point.

she holds, but blamelessness and epistemic justification are not the same. Thus, those who feel that Biv's beliefs are justified, and that (NED) is true, conflate blamelessness and epistemic justification. But this response to the *New Evil Demon* is another story for another occasion.[35,36]

References

BonJour, L. (1980). "Externalist Theories of Empirical Knowledge." Midwest Studies in Philosophy **5**: 53-73.

Cohen, S. (1984). "Justification and Truth." Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition **46**(3): 279-295.

Comesaña, J. (2002). "The Diagonal and the Demon." Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition **110**(3): 249-266.

Comesaña, J. (2010). "Evidentialist Reliabilism." Noûs **44**(4): 571-600.

Goldman, A. I. (1988). "Strong and Weak Justification." Philosophical Perspectives **2**: 51-69.

Grice, H. P. (1989). Studies in the Way of Words. Cambridge, Mass., Harvard UP.

Hawthorne, J. and O. Magidor (2009). "Assertion, Context, and Epistemic Accessibility." Mind **118**(470): 377-397.

Heim, I. and A. Kratzer (1998). Semantics in Generative Grammar. Oxford, Blackwell.

Lewis, D. (1970). "Anselm and Actuality." Nous **4**(2): 175-188.

Lewis, D. (1973). Counterfactuals. Oxford, Blackwell.

Soames, S. (2005). Reference and description: the case against two-dimensionalism. Princeton, N.J. ; Oxford, Princeton University Press.

Sosa, E. (1993). "Review: Proper Functionalism and Virtue Epistemology." Noûs **27**(1): 51-65.

Sosa, E. (2001). "Goldman's Reliabilism and Virtue Epistemology." Philosophical Topics **29**(1/2): 383-400.

Stalnaker, R. (1978). "Assertion." Syntax and Semantics **9**: 315-332. Reprinted in: Stalnaker, R. (1999). Context and Content, Oxford: OUP, p. 78-95. Page references are to reprint edition.

Stalnaker, R. (1999). Context and Content: Essays on Intentionality in Speech and Thought. Oxford, OUP.

Sutton, J. (2005). "Stick To What You Know." Noûs **39**(3): 359-396.

Sutton, J. (2007). Without Justification. Cambridge, Massachusetts, MIT Press.

Williamson, T. (1997). "Knowledge as Evidence." Mind **106**(424): 717-741.

Williamson, T. (2000). Knowledge and Its Limits. Oxford, OUP.

[35] But see (Sutton 2005; 2007, pp. 29-35) for interesting discussion of a version of this view. We also note that there is a similarity between our view and Goldman's (1988), with the crucial difference, however, that we do not consider blamelessness a type of epistemic justification (Goldman's 'weak justification' is partly defined in terms of blamelessness). Thus, on our account, 'justified' is neither ambiguous, nor polysemous, and also not indexical in the way suggested by Comesaña.

[36] We should also note that we do not want to suggest that we are reliabilists. The problem of the *New Evil Demon*, however, also arises for a view that we do find attractive—namely, for evidentialist accounts of epistemic justification paired with the Williamsonian (2000; Williamson 1997) view that E=K.