Reliabilism and Defeat

Word count: 13,254

May 13, 2025

Abstract

According to reliabilism, whether a belief is justified, or amounts to knowledge, depends on whether similar beliefs are, or would be, true. We consider familiar arguments that reliabilism is incompatible with defeat, and show how reliabilists can respond to them by understanding similarity in a flexible way. This alternative conception of similarity also allows reliabilists to provide a unified treatment of defeat, including higher-order defeat. However, it also over-generates justification when you learn but ignore excellent evidence for an otherwise unjustified belief. We argue that this problem is much harder, and calls for a structural revision of reliabilism. We propose what we take to be the best such revision.

1 Reliabilist Reductions

According to reliabilist theories, whether a belief is justified, or amounts to knowledge, depends on whether similar beliefs are, or would be, true.[1] Reductive theories of this kind are attractive for two reasons. First, they explain how epistemic states are related to truth — a feature that seems to set them apart from moral or aesthetic states, and gives them their distinct epistemic flavor. Second, they spell out how epistemic states are grounded in, or identical to, non-epistemic facts. By doing this, they tell us what the sources of the epistemic domain are.

How informative reliabilist theories of epistemic states are depends on how well the relevant notion of similarity between beliefs is understood.

[1] Ramsey (1931 [1926]) was, to our knowledge, the first to propose reliabilist accounts of knowledge and reasonable credence. (He took a belief to be knowledge when it is "(i) true, (ii) certain, and (iii) obtained by a reliable process" (1931 [1926], 258) and proposed that "the best degree of confidence to place in a certain specific memory feeling" depends on "how often when that feeling occurs the event whose image it attaches to has actually taken place." (1931 [1926], 92).) After Goldman (1979) popularized reliabilism about justification, the literature has exploded; see e.g. Alston (1980) and Lyons (2009) on full belief, and Dunn (2015), Tang (2016), and Pettigrew (2021) on credence. On safety theories see Williamson (2000, 2009a), Sosa (1999) and Hawthorne & Dietz (2023); on normality see Goldman (1986, 107), Stalnaker (2006), Smith (2010), Greco (2014), and Goodman & Salow (2023a,b).

This is a vexed issue. One popular claim about (the relevant kind of) similarity is that it is causal-historical: whether or not two beliefs are similar depends on what caused them to be formed and sustained. Here are two popular theories that fit this mold:

Method Reliabilism about Justification. For a belief to be justified is for the chance-expectation of the ratio of true beliefs out of those formed by the same method to be sufficiently high.

The Safety Theory of Knowledge. For a belief to be knowledge is for there to be no easy possibility (i.e. possibility that could easily have obtained) where you falsely believe something on a similar basis.

As vague as these theories of justification and knowledge may be, we think some account in their vicinity is likely to be correct.

Our topic in this paper will be a tension between such reliabilist theories of epistemic states and the possibility of defeat.[2] In a *defeat case*, someone loses an epistemic state because they receive some counter-evidence to it. To get a feel for the phenomenon, consider an example:

Measuring Mars. Una and Bianca have been sent to Mars with a reliable thermometer. They measure the temperature, and form the belief that the measured value is correct. Bianca (but not Una) then takes a second measurement, which disagrees with the first. Bianca stores its value, but ignores it and keeps her belief.

The characteristic defeatist judgment is that while initially both Una's and Bianca's beliefs are justified, Bianca's belief becomes unjustified when she takes the second measurement.[3] Of course Una and Bianca would make mistakes in exactly the same scenarios — when and only when the first measurement is mistaken. It is somewhat puzzling, then, in what sense it is more likely, could happen more easily, or would be more normal for a belief like Bianca's to be wrong. As we'll see in §2, epistemologists give different accounts of what exactly the tension consists in. One worry we will consider assumes a causal-historical account of similarity, and observes that the second measurement does not seem to be a cause of Bianca's belief, and hence cannot affect which beliefs are similar to Bianca's. A second worry is that even if ignored counter-evidence *can* matter to similarity, this may not be enough to make the belief count as unreliably formed.

[2]Goldman (1979) recognized the difficulty, but Lasonen-Aarnio (2010, 2014) prominently sharpened it, and shaped our conception of it. See also Baker-Hytch & Benton (2015), Beddor (2015, 2021), Bergmann (2005), Constantin (2020), Fraser (forthcoming), Grundmann (2009), Hirvelä (2023), and Loughrist (2021).

[3]Notice that the temporal order of the presentation isn't necessary for the point. Even if Bianca first took the ignored measurement, or took both measurements at the same time, her belief would be unjustified.

Epistemologists have reacted to this tension either by rejecting or seriously modifying reliabilism, or else by concluding that defeat is impossible or less common than generally assumed. The first contribution of this paper is to explore a more conservative solution to the problem. Rather than rule out reliabilist theories of epistemic states, defeat cases teach us something interesting about the notion of causal-historical similarity that these theories use. Very roughly, similarity of beliefs is not just a matter of similarity in their causal history, but also in the accompanying overall state of mind. Given that one of the major worries for reliabilism is that its notion of similarity is under-constrained, we should welcome any opportunity to understand it better.[4] Whether or not you end up agreeing with our take on similarity, we hope to convince you that similarity is a sufficiently flexible parameter that reliabilist theories have room for defeat.

For an analogy, consider the use of similarity in the semantics of counterfactuals. Lewis (1973) proposed (roughly) that 'If p had been the case, it would have been the case that q' is true at a world just in case q is true at the most similar p-world(s). Fine (1975) objected that intuitively 'If Nixon had pressed the button, there would have been a nuclear holocaust' may well be true, but any world involving a nuclear holocaust is (hopefully) nothing like ours. Rather than ruling out the similarity-semantics for counterfactuals, this objection is widely taken to tell us something about the relevant notion of similarity: what matters to counterfactuals is similarity with respect to the past (and specifically the *causal* history), and perhaps the laws of nature, not with respect to the future.[5] We take a similar attitude to defeat cases: rather than being counterexamples to reliabilism, they tell us something about how to fill in the parameters of reliabilist theories.

By rethinking similarity, we can treat defeat as a unified phenomenon. In particular, higher-order defeat can be treated just like other kinds of defeat. This is important because higher-order defeat has proven hard to square with natural thoughts about evidence, justification and rationality — so hard, in fact, that some philosophers have concluded that higher-order defeat is impossible, or requires one to ignore parts of one's evidence.[6]

But this solution also draws attention to a novel, major difficulty for reliabilists. If we rethink similarity in the way sketched above, then to be similar to Una's belief, a belief must be formed by taking one measurement of a comparably reliable thermometer. To be similar to Bianca's belief, a belief must be formed by taking one measurement of a comparably reliable thermometer, *while ignoring a second disagreeing measurement*. Could this be why only Una's belief ends up justified? We argue that the story needs to be more complicated. If similarity required similarity in the evidence

[4]This worry is known as the generality problem, see (Conee & Feldman, 1998).

[5]See Lewis (1979), Kaufmann (2013), Dorr (2016), and Holguin & Teitel (manuscript).

[6]See, for example, Lasonen-Aarnio (2014), and Christensen (2010).

one ignores, then an unjustified belief could become justified because one receives but ignores evidence in its favor. Call cases of this sort *uplift cases*. If we confine ourselves to *one* notion of similarity, it is hard to see how we can predict defeat without predicting uplift, too.

Once reliabilists are flexible enough about similarity, they can account for defeat by allowing ignored evidence to matter to similarity. We are skeptical that there is a solution of this kind to the problem of uplift. What uplift requires is a *structural* change in how we use similarity to account for justification and knowledge. Our ultimate view differs from simple reliabilist views like Method Reliabilism in using two notions of similarity, and two resulting notions of reliability. Nevertheless we think it retains much of the spirit of reliabilism, by reducing epistemic states to facts about objective notions like causal-historical similarity and chance.

Here's a road-map for the paper. §2 presents the tension between reliablism and defeat. §3 uses defeat cases to draw lessons about the nature of the notion of similarity relevant for reliability, and proposes a reliabilist view that can accommodate defeat. This view has the attraction of treating defeat as a unified phenomenon. §4 argues that the problem is even worse than has been appreciated so far, and that natural attempts to predict defeat — ours included — also imply uplift. We then suggest a way to complicate the structure of our view of justification in a way that predicts defeat without predicting uplift. §5 concludes with some broader lessons for the relationship between epistemic states and objective notions like chance and causal-historical similarity.

2 Defeating Reliability?

In this section, we'll explain the tension between reliabilism and defeat. We'll first make the intuitive case for defeat, and then consider a few arguments that reliabilism cannot accommodate it.

Perhaps the most convincing examples of defeat are cases of justification reversal, where you start out with a justified false belief, but are later put in a position to know that your belief is false:

Justification Reversal. You believe that it's raining based on hearing the weather report. When you look outside you can clearly see that it's in fact not raining, but stubbornly keep your belief that it's raining on the same basis.

We take the intuitive judgment about such a case to be that your belief is initially justified, but becomes unjustified once you look outside.[7]

[7]If you deny the possibility of justified false belief, or even justified belief without knowledge (Sutton, 2005), please focus on defeat cases like *Miracle Draw* below.

Completely parallel cases of knowledge reversal are of course not possible, since one cannot start out knowing something false. But there are intuitively compelling examples of defeat for knowledge, too:[8]

Miracle Draw. You look into an urn, and see that it contains a red ball, a black ball, and nothing else. On this basis, you form the belief that the urn contains a black ball and a red ball. Afterwards, you randomly draw balls from the urn with replacement. Incredibly, you draw red 1000 times in a row. You stubbornly keep your belief that there is a black ball in the bag.

We take the intuitive judgment to be that your belief is initially justified, and arguably even knowledge, but becomes unjustified once you draw red 1000 times.

In addition to these intuitive judgments, there are two more reasons to want to accommodate defeat. The first reason has to do with the downstream consequences of denying defeat. If I can know that there is a black ball in the bag after the drawing red 1000 times, then presumably I can also know that the 1000 draws are *misleading*. Similarly, in **Measuring Mars**, if Bianca is justified in thinking that the first measurement is correct, presumably she is also justified in thinking that the second measurement is misleading. But such inferences seem to involve unacceptable "bootstrapping" (Fraser, forthcoming).

The second reason to accept defeat comes from the realization that almost whenever we form a belief, our evidence is far too complex to pay attention to all of it, and so we constantly have to select which evidence to attend to. If we deny the possibility of defeat, we can't explain why a belief that p is generally unjustified if it results from selectively only attending to your reasons *for* p and ignoring lots of evidence against p. If justified belief is not undermined when you *learn* strong counter-evidence, it should also not be undermined when you *already have* strong counter-evidence. But that's absurd.

With this case for defeat on the table, let us examine whether reliabilist theories of epistemic states can accommodate the phenomenon. Since the news generally reports the weather correctly, reliabilists about justification can predict that you are initially justified in believing that it is raining in **Justification Reversal**. Similarly, since you would not easily misperceive the color of the balls, reliabilists about knowledge can predict that you start out knowing that there is a black ball in the bag in **Miracle Draw**. However, they run into trouble predicting that your beliefs cease to be justified or knowledge once you collect further evidence.

As Lasonen-Aarnio (2010) presents it, the difficulty is that it is unclear how simply retaining a reliable or safe belief could make it unreliable or

[8]This case is from Williamson (2000, 200f.).

unsafe. Now of course in defeat cases you don't just retain your belief, you also learn some evidence against it. Nevertheless, there is some intuitive sense in which collecting but ignoring some counter-evidence is no less reliable or safe than never collecting the counter-evidence to begin with. For example, there is some pressure to think that Bianca's belief in **Measuring Mars** is no less safe or reliable than Una's. Bianca and Una would get the temperature wrong in exactly the same scenarios.

We will now consider three ways to substantiate the challenge. The first way appeals to the commonly held idea that similarity is a matter of causal history. For method reliabilists, this follows from the popular idea that the *method* by which a belief is formed or sustained must be a feature of its causal history.[9] There are excellent reasons for thinking that features of the causal history are really important to the relevant notion of similarity. The fact that you could easily have falsely believed what you do, but on very different grounds, is no impediment to knowledge or justification. To see this, consider:[10]

Nozick's grandma. A grandma forms the belief that her grandchild is alive when she sees him. But the grandchild almost died in the morning, and if he had died, the parents would have told the grandma that the grandchild is well anyways and she would have believed them.[11]

The common verdict — which we take to be correct — is that the grandma is justified in believing that her grandson is alive. Given reliablism, this implies that most of the beliefs similar to the grandma's actual belief are true. This suggests that beliefs that have very different causal history — for example, those based on testimony from the parents — aren't similar to the grandma's belief, since many of those are false. So causal-historical similarity seems to be a necessary condition for similarity *tout court*.

A natural thought here would be to take causal historical similarity to also be sufficient for similarity *tout court*, and this is indeed Goldman (1979, 10)'s conclusion: "the justificational status of a belief is a function of the reliability of the process or processes that cause it." This stronger claim

[9]The causal conception of methods is at the root of reliabilism (see Ramsey (1931 [1926], 258) and Goldman (1979, 10), who motivates his theory by observing that "the justificational status of a belief is a function of the process or processes that cause it"). In the context of defeat, it has been discussed by Baker-Hytch & Benton (2015, 46f.), Beddor (2015, fn. 6) and Loughrist (2021). Loughrist (2021) argues that causally irrelevant features can be part of the method. Suppose an agent thinks that a gray object is gray because it looks gray in dim lighting. Loughrist (2021, 13856) argues that the dim lighting is part of the method, but not a cause of the agent's belief since they would still believe that the object is gray if the lighting was bright. But it is natural to think that their belief is causally over-determined — both the color and the lighting are sufficient causes of the belief, and if either had failed to occur, the other would still have caused the belief.

[10]See for example Nozick (1981, 179), Williamson (2009a, 20f.), Kripke (2011, 164f., 206f.), and Hawthorne & Dietz (2023, 151f.). Sometimes people prefer to talk in terms of the *basis*, *grounds*, or *reasons* of one's belief, rather than the method that produced the belief.

[11]Nozick (see 1981, 179).

is supported by the observation that things outside a belief's causal history usually don't seem to matter for its justification. Furthermore, we seem to be able to determine that your beliefs are unjustified by observing that they result from unreliable processes such as wishful thinking, guessing, or confused reasoning. Again, this suggests that only the reliability of the processes causing your belief matter to its justificatory status.[12]

However, if the causal history of a belief is all that matters to its reliability, and hence to its justificatory status, we run into trouble in defeat cases. The counter-evidence one receives but ignores in defeat cases need not be part of the causal history of the defeated belief. For example, we can imagine you to be so stubborn that you would never reconsider your belief about the weather, not even in the face of decisive perceptual refutation. The perceptual evidence would then appear not to be part of the method by which your belief that it is raining is formed or sustained, and hence would not affect its justificatory status for the method reliabilist.[13] Hence, if the belief was initially justified, as seems plausible, the method reliabilist would be committed to saying that it remains justified even in the face of the perceptual counter-evidence. Parallel difficulties arise for the safety theory of knowledge if the basis of a belief must be a feature of its causal history.[14]

A second way to substantiate of the challenge to reliabilism from defeat appeals to *truth-entailing* methods. Seeing, hearing, foreseeing, and remembering prima facie all entail the truth of what's seen, heard, foreseen, or remembered. If the reliabilist notion of method lines up with our intuitive classifications of how we come to know things, methods can then be truth-entailing in the sense that any belief resulting from them must be true (Nagel, 2021). Now here is the catch: if φ-ing is truth-entailing, then so is φ-ing while ignoring counter-evidence.[15] And beliefs formed by truth-entailing methods will trivially count as justified (or knowledge). For example, if your belief in **Miracle Draw** results even just partly from remembering the color, then it will come out reliable even if it results from remembering it *and* ignoring the results of the draws with replacement.

A third difficulty in accommodating defeat has to do with higher-order defeat in particular. Canonically, the defeat literature distinguishes be-

[12]Considerations like these made Goldman (1979, 9f.) accept Method Reliabilism in the first place.

[13]Constantin (2020) assumes that when you learn a defeater, the basis of the belief is replaced by a process that compares the support for the belief with the support for the defeater. We agree with Lasonen-Aarnio (2010) that it isn't clear that this would *have to* happen.

[14]Causation is also widely but not universally thought to be necessary for *basing* (see e.g. (Korcz, 2021, §1, §4), Huemer (2001)). (It has been argued that if a cause c of your belief b preempts something else a from causing b, a may still be part of the basis of b (Swain, 1979). We need not rule out this possibility—defeaters normally are not just not causes, they are also not being preempted from causing your belief.)

[15]See Lasonen-Aarnio (2010, 8) and Baker-Hytch & Benton (2015, 47).

tween three kinds of counter-evidence based on what the evidence directly or primarily tells against: counter-evidence counts as rebutting when it tells directly against the proposition initially believed, as undercutting when it tells directly against the connection between the basis for the initial belief and the belief, and as higher-order when it tells directly against the agent's ability to assess their evidence on the matter.[16] Higher-order counter-evidence is thought to pose a special difficulty: prima facie, the conjunction of your first-order evidence and your higher-order counter-evidence *still* supports your original belief. For example, consider

Hypoxia. You're on your way flying a plane to New York. Looking at your dials (D), you compute that you have enough fuel to make it (S). Right after, flight control warns you that you are at a height that puts you at a risk of hypoxia (H), a condition that would make you incapable of computing whether you have enough fuel from your dials but feels the same "from the inside". In fact, you aren't hypoxic and calculated correctly.[17]

It seems that the probability that you have sufficient fuel, given what your dials say, is high ($P(S \mid D)$ is high), just like the probability that the wall is red, given that it appears red, is also high ($P(R \mid A)$ is high). However, the probability that you have sufficient fuel, given what your dials say and that you are hypoxic, is *still* high ($P(S \mid D \wedge H)$ is still high), while the probability that the wall is red, given that it appears red and there is trick lighting, is much lower ($P(R \mid A \wedge T)$ is much lower). This contrast isn't specific to reliabilism, and is sometimes thought to make higher-order defeat a hard case for everyone. But there is a problem for reliabilists, too. If the method *inferring S from D* is reliable, then presumably so is the method *inferring S from* $D \wedge H$. So higher-order defeat in a way presents an especially recalcitrant version of the second problem.[18]

Reliabilism's struggles with defeat are well known, and have provoked three kinds of responses. Some reliabilists, such as Lasonen-Aarnio (2010, 2014), Baker-Hytch & Benton (2015), or Williamson (2024, 61), feel so committed to their favored reliabilist or safety theories of justification and knowledge that they conclude that defeat is impossible, or at any rate less common than generally believed.[19] We dislike this solution because

[16] See Pollock & Cruz (1999). The distinction is less clear than one might wish it to be, just like the relation of evidence telling *directly* or *primarily* against one thing. Since the distinction is not part of our preferred theory, we will not try to make it precise. See Kotzen (2019) for discussion.

[17] Elga (m.s., 3f.).

[18] A fourth and last sharpening of the challenge builds on substantive attempts to pin down which kind of similarity matters to for the purposes of reliabilism (see Beddor, 2015, 147f. and Grundmann, 2009, 68). We find this sharpening less compelling, since we do not find their assumptions about similarity inherently plausible.

[19] These theorists tend to offer alternative senses in which seemingly defeated beliefs, or

we are not only confident of the judgment defeated beliefs are unjustified, but also convinced that it can be accommodated within broadly reliabilist and safety theoretic frameworks — after all, the intuition that cases of defeat elicit isn't just that subjects who get defeating evidence are no longer justified, but also that they are no longer likely to get things right. When Bianca ignores ignores the second measurement, there is an intuitive sense in which beliefs *like Bianca's* aren't likely to be true. There is another sense in which beliefs like Bianca's are likely to be true, namely one in which any belief based on reading a generally reliable thermometer counts as similar. But intuitively that's not the relevant sense of similarity.

A different line of response, pioneered by Goldman (1979), is to add additional structure to reliabilist theories of justification and knowledge which is sensitive to defeat.[20] His strategy for responding to defeat builds on the observation that when a subject ignores counter-evidence, there is an alternative reliable process they could have used, namely responding to their total evidence. Had they used this alternative reliable process, they would not have given up their belief. Roughly, Goldman (1979, 20) requires for justification not only that your belief be formed by a reliable process, but also that there be no alternative reliable process of the kind just mentioned. Underlying this modification is a different perspective on justification: for a belief to be justified is for it to have resulted from an *optimal* belief-forming process, one better than any alternative way of responding to one's evidence. We think that this alternative perspective isn't particularly attractive,[21] since justification ends up not a matter of being *reliable enough*, but *as reliable as one can*.[22] Importantly, this means that justification is no longer a matter of whether similar beliefs are true. Instead, it also depends on how many possible beliefs (or other doxastic states) formed in rather different ways would be true. And while we ultimately agree that some structural modification of reliabilism is required (albeit a different one than the one proposed by Goldman), we think that we should first see how far we can get by revising our conception of similarity.

A third option, of course, is to reject Reliabilism. Naturally, internalists like Conee (1992) and BonJour (1985), and some phenomenalist Bayesians such as Schoenfield (m.s.), think this is the right lesson. These alternative theories of justification usually account for its defeasibility by appealing to another epistemic notion, such as evidence. But this just pushes the bump in the rug: arguably, evidence itself, like justification, is defeasible, so the problem would reappear for evidence.

---

the agents holding them, are epistemically sub-par. See Engel (1992, 139f.), Lasonen-Aarnio (2010, 15ff.), Baker-Hytch & Benton (2015, §5) and Williamson (2024, 61f.).

[20] See also Plantinga (2000), Bergmann (2005), Grundmann (2009), and Beddor (2021).

[21] Rejecting Goldman's strategy for evaluating belief is compatible with accepting it in the case of suspension of judgment.

[22] See Goldman (1986, 104).

To see this, suppose that for a belief to be justified is for it to be supported by your evidence, and assume that evidence was incremental (i.e. indefeasible): if something is part of your evidence at one time, it remains part of your evidence later. To handle rebutting and undercutting defeat, the view takes evidence to be sparse. Now consider:

> **Red Wall**. You're looking at a red wall, and come to believe that it is red. When a reliable friend tells you afterwards that the lighting in the room is misleading, making the wall seem red regardless of its true color, you stubbornly hold on to your belief. Though generally reliable, your friend is mistaken.[23]

When you look at a red wall, the view takes your evidence to include only that the wall appears red, not that it is red. On its own, this evidence makes it likely that the wall is red, but together with the (appearance of) trick lighting it may no longer make it likely.[24] That's why, after learning about the (appearance of) trick lighting, you are no longer justified in believing that the wall is red. Whatever the merits of this account of first-order defeat, it breaks down for higher-order defeat.[25] Call whatever the internalist takes to be your evidence E. Surely you are justified in believing E. But you might receive excellent evidence that E isn't your evidence. Intuitively, this would defeat the justification of beliefs inferred from E, including the belief in E itself. But the internalist cannot deny that E would remain part of your evidence, since they take evidence to be incremental, and anything that is part of your evidence is supported by it. Using Williamson (2000, 219)'s memorable slogan, we can put the problem for Bayesians as follows: they cannot handle higher-defeat because they "have forgotten forgetting".[26]

The upshot of this discussion isn't merely that internalism, like reliabilism, struggles to account for some defeat cases. Rather, its failure teaches us that in order to account for defeat in general, our basic epistemic state—whether it's justification, knowledge, or evidence—needs to already be defeasible. A good theory of defeat should explain how it enters the epistemic domain.

[23] From Chisholm (1966, 48).

[24] See Swinburne (2001, 29f.).

[25] See Schoenfield (m.s.) on this very problem and White (2009, 238f.) on an analogous issue in the case of peer disagreement.

[26] Indeed, the problem of higher-order evidence becomes more pressing once we adopt a sparse theory of evidence. The sparser our evidence, the more distance needs to be covered in moving from the evidence to beliefs, and so the more room for uncertainty whether you went wrong in moving from the evidence to beliefs.

3 Defeating Defeat

Beliefs held up in the face of strong counter-evidence are intuitively not just unjustified, but also of a kind that is generally unlikely to be true. This is why we find extant responses to the problem of defeat dissatisfying, and why we see it as an opportunity to better understand the sense of 'like' in which beliefs like defeated beliefs are unlikely to be true.

"Similarity" is an incredibly context-sensitive notion. In so far as reliabilists make use of the idea that some beliefs are similar, it is open to them to specify what sense of similarity they have in mind. Whatever opinions we have about concepts like "method", "basis", or "ground", nothing forces the reliabilist to cash out similarity in terms of them.

We will now argue that defeat teaches us three lessons about the relevant notion of similarity. Each can be seen as a response to a different way to substantiate the worry that reliabilism cannot accommodate defeat surveyed in §2. The first lesson responds to the worry that we should take similarity of beliefs to be fully determined by their causal histories. The second lesson responds to the worry that truth-entailing methods are similar to one another in being truth-entailing. The third lesson responds to similar problems for higher-order defeat in particular. Once all three lessons about similarity are in place, we'll sketch a version of reliablism that is defeat-friendly.

Let's start with the first lesson. Our discussion in §2 suggested that reliablism is in tension with defeat if similarity, in the operative sense, is fully determined by the causal history of the relevant belief. Our first lesson is a direct result of this discussion. Defeat cases teach us that the following principle is false:

> **Belief Causal History.** If two beliefs have the same causal history, then they are equally similar to any third belief.[27]

However, it would be unwise to throw out the baby with the bathwater. Even if the causal history of the belief is not the only similarity-relevant feature, there are excellent reasons, sketched in §2, to think that the it is important. Furthermore, methodologically, it seems desirable to stay as close to **Belief Causal History** in making room for defeat in order to keep our theory *predictive*, and *faithful* to the spirit of Reliabilism.

Rather than looking at the *local* causal history of your belief only, we could, in determining similarity, consider the *global* causal history of your entire state of mind, including any counter-evidence you have received:

> **Mental Causal History.** If two beliefs are part of overall states of mind with the same causal history,[28] and occupy the same position in that

[27] Goldman (1979, 10). See also footnote 9.

[28] When we say that the causal histories of the two beliefs are the same, we mean that they are qualitatively the same.

causal history, then they are equally similar to any third belief.

Besides focusing on the causal history of the overall state of mind, **Mental Causal History** differs from **Belief Causal History** in requiring not just sameness of the relevant causal history, but also that the belief occupy the same position within this causal history. This is needed to avoid the trivializing consequence that either none of all your beliefs at a given time are justified.

Mental Causal History allows features outside the causal history of beliefs to affect their similarity, so long as they are part of the causal history of your overall state of mind at the time. Since counter-evidence will be reflected in your overall state of mind, we are optimistic that this view is consistent with defeat. But clearly, we need to make further assumptions about similarity to also predict defeat. This is the job of our next two lesson.

Making counter-evidence relevant to similarity will make defeated beliefs come out unjustified only if we do not hold fixed certain other features of the causal history, such as whether the belief is true. This would fail, for example, on an externalist version of **Method Reliabilism** which, when things work out, takes the method by which your beliefs are formed to be truth-entailing. We know that we cannot identify similarity with sameness of method, since that would entail **Belief Causal History**. But we can consider the weaker idea that two beliefs are maximally similar *only if* they are formed by the same method, while allowing other features of your state of mind to influence similarity, too. A view of this kind will not be able to accommodate defeat.

To see this, consider again **Miracle Draw**. If the relevant method you use there is *perception*, understood in a truth-entailing way, then since all similar beliefs are formed by the same method, all similar beliefs will be true. This means that the chance that a belief is true, given that it is similar, will always be 1, no matter what else goes into similarity. This suggests that if we want to accommodate defeat by adopting the right notion of similarity, whether a belief is true, or formed by a truth-entailing method, cannot be necessary conditions on similarity. To put this in a more general form, whether two beliefs are true cannot be too important to their similarity.[29]

Note that even if one was skeptical of the possibility of defeat, one should still think that whether your belief is true cannot in general be too important to similarity. Otherwise cases of lucky success — cases where you form a true belief, but had a high chance of forming a false belief — would always count as very dissimilar from cases of failure, and so would always come out as likely to be true, since likelihood depends

[29]Problems for defeat from "too" externalist individuation of methods are discussed by Lasonen-Aarnio (2010, 8) and Baker-Hytch & Benton (2015, 47).

on chance and similarity.[30]

One might worry that there is no way to determine how important features of the causal history are to similarity without using epistemic notions.[31] What if the only natural thing in common to the causes that matter is something epistemic, such as that they are all part of your evidence? In response, we should first note that even if there was no natural non-epistemic way to pick out which features matter to similarity, there will still be some interesting implications to our theory, such as implicatures about the logic of justification (or whichever epistemic state we are giving a theory of).[32] But, second, we are hopeful that there *is* something natural in common to all of the relevant causes which isn't epistemic. While we do not know what that is, we can substantiate our optimism by giving some constraints that only rely on the causal structure, and not its epistemic upshots.

In stating these constraints, we assume that we can represent the causal history as consisting of two parts, *laws* and *history*. The *laws* specify how later parts of the causal chain depend on earlier parts, such as how likely a wall is to appear red to you, given that it is red. (Laws, in our sense, will not be laws of nature, but high-level generalizations about your perception, memory, and such.) The *history* specifies the values of the different variables part of the causal chain, such as whether the wall is red, whether it appears red, and so on. The laws seems central to whether your belief is justified; for example, how reliable your vision is seems really important to whether and which of your perceptual beliefs are justified.

What about the history? Given our commitment to **Mental Causal History**, a natural idea is to exploit the structure of the causal chains leading up to your current state of mind: all else being equal, the closer a feature is within a causal chain leading up to your current state of mind, the more important it is. Since the perceptual experiences, and memories thereof, causing you to think that one of the balls is red are closer to the belief in the causal chain, this constraint suggests that similarity of experiences and memories matters more to justification than whether or not your beliefs are true.[33]

Upshot: to determine whether two beliefs are similar, we have to answer two questions. First, do the causal chains from which those beliefs'

[30] One *could* adopt a view on which whether your belief is true is important to similarity only when your belief had a low chance of being false anyways, but in this case it seems more natural to simply hold fixed whatever makes the belief unlikely to be false.

[31] Thanks to [redacted] for raising this worry.

[32] See Williamson (2009b, 306, 312) and Hawthorne & Dietz (2023, §2.1).

[33] Nozick (1981, 184f.) assumes that "any method experientially the same, the same "from the inside," will count as the same method." Hawthorne (2007, 209f.) briefly suggests that giving an important role to experiential similarity can help with defeat, and Baker-Hytch & Benton (2015, §4.2) consider this option but object that it would result in skeptical consequences. We avoid such skeptical consequences by taking skeptical scenarios to have low chance, even if the beliefs you form in them are very similar.

states of minds result have similar laws? That is, do different parts of the causal chain depend on one another in similar ways? Second, do they have similar late history, that is are the actual values of variables appearing towards the end similar? This is our second lesson.

Our third and final lesson concerns the third difficulty mentioned in §2, which is really an especially tough version of the second difficulty involving higher-order defeat. The problem with higher-order defeat is that it seems that the original method, even when combined with the presence of higher-order counter-evidence, is still a reliable way to form beliefs. For example, *inferring from these dial readings that you have enough fuel, while being hypoxic* is arguably just as reliable as *inferring from these dial readings that you have enough fuel*.

To solve this problem, it is helpful to consider beliefs in necessary truths for a moment.

Bag of Tricks. Your logician friend has prepared a bag for you that contains one hundred little snippets, half with theorems and half with anti-theorems inscribed. You randomly draw a snippet, and form the belief that the formula inscribed is true without examining yourself whether it is true. In fact, you picked a theorem.

Since you in fact picked up a theorem, any case where you form the very same belief will be a case where you form a true belief. Since your belief is not reliably formed, an immediate upshot of cases like this is that similar cases must include ones where you end up forming a different belief — for example, ones where you draw an anti-theorem, and believe it. Unreliable mathematical inference to a true conclusion will have to be treated similarly. Reliabilists are thus anyways committed to *content variability*: similar beliefs need not have the same content.[34]

Once content variability is allowed, there is no special difficulty with higher-order defeat anymore. Your inference in **Hypoxia** resembles rare but possible cases of (quasi-)inference to an unsupported conclusion.[35] For example, in some similar cases you end up forming the belief that you have enough fuel to make it to a destination even further away. Since such cases are rare enough, they do not prevent your original belief from being justified or even knowledge. But once we zoom in on the cases where you are told that you are hypoxic, a lot more of the (quasi-)inferences involved go wrong, and that's why your later belief is no longer justified and fails to be knowledge.

[34] See Williamson (2009a, 23f.).

[35] It is unsurprising that we should have to look to cases of failures of logical omniscience to account for defeat. For an agent who was never at any risk of irrationally responding to their evidence, the best response to evidence E would include being certain that they rationally responded to E, leaving no room for higher-order uncertainty. It's only because we can respond to evidence irrationally that higher-order defeat is possible for us.

Upshot: beliefs can be similar even when their contents are different. In higher-order defeat cases, it may be true that most beliefs formed in the same way with the same content are true, but many similar beliefs are still false, because they have different contents. That's our third lesson.[36]

We propose that reliabilists should care about a notion of similarity which considers both whether beliefs are parts of a global state of mind with a similar causal history, and whether they occupy a similar position in that causal history, but without holding fixed their contents. We use similarity in a place where you may have expected us to use *identity*. Method Reliabilists sometimes take beliefs to be similar when they are formed by the *same* method. But arguably whether two beliefs are formed by the same method is in turn a matter of causal-historical similarity. Even worse, it suggests (in our view, misleadingly) that the relevant notion of causal-historical similarity is an equivalence relation.

With this in mind, here's our first stab at a defeat-friendly reliabilism:

Global Reliabilism about Justification. For a belief to be justified is for the chance-expectation of the ratio of true beliefs out of those *globally similar* to be sufficiently high.

It would be good to understand better what global similarity is, just like it would in general be good to be able to solve the generality problem. But for our purposes here, what's important is that Global Reliabilism is compatible with defeat. To see this, consider a simple defeat case like **Measuring Mars**. The causal history of Una's entire state of mind includes the first measurement, as well as some "laws" connecting the environmental temperature to thermometer readings. Beliefs formed against the backdrop of an overall state of mind with similar causal history tend to be true, and so Una's belief is justified and (if things work out) knowledge. The causal history of Bianca's entire state of mind, on the other hand, includes both the first *and* the second measurement, and similar background "laws" about the thermometer. Beliefs formed against this more inclusive backdrop are only true about half the time, and hence Bianca's belief is unjustified (and falls short of knowledge).

The important aspect of Global Reliabilism for our purposes is that it considers *global* similarity. Various alternative recipes for defining justification from chance and similarity would work just as well for us, provided they use global similarity. In particular, we might take a belief to be justified when it's truth-ratio is sufficiently high across the *most likely* worlds, or the worlds *at least as likely* or *not sufficiently less likely* than the actual

[36] In §5 we will ultimately argue that similarity is degreed. This leaves open a view on which inferential beliefs are much less similar, in the relevant sense, to quasi-inferential beliefs, than quasi-inferential beliefs are to inferential beliefs.

world.[37] These techniques are in fact required to give a plausible theory of knowledge. Since what's known is true, the ratio definition will not work unless the threshold in question is one, but this would result in an indefeasible view of knowledge *unless* one uses something like these techniques. We will not get into these details, since they are orthogonal to our questions about defeat, but see the references in the previous footnote.

One striking feature of Global Reliabilism is that it treats what have traditionally been considered different kinds of defeat the same way. Recall that the defeat literature distinguishes between rebutting, undercutting, and higher-order counter-evidence. From a reliabilist point of view, there is no deep difference between these three kinds of counter-evidence: since reliabilists do not care about the particular content of the evidence, but only how it correlates with getting things right, they have no need for such a distinction. For reliabilists, there is no deep difference between evidence that your thermometer is malfunctioning, that your vision is malfunctioning, or that your own abilities for evaluating your evidence are malfunctioning. They are all just different instruments, though more or less central to one's overall epistemic functioning.

This implication of reliabilism seems correct — it seems natural to treat all cases of defeat in the same way, as our intuitions don't differ between them, and as it's plausible to describe every case of defeat as a case where the subject isn't likely to get things right anymore (whether this likelihood is understood objectively or subjectively).[38]

Global Reliabilism contrasts here with Bayesian accounts of defeat. If a belief is justified in virtue of its content being likely on your evidence, then whenever your original evidence in fact entails its content, and your new evidence strengthens your original evidence, your belief will remain justified at the later time. This makes trouble in cases of higher-order defeat like **Hypoxia**.[39] Bayesian accounts of justification fail to model that humans are sometimes imperfectly sensitive to evidential support relations, which is why they struggle to account for failures of logical omniscience. This very same failing also explains why Bayesians struggle with higher-order defeat, since such cases also crucially depend on our imperfect sensitivity to evidential support relations.

A second important feature of Global Reliabilism is its reductive nature. It not only attempts to predict the result that epistemic states are defeasible, but also to explain in virtue of what they are defeasible. Think of an account of justification which takes it to be explained by evidence,

[37] See Goldman (1986, 107), Stalnaker (2006), Smith (2010), Greco (2014), and Goodman & Salow (2023a,b).

[38] We are in agreement with Maria Lasonen-Aarnio (2014, 315) here: "I see no significant difference between intuitions elicited by more familiar cases of defeat and those elicited by cases involving higher-order evidence."

[39] See Schoenfield (m.s.).

but takes evidence to be irreducible. This kind of account might be able to explain why justification is defeasible, and it might be able to predict that evidence is defeasible too — by sheer stipulation — but it's hard to see how it can explain why evidence is defeasible, given that evidence isn't explained in any other terms. The reliabilist ambition is *reductive*, and so we should try to find the root of defeasiblity in epistemology. On Global Reliabilism, the justification for your belief is defeated when you receive strong counter-evidence because it stops being similar to beliefs held in the absence of comparable counter-evidence. Again, identifying what exactly this notion of similarity is can be hard, but we provided some natural constraints, and there's no principled reason why no notion of similarity could satisfy all the desiderata.

Third, Global Reliabilism is holistic in two senses. On the one hand, it doesn't treat any part of one's state of mind as directly supporting or undermining any particular conclusion, as some views of defeat in terms of reasons do (see §4). Rather, it has the desirable implication that which evidence supports which conclusion, or undermine the support of some evidence for some conclusion, depends on one's background evidence.[40] On the other hand, Global Reliabilism doesn't predict a uniform verdict about higher-order evidence across cases, as either always creating defeat, or as always losing to the first-order considerations and leading to no defeat. Whether higher-order counter-evidence leads to defeat in a particular case depends on whether, given all other similarity-relevant features, it makes too few of the similar beliefs be true in expectation. This means that whether higher-order defeat occurs depends both on the counter-evidence itself and on the rest of the information available to the subject, rather than on merely whether the counter-evidence is higher-order. And this is as it should be: if what we care about when we care about justification is our likelihood to get things right, then a piece of counter-evidence should matter only in as much as it affects this likelihood in the relevant case.

For all these reasons, it would have been great if Global Reliabilism was correct. Unfortunately, it is not. In the next section, we first argue that accommodating defeat in this way introduces a revenge problem. Then, we suggest that the best way to solve this problem is by complicating the structure of our reliabilist view.

[40] To see most vividly why this is a desirable feature, consider the following case: your friend gives you, at two different occasions, two different excuses for why she missed your birthday party. Each excuse on its own might be pro-evidence for the conclusion that she wanted to come to the party (but couldn't), but both excuses together act as counter-evidence to this claim — coming up with too many excuses gives rise to the suspicion that she didn't want to be there.

4 The Symmetry Problem

Much of the debate about reliabilism and defeat has centered on the question whether ignored evidence can, after all, make a difference to the method by which one's belief is formed or sustained, or the basis on which it is held.[41] We think that this is misguided, to some extent. The core of the reliabilist project, as we see it, is the thought that whether a belief is justified, or amounts to knowledge, depends on whether similar beliefs are, or would be, true. Whatever we think about the notion of *method* or *basis*, the notion of similarity involved in the statement of reliabilism is flexible enough that it may well be affected by ignored evidence. This is exactly what §3 showed. However, we think that once you allow ignored evidence to affect similarity, a hard problem appears. If we account for defeat by letting ignored evidence influence similarity, ignored evidence will be able to affect one's justification not only for the worse, but also for the better. And this symmetry between defeating and improving one's justification is really hard to break. We will explain the problem, and propose a way to structurally modify reliabilism to break the symmetry.

Assume, then, the view sketched in §3, which means that we can require that for two beliefs to be similar, the evidence *ignored* in forming and sustaining them must be similar. The problem with this proposal is this: if similarity required similarity in the evidence one ignores, then an unjustified belief could become justified because one receives but ignores evidence in its favor. But it seems like it couldn't, so evidence one ignores in forming beliefs does not always seem to influence their similarity. To see this, consider the inverse case of defeat, one where you start out having an unjustified belief, and gain some excellent evidence in its favor that you ignore. We will call examples of this sort *uplift* cases. Our intuition is that in uplift cases, your belief typically does not become justified. Here is an example to elicit intuitions:

Wishful Thinking. You dislike somebody, which makes you believe that something they said was dumb. When you gain excellent reasons to think it was indeed dumb, you ignore this evidence, and keep your belief based on your dislike.

The proposed account of defeat is prone to over-generate justification in uplift cases, predicting that beliefs become justified when you learn but ignore evidence in their favor. Since in many cases where you ignore comparable evidence in favor of your belief, you get things right, your belief would be predicted to be reliably formed and hence justified. So once we let ignored evidence play a role in determining the justificatory status of a

[41] See Lasonen-Aarnio (2010, 5-8), Baker-Hytch & Benton (2015, 45-9), and Beddor (2015, 147f.).

belief, we're left with a worry: if ignored evidence can affect one's justification for the worse, why can't it affect one's justification for the better? Is there a principled way to break the symmetry?

Other reliabilists face this problem, too. Swinburne (2001, 29) requires that in order for beliefs to be similar, they have to be formed by a similar method *in a similar environment*. Ignored evidence is allowed to affect similarity by changing the environment, even when it does not change the method by which the belief is formed. But whatever Swinburne's notion of an "environment" is, it is not clear why ignored evidence would be part of the environment only when it is evidence against one's belief.

In fact, the general problem of specifying how ignored evidence affects one's justification is not specific to reliabilism. Indeed, a version of this problem has been discussed in the context of evidentialist theories of justification.[42] On the one hand, since our evidence is often very complex, it is unrealistic to expect people to always be sensitive to all their evidence. On the other hand, it is natural to think that blatantly ignoring decisive counter-evidence to one's belief, as in **Justification Reversal**, makes beliefs unjustified. Even ignoring our reliabilist commitments, drawing a principled and nevertheless plausible line here is a tall order.

To appreciate the difficulty, consider some natural initial attempts to specify how a belief must be related to one's evidence to be justified. One natural thought is that it must not only be based on a subset of one's evidence which does in fact support the belief, but also that one's total evidence supports the belief. One may even suggest that a subject which only fulfills the second conjunct is propositionally justified, while a subject who fulfills both is doxastically justified.

This simple conjunctive condition, however initially promising, seems too weak, as illustrated by the following example:

Coins. You are about to flip a random coin many times. You believe that it will land heads about half the time because you think it is fair. Once you flip it a few times, it alternates: on every odd flip it lands heads, on every even flip it lands tails. You keep believing that the coin is fair, and on this basis keep believing that the coin will land heads about half the time.[43]

In fact, you total evidence here does support that the coin will land heads about 50% of the time, since it supports that the coin is alternating. But

[42] "[A] well-founded attitude need not be based on a person's whole body of evidence. What seems required is that the person base a well-founded attitude on a justifying part of the person's evidence, and that he not ignore any evidence he has that defeats the justifying power of the evidence he does base his attitude on. It might be that his defeating evidence is itself defeated by a still wider body of his evidence. In such a case, the person's attitude is well-founded only if he takes the wider body into account." (Feldman & Conee, 1985, 33, n.21)

[43] Thanks to [redacted], who helped constructing this case.

intuitively we would not want to ascribe justification or knowledge to you.

A second natural thought, pursued by Feldman & Conee (1985), would be to require that for a belief to be justified, it must be based on a subset of one's evidence such that every superset of it that is still a subset of one's total evidence supports the belief. This condition is strictly stronger than the conjunctive proposal, and avoids the counterexample, but arguably at the cost of being too strong. Here is another case to bring this out:

Students. You are teaching a class of 40 students, and you are wondering if the class is going well. You *could* satisfy yourself that it is going well by considering all the students individually, and then ensuring that the percentage of students doing well is above 80%. Instead, you randomly select 10 students, and compute that 90% of them are doing well. If in fact 34 of your 40 students are doing well, this would seem like another way to form a justified belief that your class is going well.

On Feldman & Conee (1985)'s proposal your conclusion would be unjustified, since there is a superset of your evidence — evidence additionally specifying the progress of the 5 other students who are struggling — which doesn't support your belief. On this evidence, only 60% of your students are doing well, and so it does not support the conclusion that your class is going well. But we take your belief to be justified, and so reject Feldman & Conee (1985)'s proposal. When your evidence is sufficiently complex, taking evidential shortcuts by randomly sampling one's evidence seems compatible with justification even if some (carefully gerrymandered) superset of the evidence one looks at fails to support one's belief.

There is a wide range of possible views in the vicinity here which evidentialists could try out, but for now we wish to make a more limited point. Whether one is an evidentialist or a reliabilist, it is hard to give a plausible informative theory of which ways of ignoring evidence are compatible with justified belief.

We think that the contrast between defeat and uplift reveals that reliabilists need more structure to their view. Similarity of beliefs cannot be a matter of similarity of method only, since that fails to account for defeat. It cannot be a matter of global similarity in general, since that would predict uplift. While there are various intermediate notions of similarity one might try out, we have not been able to find one that rules out defeat without predicting uplift. Instead, we think uplift calls for a structural modification to the reliabilist view.

Once we structurally complicate reliabilism, there is a wide array of options to choose from. We might consider alternative ways the subject *could have* formed a belief, bring *reasons* into reliabilist theorizing, or consider whether beliefs result from dispositions conducive towards justification,

knowledge, or rationality. We will briefly explain what we do and do not like about these theoretical options, and then propose our own.

As explained in §2, the response to defeat with most historical precedent considers other ways that an agent might have formed beliefs. Abstractly, perhaps defeated beliefs are unjustified because there is another way the agent might have formed doxastic attitudes that achieves more accuracy (in chance-expectation, in the long run, or some such). A good feature of such comparative views is that they can explain straightforwardly what's wrong with an agent who wonders whether p is true, has excellent evidence in favor of p, and yet suspends judgment as to whether p. Their suspension of judgment is unjustified because forming the belief that p would on average result in more accurate beliefs.

What we don't like about such comparative views of justification is that they make justification a matter of optimizing, rather than satisficing. It seems to us that you can know something in one way even if there is an even more reliable or safer way to arrive at the same conclusion. For example, it seems perfectly okay in **Students** to arrive at a conclusion about all students by considering your evidence about a random sample of students. One might try to avoid this problem by restricting attention to alternative reliable processes that would have resulted in a different doxastic attitude. We think that this modification still spells trouble in cases like **Students**, where you are intuitively taking a *sufficiently* safe evidential short-cut. Most but crucially not all ways of taking more evidence into account would result in the same doxastic attitude, and so there is another way you might have formed doxastic attitudes that would have resulted in suspension of judgment and would lead to more accuracy on average.

A second sort of response tries to bring *reasons* into the reliabilist framework.[44] There are well-established theories of defeat building on the ideology of *reasons* that primarily or directly support conclusions, where further reasons can undermine the relationships between reasons and the conclusions they support, and conclusions can in turn primarily or directly support some further conclusions. We would like to do without the ideology of *reasons*, *directly supporting*, and *undermining*. There are two reasons for this. First, we want a reductive theory, and all three notions strike us as paradigm epistemic notions, and ones we do not obviously understand better than justification itself. Second, we worry that the reasons framework posits structure that may not be there. Which reasons support which conclusions, or undermine the support of other reasons for some conclusion, depends on one's background evidence.[45] Moreover, how one's evidence should be broken up into reasons may depend on which other reasons one has.

[44]See Beddor (2015).

[45]See footnote 40.

Another resource philosophers sometimes bring in are *success-conducive dispositions*, where success could be understood as knowledge, justification, or rationality.[46] In a way reliabilists already have a concept like this, but their measure of success is truth. Success-conducive dispositions of the new kind can employed in different ways. Some philosophers use it to define a secondary evaluative notion, reasonableness, which is independent of justification and what defeated beliefs really lack. Others use success-conducive dispositions to understand justification itself, and maintain that defeated beliefs are unjustified.[47]

Views which only use success-conducive dispositions to define a secondary form of evaluation, different from justification, seem unattractive to us because they give us the resources to define a notion of justification which would seem to track case judgments better. Call what theorists of the first kind take to be justification *proto-justification*. We worry that we can improve on the proto-justification-theory by letting justification be *reasonable proto-justification*, i.e. roughly proto-justified beliefs resulting from proto-justification-conducive dispositions. Indeed, Wedgwood (2022) proposes an account of justification with similar structure. A worry about theories of both kinds concerns the claim that a belief can be justified only if it is the result of a success-conducive disposition. If dispositions are understood as temporally and modally stable feature of a person, then it seems to us that someone could have justified beliefs that do not result from success-conducive dispositions. Just like someone who is not in general disposed to do morally good acts may nevertheless be justified in doing something good, someone who is not in general disposed to form epistemically good beliefs may be justified in forming an epistemically good belief.[48] If dispositions are allowed to be temporally and modally unstable, then the view becomes a lot closer to our preferred account.

Our own best attempt at solving the symmetry problem takes the missing ingredient in reliabilism to be that of explanatory reason. Let's distinguish between two notions of reliability. Let a belief be globally reliable iff the chance-expectation of the ratio of true beliefs out of those *globally similar* to it is high, where beliefs are globally similar when they are parts of a global state of mind with a similar causal history, and occupy a similar position in that causal history. Let a belief be locally reliable iff the chance-expectation of the ratio of true beliefs out of those *locally similar* to it is high, where beliefs are locally similar when they have similar causal history. (What we call 'local reliability' is the notion of reliability used by Method Reliabilism to define justification.) With those in hand, we propose the following account of justification:

[46] See Lasonen-Aarnio (2010, 2021) and Wedgwood (2022).

[47] In the latter case, it is important that the notion of success isn't doxastic but propositional justification.

[48] Markovits (2010, 210) makes this point about moral worth.

Explanatory Reliabilism For a belief to be justified is for it to be globally reliable and locally reliable for a (sufficiently) shared reason.

Notice that 'reason' above should be read as explanatory, rather than normative, reason. Notice further that this view leaves it open how exactly the two kinds of reliability are connected. This is a feature, not a bug: we think that what's needed is an explanatory alignment, and this can be achieved either by the local reliability explaining the global one, or vice versa, or when the two have a common explanation.

To see how the view can solve the symmetry problem, consider the different kinds of cases we discussed throughout the paper. In normal defeat cases, such as **Measuring Mars**, **Explanatory Reliabilism** implies that Bianca isn't justified because she's not globally reliable. In normal uplift cases, such as **Wishful Thinking**, **Explanatory Reliabilism** implies that you aren't justified because you're not locally reliable. And finally, in weird cases such as **Coins**, the view implies that what is wrong with your belief is that although it is both locally and globally reliable, the reasons for these two kinds of reliability are divergent. It's only accidentally that the two causal-historical routes give the same verdict. And finally, **Explanatory Reliabilism** implies that you justified in **Students**, and similarly in other random sampling cases, because in those cases there's an explanatory connection between the local causal-history of one's belief and its global causal history.

On the one hand, we are somewhat disappointed that we couldn't find a solution to the symmetry problem that doesn't require making structural changes to reliabilism. On the other hand, we do think that **Explanatory Reliabilism** is intuitively compelling, and that it keeps with what we take to be the core motivations of reliabilism: close connection to the truth and reductionism. We conclude the paper by discussing the appeal of the reduction base in more detail in §5.

5 The Epistemic and the Objective

One of the central ambitions of reliabilist theories of epistemic states is to explain how epistemic facts supervene on non-epistemic facts. On our preferred theory, epistemic facts reduce to facts about *chance*, *causal-historical similarity*, *explanation*, and *which beliefs are true*. In this section, we try to motivate two of the components of the reduction base: chance and causal-historical similarity. (The truth component was motivated in §1, and the explanation component was motivated in §4.)

To motivate the appeal to chance, we first argue that we need an asymmetric feature in our reduction base.[49] Consider the relation of being com-

[49]Goodman & Salow (2023b, 97-98) nicely motivate this, too.

patible with your knowledge — that is, Rwv holds iff you don't know in w that you are not in v.[50] One salient feature of this relation is that it is asymmetric: if things go well, you can know quite a lot, for example that a fair coin will not land heads 100 times in a row, or that it will land heads around half the time. But if you had been unlucky enough to face a fair coin that will land heads 100 times in a row, you would have known much less, for example you would not have known that the coin would not land heads around half the time. Though it may be harder to elicit intuitions here, similar thoughts seem plausible in the case of skeptical scenarios. This asymmetry in your knowledge suggests that our supervenience base for knowledge needs to include an asymmetric component. Since similarity relations are generally symmetric, a purely similarity-based theory of knowledge will struggle to break the symmetry between skeptical and non-skeptical possibilities. On our theory, chance plays this symmetry-breaking role: the coin is much more likely to land heads around half the time than it is to only land heads.

Why think epistemic facts supervene partly on chance facts, and not on some alternative asymmetric modal component? Imagine a conspiracy theorist who, prior to learning any information, failed to treat chance as an expert: they expect to live in a world where high-chance events rarely happen, and low chance events happen all the time. The credences of our conspiracy theorist, and the accompanying beliefs, seem unjustified. What exactly is wrong with the beliefs of such a person? Of course, their beliefs have a high chance of being inaccurate. But, given their distrust of chance, our conspiracy theorist would treat this not as a reason to revise them, but as an extra reason to have them, since they think high chance events rarely happen. So whatever the epistemic fault involved, it is not one apparent in view of their own beliefs.[51] On our theory, justification is, by its nature, tied to chance: whether or not a belief is justified depends on the chance that similar beliefs would be accurate. We propose, then, that what is wrong with our chance conspiracy theorist is that their beliefs are unjustified, because disbelieving something on the basis that one knows it to have high chance results in false beliefs most of the time. To generalize, in order to predict that epistemic states such as knowledge, justification, or rational credence are connected to chance in the right way, we need to

[50] w and v here should be understood as centered possible worlds.

[51] Pettigrew (2012) tries to get around this problem by assuming that the fundamental epistemic value is having one's credences match the chances. But as Pettigrew (2016, §9.4) points out, the fundamental epistemic value seems to be accuracy, understood as closeness to the truth, and we care about having our credences match the chances only as a means to having accurate beliefs.

think that they themselves are partly determined by chance facts.[52][53] A second reason to want to bring in chance is that knowledge, justification, and the like are things that we care about, and so their supervenience base should consist of things that matter. In so far as one cares whether a belief is true, it is also natural to instrumentally care whether a belief is formed in a way that is likely to result in true beliefs.

Next, we want to motivate the appeal to similarity. Another salient property of compatibility with one's knowledge is that it arguably fails to be transitive. When I see a blue piece of cloth, I will not be able to identify the shade of blue exactly: for all I know it is a little brighter or darker than it really is. Had the cloth been a little darker, for all I know it could have been a little darker still. And so on. And yet, I do know that the cloth is not *much* darker than it is.[54] Similarity, of course, fails to be transitive in just the same way: a light shade of blue is similar to a slightly darker shade of blue, which is in turn similar to a slightly darker shade of blue still, and so on. And yet, it is not similar to a *much* darker shade of blue. Similarity thus seems like a natural candidate to explain why compatibility with one's knowledge is not transitive.

A further advantage of appealing to similarity is in the treatment of beliefs in necessary propositions, as was briefly discussed in the third lesson of §3. Since necessary propositions could not have been false, there is a salient sense in which a belief in a necessary proposition could not have been mistaken, at least if beliefs are individuated in terms of their content. Bringing similarity into the mix can help here: on our view, what matters is not whether the very same belief could have been false, but whether enough similar beliefs are false, in expectation. To see how this helps, recall **Bag of Tricks**. If I were to randomly draw a snippet from the bag, and come to believe the mathematical claim inscribed, I would fail to know the mathematical claim even if it was a theorem. The natural explanation of this fact that we propose in §3 is that my belief is similar to a belief with

[52] One of the most plausible constraints on rational credence is the principal principle, which says roughly that conditional on the chance of p being t, you should assign p a credence of t. (Roughly, because some care is needed in cases of "inadmissible" evidence, that is evidence that goes beyond the information the chances have. See Lewis (1980).) Less precise but structurally similar connections between knowledge and chance seem plausible, too. If we want such systematic connections between epistemic facts and chance facts to fall out of our reduction of epistemic facts to non-epistemic facts, we will need to bring chance into the mix in some way.

[53] Williamson (2009a, 14f.) worries that chance ideology cannot be the right ingredient to a reliabilist reduction of knowledge because it would trivialize knowledge in a deterministic world. We think that if it turned out that our world is deterministic, we would still be saying something true when we say of an ordinary coin that it has half a chance of landing heads. The existence of non-trivial chance, in the sense we are interested in, is compatible with determinism. For example, even if the microscopic facts, together with the laws of nature, metaphysically necessitate the future, the macroscopic facts only together with the laws need not, and so conditioning on those only may result in non-trivial chances.

[54] See Williamson (2000), who uses the non-transitivity of indiscriminability to argue for the possibility that one can know something without knowing that one knows it.

a different content, one that I could, with substantial chance, have formed had I drawn a snippet with an anti-theorem inscribed, and hence would have believed a falsehood.

Throughout the paper, we have treated chance as degreed and similarity as binary. Ultimately, we think that both notions should be treated as degreed. To illustrate why, consider minimal pairs of cases where you are prone to make certain kinds of mistakes. In the first case, the mistake would have arisen with small probability in a fairly similar case. In the second, it would have arisen with slightly higher probability in a slightly less similar case. Intuitively, assuming all else is equal, your beliefs in two such cases should be similarly justified. For example:

Tree. Mr. Magoo is looking at a tree in the distance, trying to estimate its height. In fact, the tree is 30.5 meters tall. Mr. Magoo forms the belief that the tree is between 30 and 31 meters tall.

(a) There is a small probability p that Mr. Magoo would have made a mistake had the tree been a few centimeters ϵ taller than it is.

(b) There is a slightly higher probability $p' > p$ that Mr. Magoo would have made a mistake had the tree been a few more centimeters $\epsilon' > \epsilon$ taller than it is.

Intuitively, an unlikely mistake in a similar case is comparable to a slightly more likely mistake in a slightly less similar case. This motivates us to think that chance and similarity are really degreed, and can be traded off against one another. We have ignored this complication above since it is, as far as we can tell, orthogonal to the difficulties reliabilists face with defeat cases. But it should be taken into account in the ultimate account of knowledge, justification, and the like.

References

Alston, William P. 1980. Level-confusions in epistemology. *Midwest Studies in Philosophy* 5(1). 135–150. doi:10.1111/j.1475-4975.1980.tb00401.x.

Baker-Hytch, Max & Matthew A. Benton. 2015. Defeatism defeated. *Philosophical Perspectives* 29(1). 40–66. doi:10.1111/phpe.12056.

Beddor, Bob. 2015. Process reliabilism's troubles with defeat. *Philosophical Quarterly* 65(259). 145–159. doi:10.1093/pq/pqu075.

Beddor, Bob. 2021. Reasons for reliabilism. In Jessica Brown & Mona Simion (eds.), *Reasons, justification, and defeat*, 146–176. Oxford University Press.

Bergmann, Michael. 2005. Defeaters and higher-level requirements. *Philosophical Quarterly* 55(220). 419–436. doi:10.1111/j.0031-8094.2005.00408.x.

BonJour, Laurence. 1985. *The structure of empirical knowledge*. Cambridge: Harvard University Press.

Chisholm, Roderick M. 1966. *Theory of knowledge*. Englewood Cliffs, N.J., Prentice-Hall.

Christensen, David. 2010. Higher order evidence. *Philosophy and Phenomenological Research* 81(1). 185–215. doi:10.1111/j.1933-1592.2010.00366.x.

Conee, E. & R. Feldman. 1998. The generality problem for reliabilism. *Philosophical Studies* 89(1). 1–29. doi:10.1023/a:1004243308503.

Conee, Earl. 1992. The truth connection. *Philosophy and Phenomenological Research* 52(3). 657–669.
Constantin, Jan. 2020. Replacement and reasoning: A reliabilist account of epistemic defeat. *Synthese* 197(8). 3437–3457. doi:10.1007/s11229-018-01895-y.
Dorr, Cian. 2016. Against counterfactual miracles. *Philosophical Review* 125(2). 241–286. doi:10.1215/00318108-3453187.
Dunn, Jeff. 2015. Reliability for degrees of belief. *Philosophical Studies* 172(7). 1929–1952. doi:10.1007/s11098-014-0380-2.
Elga, Adam. m.s. Lucky to be rational.
Engel, Mylan. 1992. Personal and doxastic justification in epistemology. *Philosophical Studies* 67(2). 133–150. doi:10.1007/bf00373694.
Feldman, Richard & Earl Conee. 1985. Evidentialism. *Philosophical Studies* 48(1). 15–34. doi:10.1007/bf00372404.
Fine, Kit. 1975. Critical notice of lewis, counterfactuals. *Mind* 84(335). 451–458.
Fraser, Rachel. forthcoming. The will in belief. In *Oxford studies in epistemology*, Oxford University Press Oxford.
Goldman, Alvin I. 1979. What is justified belief? In George Pappas (ed.), *Justification and knowledge*, 1–25. D. Reidel.
Goldman, Alvin I. 1986. *Epistemology and cognition*. Cambridge: Harvard University Press.
Goodman, Jeremy & Bernhard Salow. 2023a. Belief revision normalized.
Goodman, Jeremy & Bernhard Salow. 2023b. Epistemology normalized. *Philosophical Review* 132(1). 89–145. doi:10.1215/00318108-10123787.
Greco, Daniel. 2014. Could KK be OK? *Journal of Philosophy* 111(4). 169–197. doi:10.5840/jphil2014111411.
Grundmann, Thomas. 2009. Reliabilism and the problem of defeaters. *Grazer Philosophische Studien* 79(1). 65–76. doi:10.1163/18756735-90000857.
Hawthorne, John. 2007. A priority and externalism. In Sanford Goldberg (ed.), *Internalism and externalism in semantics and epistemology*, 201–218. Oxford University Press.
Hawthorne, John & Christina H Dietz. 2023. The safety conception of knowledge. In Luis R. G. Oliveira (ed.), *Externalism about knowledge*, 150–185. Oxford University Press.
Hirvelä, Jaakko. 2023. A virtue reliabilist error-theory of defeat. *Erkenntnis* 88(6). 2449–2466. doi:10.1007/s10670-021-00462-1.
Holguin, Ben & Trevor Teitel. manuscript. On the plurality of counterfactuals.
Huemer, Michael (ed.). 2001. *Skepticism and the veil of perception*. Lanham: Rowman & Littlefield.
Kaufmann, Stefan. 2013. Causal premise semantics. *Cognitive Science* 37(6). 1136–1170. doi:10.1111/cogs.12063.
Korcz, Keith Allen. 2021. The Epistemic Basing Relation. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Metaphysics Research Lab, Stanford University Spring 2021 edn.
Kotzen, Matthew. 2019. A formal account of epistemic defeat. In Rodrigo Borges, Branden Fitelson & Cherie Braden (eds.), *Knowledge, scepticism, and defeat: Themes from klein*, Springer Verlag.
Kripke, Saul A. 2011. *Philosophical troubles: Collected papers, volume 1*, vol. 1. OUP USA.
Lasonen-Aarnio, Maria. 2010. Unreasonable knowledge. *Philosophical Perspectives* 24(1). 1–21. doi:10.1111/j.1520-8583.2010.00183.x.
Lasonen-Aarnio, Maria. 2014. Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research* 88(2). 314–345. doi:10.1111/phpr.12090.
Lasonen-Aarnio, Maria. 2021. Dispositional evaluations and defeat. In Jessica Brown & Mona Simion (eds.), *Reasons, justification, and defeat*, 91–115. Oxford University Press.
Lewis, David. 1979. Counterfactual dependence and time?s arrow. *Noûs* 13(4). 455–476. doi:10.2307/2215339.
Lewis, David. 1980. A subjectivist's guide to objective chance. In Glenn Pearce William L. Harper, Robert Stalnaker (ed.), *Ifs: Conditionals, belief, decision, chance and time*, 267–297. Springer.
Lewis, David Kellogg. 1973. *Counterfactuals*. Cambridge, MA: Blackwell.
Loughrist, Tim. 2021. Defeaters and the generality problem. *Synthese* 199(5). 13845–13860. doi:10.1007/s11229-021-03400-4.
Lyons, Jack C. 2009. *Perception and basic beliefs: Zombies, modules and the problem of the external world*. New York, US: Oxford University Press.

Markovits, Julia. 2010. Acting for the right reasons. *Philosophical Review* 119(2). 201–242. doi:10.1215/00318108-2009-037.
Nagel, Jennifer. 2021. Losing knowledge by thinking about thinking. In Jessica Brown & Mona Simion (eds.), *Reasons, justification, and defeat*, 69–92. Oxford University Press.
Nozick, Robert. 1981. *Philosophical explanations*. Cambridge, Mass.: Harvard University Press.
Pettigrew, Richard. 2012. Accuracy, chance, and the principal principle. *Philosophical Review* 121(2). 241–275. doi:10.1215/00318108-1539098.
Pettigrew, Richard. 2016. *Accuracy and the laws of credence*. New York, NY.: Oxford University Press UK.
Pettigrew, Richard. 2021. What is justified credence? *Episteme* 18(1). 16–30. doi:10.1017/epi.2018.50.
Plantinga, Alvin. 2000. Defeaters and defeat. In Alvin Plantinga (ed.), *Warranted christian belief*, Oxford University Press USA.
Pollock, John & Joe Cruz. 1999. *Contemporary theories of knowledge, 2nd edition*. Rowman & Littlefield.
Ramsey, Frank. 1931 [1926]. Truth and probability. In R.B. Braithwaite (ed.), *The foundations of mathematics and other logical essays*, London: Kegan Paul, Trench, Trubner, & Co.
Schoenfield, Miriam. m.s. Higher order troubles for higher order defeat. Manuscript.
Smith, Martin. 2010. What else justification could be. *Noûs* 44(1). 10–31. doi:10.1111/j.1468-0068.2009.00729.x.
Sosa, Ernest. 1999. How must knowledge be modally related to what is known? *Philosophical Topics* 26(1-2). 373–384. doi:10.5840/philtopics1999261/229.
Stalnaker, Robert. 2006. On logics of knowledge and belief. *Philosophical Studies* 128(1). 169–199. doi:10.1007/s11098-005-4062-y.
Sutton, Jonathan. 2005. Stick to what you know. *Noûs* 39(3). 359–396. doi:10.1111/j.0029-4624.2005.00506.x.
Swain, Marshall. 1979. Justification and the basis of belief. In George Pappas (ed.), *Justification and knowledge: New studies in epistemology*, 25–50. D. Reidel.
Swinburne, Richard. 2001. *Epistemic justification*. New York: Oxford University Press.
Tang, Weng Hong. 2016. Reliability theories of justified credence. *Mind* 125(497). 63–94. doi:10.1093/mind/fzv199.
Wedgwood, Ralph. 2022. Doxastic rationality. In Paul Silva & Luis R. G. Oliveira (eds.), *Propositional and doxastic justification: New essays on their nature and significance*, 219–240. Routledge.
White, Roger. 2009. On treating oneself and others as thermometers. *Episteme* 6(3). 233–250. doi:10.3366/e1742360009000689.
Williamson, Timothy. 2000. *Knowledge and its limits*. Oxford University Press.
Williamson, Timothy. 2009a. Probability and danger. In *Amherst lecture in philosophy*, 1–35.
Williamson, Timothy. 2009b. Reply to goldman. In Duncan Pritchard & Patrick Greenough (eds.), *Williamson on knowledge*, 305–312. Oxford: Oxford University Press.
Williamson, Timothy. 2024. Modal epistemology. In Jacques-Henri Vollet Artūrs Logins (ed.), *Putting knowledge to work: New directions for knowledge-first epistemology*, Oxford: OUP.