RELIABILISTS SHOULD STILL FEAR THE DEMON

B.J.C. MADISON

ABSTRACT: In its most basic form, Simple Reliabilism states that: a belief is justified *iff* it is formed as the result of a reliable belief-forming process. But so-called New Evil Demon (NED) cases have been given as counterexamples. A common response has been to complicate reliabilism from its simplest form to accommodate the basic reliabilist position, while at the same time granting the force of NED intuitions. But what if despite initial appearances, Simple Reliabilism, without qualification, is compatible with the NED intuition? What we can call the Dispositionalist Response to the New Evil Demon problem is fascinating because it contends just that: Simple Reliabilism *is* fully compatible with the NED intuition. It is claimed that all we need to do to recognize their compatibility is appreciate that reliability is a dispositional property. In this paper I shall critically evaluate the Dispositionalist proposal.

KEYWORDS: epistemic justification, New Evil Demon Problem, reliabilism

1. Introduction

Reliabilism is a family of views about the nature of epistemic justification. In its most basic form, Simple ("Crude") Reliabilism states that: a belief is justified *iff* it is formed as the result of a reliable belief-forming process. To be a little more precise, Simple Reliabilism is understood to be an instance of *Lone World Reliabilism*, the class of views which hold that: "S's belief that *p* is justified *iff* S's belief that *p* is the output of a belief-formation process type that is reliable in *w*."[1] Specifically, Simple Reliabilism is also an instance of *Same World Reliabilism* which "identifies *w* as the world in which S forms the belief. A process's performance in a world determines its reliability there and thus, on Same World Reliabilism, determines the justificatory status of its outputs *there*."[2]

Reliabilism faces a number of objections, and several have been conflated under the label of the 'Generality Problem.' Perhaps the most discussed aspect of the Generality Problem concerns what we can call Reliability-of-What: which process

[1] Matthew Frise, "The Reliability Problem For Reliabilism," *Philosophical Studies* 175 (2018): 923-945.

[2] Ibid.

type is the relevant one to assess for reliability? The charge is that the reliabilist cannot provide an answer in a principled and non ad hoc way.[3]

But other aspects of the Generality Problem are no less pressing. For instance, reliabilists need an answer to the issue of Reliability-When: what is the relevant temporal interval to assess reliability?[4] In addition to Reliability-of-What and Reliability-When, reliabilists also need to address the issue of Reliability-Where: which worlds must the process be reliable in?[5] As noted above, Simple Reliabilism answers this question straightforwardly by endorsing Same World Reliabilism: what matters is that the process used is reliable in the very same world that it is actually used in. But many philosophers, including those sympathetic to some kind of reliabilism generally, think that this cannot be correct, due to what has become known as the New Evil Demon problem.

So-called New Evil Demon (hereafter 'NED') cases aim to show that Simple Reliabilism is false, since subjects with false unreliably produced beliefs can nonetheless still have justified beliefs.[6] From this it is concluded that reliability is not necessary for epistemic justification. For example, take a subject here in our actual world who has intuitively justified beliefs: she justifiably believes that there is a cat before her on the basis of seeming to see a cat before her; she justifiably believes what she had for breakfast on the basis of seeming to remember what she had for breakfast, and so on. Now compare this subject with her counterpart in a possible world inhabited by an evil demon of great power, so great that he could ensure that the subjects of that world have beliefs about the external world that are false, based on subjectively indistinguishable non-veridical perceptual experiences. The demon also ensures that the subject's memory beliefs are false, based on subjectively indistinguishable non-veridical memory experiences, and so on.

The subjects in the demon world, we can suppose, have all the same non-factive mental states as their non-deceived counterparts, but their beliefs are by and large false, and so presumably unreliably produced. But do the subjects of that demon world have *justified* beliefs? If so, are their beliefs justified to the exact same extent as their counterparts in a non-demon world? What I shall call the New Evil

[3] e.g. Richard Feldman and Earl Conee, "The Generality Problem for Reliabilism," *Philosophical Studies* 89 (1998): 1-29.

[4] Matthew Frise calls this the Temporality Problem for reliabilism. See Frise, "Generality,". See also Brian Weatherson, "The Temporal Generality Problem," *Logos & Episteme* 3 (2012): 117-122.

[5] Scott Sturgeon distinguishes between the reliability-of-what vs. reliability-where questions that arise for process reliabilism. See Scott Sturgeon, *Matters of Mind* (London: Routledge, 2000), 96.

[6] E.g. Keith Lehrer and Stewart Cohen, "Justification, Truth, and Coherence," *Synthese* 55 (1983): 191-207; Stewart Cohen, "Justification and Truth," *Philosophical Studies* 46 (1984): 279-95.

Demon intuition is an evaluative judgment about *sameness*: our counterparts have all the same justified beliefs as we do, despite their falsity.

Three main responses to NED cases are common. What we can call The Committed Reliabilist response maintains that we should keep Simple Reliabilism but reject NED intuitions.[7] At the other end of the spectrum, what we can call The Committed Internalist response suggests that we keep NED intuitions and reject all forms of reliabilism, thus maintaining that reliability is not necessary for justification.[8] Finally, one might endorse what we can call an Irenic Solution. An Irenic Solution is one which aims to preserve some form of (modified) reliabilism, while at the same time, granting the force of the NED intuition.

Most Irenic Solutions do this by departing from Same World Reliabilism and adopting some form of *Modal Reliabilism*: "S's belief that *p* is justified *iff* S's belief that *p* is the output of a belief-formation process that is reliable in *W*."[9] W is taken to be a special domain of worlds that need not include the same world the belief forming process is actually used in. Complicating reliabilism to accommodate the NED intuition has taken many forms. Notable examples include Normal Worlds Reliabilism;[10] Weak and Strong Justification;[11] Indexical Reliabilism;[12] Home World

[7] For example, Bach, Brewer, Engel, and Sutton have, for different reasons, denied that our demon world counterparts' beliefs are justified. See Kent Bach, "A Rationale for Reliabilism," *The Monist* 68 (1985): 246-63; Bill Brewer, "Foundations of Perceptual Knowledge," *American Philosophical Quarterly* 34 (1997): 41-55; Mylan Engel, "Personal and Doxastic Justification," *Philosophical Studies* 67 (1992): 133-51; Jonathan Sutton, "Stick to What You Know," *Nous* 39 (2005): 359-96; Jonathan Sutton, *Without Justification* (Cambridge, MA: MIT University Press, 2007).

[8] For example, on the basis of New Evil Demon cases, prominent internalists have denied that reliability is necessary for justification, such as Lehrer and Cohen, "Justification,"; Cohen, "Justification and Truth;" Earl Conee and Richard Feldman, *Evidentialism* (New York: Oxford University Press, 2004); Ralph Wedgwood, "Internalism Explained," *Philosophy and Phenomenological Research* 65 (2002): 349-369; Michael Huemer, "Phenomenal Conservatism and the Internalist Intuition," *American Philosophical Quarterly* 43 (2006): 147-158. Some epistemic *externalists* have even drawn this conclusion in response to the New Evil Demon Problem; see for example Michael Bergmann, *Justification without Awareness* (New York: Oxford University Press, 2006).

[9] Frise, "Generality," 940.

[10] Alvin Goldman, *Epistemology and Cognition* (Cambridge, MA: Harvard University Press, 1986).

[11] Alvin Goldman, "Strong and Weak Justification," *Philosophical Perspectives* 2 (1988): 51-69.

[12] Ernest Sosa, "Reliabilism and Intellectual Virtue," in *Knowledge in Perspective: Selected Essays in Epistemology* (New York: Cambridge University Press, 1991), 131-145; Ernest Sosa, "Goldman's Reliabilism and Virtue Epistemology," *Philosophical Topics* 29 (2001): 383-400.

Reliabilism;[13] distinguishing Personal v. Doxastic Justification;[14] Bergmann's version of Proper Functionalism,[15] among others.

But if one wants to endorse a form of reliabilism in the face of NED cases, is this research program into Modal Reliabilism needed? What if despite initial appearances, even the most basic and simple form of process reliabilism, without qualification, is compatible with the NED intuition? All these complications and refinements of reliabilism, insofar as they aim to reconcile reliabilism with the NED, would then be redundant and unmotivated.

What we can call the Dispositionalist Response to the New Evil Demon problem is fascinating because it contends just that: Simple Reliabilism *is* fully compatible with the NED intuition (and so versions of Modal Reliabilism, whatever their other virtues, are unmotivated by the NED problem). The Dispositionalist Response to the NED purports to be an Irenic Solution, but interestingly one that does not think that we need to modify Simple Reliabilism to reconcile it with New Evil Demon cases. It is claimed that all we need to do to recognize their compatibility is appreciate that reliability is a dispositional property. It is to the details of this proposal that I shall now turn.

2. The Dispositionalist Response to the New Evil Demon Problem

According to what I shall call the Dispositionalist Response, recently advanced by Umut Baysan, counterparts *can* have justified beliefs in a demon world because the beliefs *are* produced by a reliable belief forming process. It is just that the reliable disposition is blocked / masked in the demon world.[16] We are told that the key to the proposal is to recognize that 'reliable' is a dispositional concept, and *reliability* is a dispositional property.[17] In general, something can have a disposition, despite not manifesting it. Recognizing this general truth is meant to help us realize that victims in NED cases can have beliefs that are the product of reliable belief forming methods, but the reliability is simply not manifested, as the resulting beliefs are false.

To see that in general, one can be the bearer of a dispositional property, without ever manifesting the disposition in question, consider an analogy with the

[13] Brad Majors and Sarah Sawyer, "The Epistemological Argument for Content Externalism," *Philosophical Perspectives* 19 (2005): 257-280.

[14] Bach, "Rationale;" Engel, "Personal."

[15] Bergmann, *Justification without Awareness*.

[16] Umut Baysan, "A New Response to the New Evil Demon Problem," *Logos & Episteme* VIII (2017): 41-45.

[17] Baysan, "A New Response," 43.

property of fragility.[18] A vase can be fragile even if it never breaks. This could be simply because its owner is very careful and the fragile vase is never struck. But it is also true that something can be fragile even if it does not break when struck, even if struck repeatedly. A glass-faced smartphone is very fragile. It often breaks when dropped. But sometimes one is very lucky, and the face does not smash despite being dropped, even repeatedly. The phone need be no less fragile as a result.

Baysan suggests something similar with regard to reliable belief forming processes: a belief forming process may be disposed to produce true beliefs, but for whatever reason, at every attempt it may fail to do so. A belief forming process might instantiate the property of reliability without ever manifesting it. Baysan argues the following is a possible state of affairs: "(iv) *a* is a reliable belief-forming process; *a* is exercised; *a* doesn't produce true beliefs; this happens systematically."[19]

Baysan argues that if (iv) is possible, then it follows that one can be a Simple Reliabilist about epistemic justification and still hold that victims of the NED have beliefs that are justified: despite their systematic falsity, the beliefs are still produced by reliable belief forming processes – it is just that the reliability is not manifested in the demon world. If this is correct, we have an Irenic Solution that shows the compatibility of Simple Reliabilism and the NED intuition, without the need to resort to a form of Modal Reliabilism. I shall now critically evaluate the Dispositionalist proposal.

3. In What Sense Are Beliefs Produced by Reliable Faculties in Demon Worlds?

The key claim of the Dispositionalist Response is that NED victims can have beliefs that are produced by cognitive faculties that are reliable, but since reliability is a dispositional property, it can be instantiated without being manifested – and that is just what is going on in the NED case. But in what sense, if any, are beliefs produced by reliable faculties in demon worlds? Crucial here is what reliability amounts to. When applied to belief forming methods, "reliable" means, at the very least, that the method used is more likely to result in true beliefs than not. So the relevant sense of reliability here means reliably truth-conducive. But in what sense is a method more likely to produce true instead of false beliefs? At least two options initially present themselves, a frequency sense and a modal sense of reliability, both of which I shall argue are inadequate for the Dispositionalist Response to the NED.

First, reliability might be understood as the *frequency* of true beliefs generated by a process. Reliability in this sense is determined by the actual track

[18] Cf. Baysan, "A New Response," 44.

[19] Baysan, "A New Response," 45.

record of the belief forming process.[20] If the frequency results in a favorable ratio of true to false beliefs, then the process is a reliable one. New Evil Demon cases can be set up, however, so as to ensure a track record of falsity. So the belief forming processes employed in demon worlds are not reliable in this sense.

Furthermore, an actual high proportion of true beliefs to false beliefs is neither necessary nor sufficient for reliability. It is not necessary: suppose that a subject undergoes an operation that results in them having the perfect ability to add the sum of any numbers. Before they can exercise this incredible new ability, they die due to surgical complications. Nevertheless, for a time they possessed a reliable mathematical disposition. This is not because of anything to do with frequency, however, since the ability was never used. Rather, it seems that they possessed the reliable ability because if they *had* performed any possible addition, they *would* have always been correct.

Neither is an actual high proportion of true beliefs over false beliefs sufficient for reliability. This is because a process can be "luckily" truth conducive. For example, a student might be incredibly lucky in guessing correctly more often than not on a pop quiz. Even if such guessing resulted in a perfect score on the quiz, this would not make guessing a reliable process. A natural explanation of this is that while the student actually got all the correct answers, if they had used this same method more generally, it would presumably have resulted in many incorrect answers.

These considerations motivate a *modal* conception of reliability. According to a modal sense of reliability, a belief forming process is reliable just in case it *would* yield a favorable ratio of true to false beliefs (in some relevant range of circumstances; more on this later). But the NED case can be set up such that the beliefs of the victims of the New Evil Demon would not be reliably produced given a modal conception of reliability either, but according to the NED intuition, their beliefs would nonetheless still be justified.

To see this, suppose that, as a contingent matter of fact, all the worlds our counterparts occupy are also demon worlds. In that case, there would be no worlds where our counterparts exercise their cognitive faculties and the demon is not present. Now suppose that the demon is also not only extremely powerful, but it is also extremely evil, such that necessarily it will radically deceive its victims. It follows that the demon could ensure the falsity of these subjects' beliefs in all nearby worlds, if not all worlds, in which they exercise their cognitive faculties. Given all this, *if* the victims' belief forming processes were used, they would always result in

[20] And what temporal period is the relevant one to assess how successful the track record is? It is exceedingly difficult to say. See Frise, "Generality."

falsity (the demon sees to that). And yet, according to the NED intuition, our counterparts' same beliefs are just as justified as ours are.

So if not a frequency or modal conception of reliability, in what other sense might subjects in a NED case have beliefs produced by reliable faculties? Perhaps there is a different but related *dispositional* sense of reliability? Baysan writes, "A reliable belief-forming process is disposed to produce true beliefs."[21] And how should we understand what it is to be disposed to produce true beliefs? Baysan contends that "[…] the reliability of a belief-forming process is manifested in the fact that beliefs that are formed as a result of that process are mostly true, again, *in the right circumstances.*"[22] (emphasis added)

To see how being disposed to result in true beliefs might work, consider again fragility, a paradigm case of a disposition. Take the case of the vase again: what makes it the case that the vase is fragile? Not that it broke, or will break, or that it would break under *any* possible conditions, but presumably that it would break, *in the right circumstances*, and if submitted to the right kind of stimulus conditions. It is true that a vase can be fragile even if it never breaks, and even if it does not break when struck. I grant that even if wrapped in packing material, and so in those circumstances it cannot break, that it might still nonetheless be fragile. But this, I submit, is because it is at least *possible* for the vase to break (including in whatever counts as the right circumstances).

If, on the other hand, it is now metaphysically impossible for the vase to break, that is, if there are now no worlds where it can break, I submit that the vase is no longer fragile.[23] As a vivid illustration: if God exists and promises to protect the vase, and never let it break, it would then be metaphysically impossible for it to break. Given God's omnipresence, omnipotence, and that he always keeps His promises, there are no worlds where the vase breaks. In such a case, I suggest that the vase has lost a disposition; the disposition is not just 'super-masked.' Rather, the vase is no longer fragile.

[21] Baysan, "A New Response," 44.

[22] Baysan, "A New Response," 43-44.

[23] While most will surely agree that something can have a disposition that is never actually manifested, is it really a necessary condition of having a disposition that it at least has metaphysically possible manifestation conditions? On the possibility of dispositions with impossible manifestation conditions generally, see C.S. Jenkins and Daniel Nolan, "Disposition Impossible," *Nous* 46 (2012): 732-753; see also Jack Spencer, "Able to Do the Impossible," *Mind* 126 (2017): 466-497. Note that several of the examples in Jenkins and Nolan turn on logical or nomic necessity, rather than metaphysical necessity, which is at issue here. For critical discussion of Jenkins and Nolan, see Barbara Vetter, *Potentiality: From Dispositions to Modality* (Oxford: Oxford University Press, 2015), chapter 7.

Similarly, a belief forming process may be disposed to produce true beliefs, but for whatever reason at any attempt, it may fail to do so. But it must at least be *possible* for the process to produce true beliefs if it is disposed to be reliable. Even if not physically possible, it must be at least metaphysically possible. But in the version of the NED case I am considering where the demon happens to exist in all the worlds our counterparts do, given the Demon's omnipresence, his vast power, and his unwavering evil intentions to deceive, there are no worlds where the belief-forming process yields true beliefs – the demon sees to that.

I agree with Baysan that the following is a possible state of affairs: "(iv) *a* is a reliable belief-forming process; *a* is exercised; *a* doesn't produce true beliefs; this happens systematically."[24] But it does not follow from the truth of (iv) alone that a belief forming process remains reliable even if it *never* produces true beliefs; and even stronger, even if it is *impossible* for it to produce true beliefs. Like the vase no longer being fragile if it is impossible for it to break, I submit that if it is impossible for a process to yield true beliefs, it is not reliable either.

One might object that one has the lingering intuition that a vase can be fragile, even if it is now impossible that it breaks. If so, might a process be reliable even if it is impossible that it produces true beliefs? Even if this were granted in this particular case of fragility, there is a key disanalogy between vases and belief forming processes: a vase's fragility is presumably determined in part by its *intrinsic* properties (e.g. its microstructure), and if this is not changed, then the vase is still fragile, even if it is now metaphysically impossible that it breaks. After all, it is a fairly widely held view that dispositions are fixed by a thing's intrinsic features.[25]

But a belief being produced by a token process, one that is of a reliable type, are wholly *extrinsic* features both of the belief and the process. If two subjects are exact intrinsic duplicates, and have the same belief forming processes, these processes need not be equally reliable – for instance, the subjects might be in radically different environments, as the NED cases make vivid. Whether a process produces true beliefs is partly determined relationally. This means that whether a process is reliable necessarily depends on the environment in which it is used. Baysan seems to implicitly recognize this, as reliable belief-forming processes are described as tending to produce true beliefs, *in the right circumstances*. What would the right circumstances be, in the case of reliable belief forming processes? Presumably, at the very least, ones where there is no deceiving demon.

But accommodating this point is inconsistent with Simple Reliabilism as this new proposal really amounts to a closet form of *Modal Reliabilism*, which demands

[24] Baysan, "A New Response," 45.

[25] Cf. David Lewis, "Finkish Dispositions," *The Philosophical Quarterly* 47 (1997): 143-158.

not the actual reliability of the process used, but only reliability in a special domain of worlds – namely ones that lack, amongst other things, deceiving demons. This is to relativize reliabilism in just the way that Baysan wants to avoid. The upshot is that reliabilism still needs an answer to not only questions of Reliability-of-*What*, and Reliability-*When*, but also Reliability-*Where*. That is, reliabilists still need to type-individuate environments; not only the type of physical environment, but also the relevant range of possible worlds the process needs to be truth-conducive in.

In short, reliability is determined not only by the relevant belief forming process, but also the relevant environment, and the Dispositionalist response to the NED problem overlooks this.

4. A Value Problem for the Dispositionalist Response

Besides under-appreciating the issue of reliability-*where*, another problem for the Dispositionalist Response to the NED problem is accounting for the value of epistemic justification. By maintaining that beliefs can be fully justified, despite being actually all false, we lose a main benefit of Simple Reliabilism. Namely, traditionally Same World Reliabilism can offer a straightforward account of the value of justification: the value of justification is instrumental as a means to truth.

As Laurence Bonjour asks, "Why should we, as cognitive beings, *care* whether our beliefs are epistemically justified? Why is such justification something to be sought and valued?"[26] Bonjour thinks that the answer to these questions is obvious. He writes,

> Once the question is posed this way, the following answer seems obviously correct, at least in first approximation. What makes us cognitive beings at all is our capacity for belief, and the goal of our distinctively cognitive endeavors is *truth*: We want our belief to correctly and accurately depict the world. […] The basic role of justification is that of a *means* to truth, a more directly attainable mediating link between our subjective starting point and our objective goal. […] If epistemic justification were not conducive to truth in this way, if finding epistemically justified belief did not substantially increase the likelihood of finding true ones, then epistemic justification would be irrelevant to our main cognitive goal and of dubious worth. It is only if we have some reason for thinking that epistemic justification constitutes a path to truth that we as cognitive beings have any motive for preferring epistemically justified beliefs to epistemically unjustified ones. Epistemic justification is therefore in the final analysis only an instrumental value, not an intrinsic one.[27]

[26] Laurence Bonjour, *The Structure of Empirical Knowledge* (Cambridge, Mass.: Harvard University Press, 1985), 7.

[27] Bonjour, *The Structure of Empirical Knowledge*, 7-8; see also p. 157 for another clear expression

As is made clear above, Bonjour thinks that the value of justification is instrumental to the end of truth.[28] Simple Reliabilism can easily account for the value of justification in the following way: since truth is of value, and given Simple Reliabilism, beliefs formed by reliable processes are more likely to actually be true.

But according to the Dispositionalist Response, beliefs can be fully justified, even though all and always false. So one is left wondering: what is so great about epistemic justification? If it is also consistent with being the product of a reliable belief forming process that output belief tokens can all and always be false, then what is also so great about reliability? By having no answers to the questions of why justification and reliability are valuable, the Dispositionalist Response to the New Evil Demon loses one of the key advantages of Same World Reliabilism. This is of course not a decisive reason to reject the Dispositionalist Response, but it is a major strike against it. Having a clear and straightforward account of the value of justification would be a clear advantage over epistemically internalist views which reject reliability as necessary for justification.

In summary, I have argued that we have no adequate new solution to the problem of reconciling Simple Reliabilism with the NED intuition. The traditional options therefore remain. One can be a Committed Reliabilist and reject the NED intuition. One can be a Committed Internalist and reject reliability as a necessary condition on justification. Alternatively, one ought to seek an Irenic Solution by complicating the basic reliabilist proposal and developing some form of Modal Reliabilism consistent with NED cases.[29]

of this position.

[28] For a recent defense of the idea that the value of epistemic justification is not exhausted by its instrumental value, that justification is also valuable for its own sake, and that therefore truth value monism is false (i.e. that there is more of final epistemic value than mere true belief), see B.J.C. Madison, "Epistemic Value and the New Evil Demon," *Pacific Philosophical Quarterly* 98 (2017): 89-107.

[29] Thanks to an audience at New York University Abu Dhabi. Thanks also to Rhiannon James for helpful written comments on earlier drafts of this paper.