Reasons for Reliabilism

Bob Beddor

October 6, 2019

Abstract

One leading approach to justification comes from the reliabilist tradition, which maintains that a belief is justified provided that it is reliably formed. Another comes from the 'Reasons First' tradition, which claims that a belief is justified provided that it is based on reasons that support it. These two approaches are typically developed in isolation from each other; this essay motivates and defends a synthesis. On the view proposed here, justification is understood in terms of an agent's reasons for belief, which are in turn analyzed along reliabilist lines: an agent's reasons for belief are the states that serve as inputs to their reliable processes. I show that this 'Reasons First Reliabilism' allows each tradition to profit from the other's explanatory resources. It enables reliabilists to explain epistemic defeat, and it enables Reasons Firsters to give a predictive and naturalistic epistemology. I go on to compare Reasons First Reliabilism with other hybrid versions of reliabilism that have been proposed in the literature.

1 Two Approaches to Justification

What determines whether a belief is justified?

One answer comes from the reliabilist tradition:

Reliabilist Answer Whether a belief is justified depends on whether it is reliably formed.[1]

Reliabilism holds considerable appeal. First, it explains the intuitive connection between justification and truth: a reliable process is, by definition, truth-conducive. Second, it explains how justification reduces to non-epistemic properties. Justification is explained in terms of reliability, which is explained in terms of truth and falsity. Reliabilism thus offers a way of locating epistemic properties within a naturalistic worldview.[2]

[1] The *locus classicus* of reliabilism about justification is Goldman (1979). For further development and defense, see Goldman (1986, 2012); Kornblith (2002); Lyons (2009).

[2] The reductive goals of reliabilism are clearly announced by Goldman, who writes: "I want a theory of justified belief to specify in non-epistemic terms when a belief is justified" (1979: 90). As Kim (1988) notes, one motivation for seeking a reductive account is that epistemic properties supervene on natural properties. A reductive account offers to explain this supervenience.

Despite its appeal, reliabilism faces significant challenges. Many of these are by now well-known, and have elicited replies from reliabilists, which in turn have elicited counter-replies.[3] In this paper, I will bypass this well-worn terrain to focus on a challenge that has received comparatively little attention—one that attacks the heart of reliabilism's reductive aspirations.

The challenge arises from the fact that a belief can be reliably formed even though the believer has good reason to think the belief is false, or that it was unreliably formed. When this happens, the belief is *defeated*. In order to account for such cases, reliabilists need to provide a theory of defeat. And in order for this account to be faithful to reliabilism's reductive ambitions, it had better not use any epistemic terms in the *analysans*.

Providing such an account is no easy task. The standard reliabilist approach is to say that a belief is defeated when there is an alternative reliable process that would have led the believer to abandon the belief, had it been used. But, as I'll argue here, this 'Alternative Reliable Process Account' faces serious problems. This raises the worry that reliabilists are unable to explain a central facet of justification.

Faced with this difficulty, it is tempting to look beyond the reliabilist tradition for help. Another prominent account of justification comes from the 'Reasons First' tradition. This tradition offers a competing story about the grounds of epistemic justification:

Reasons First Answer Whether a belief is justified depends on whether it is supported by adequate reasons.

Unlike reliabilism, the Reasons First framework provides a promising account of defeat. Reasons Firsters hold that a belief is *prima facie* justified if it is supported by the agent's *prima facie* reasons. Defeat occurs when this support is undermined by the acquisition of further reasons. This approach to defeat has been elaborated in great detail by John Pollock, who develops a rigorous system for computing the defeat statuses of an agent's beliefs on the basis of their *prima facie* reasons.[4]

Despite these advantages, the Reasons First framework faces difficulties of its own. As standardly developed, the Reasons First framework lacks the explanatory benefits that make reliabilism attractive. It leaves unexplained the intuitive connection between justification and truth, and it does not reduce epistemic properties to non-epistemic properties. After all, Reasons Firsters take as a primitive the notion of a *reason to believe*, which is clearly an epistemic notion.

Given this tradeoff, we might hope for a theory that combines the attractions of both traditions. This paper develops one such theory: Reasons First Reliabilism. The theory is Reasons First because it explains justification and defeat in terms of the notion of a *reason to believe*, which is taken to be the most fundamental normative notion. But it is also reliabilist, because it goes on to analyze this notion in terms of

[3]For an overview of the major challenges and replies, see Goldman and Beddor (2015).

[4]See e.g., Pollock (1987, 1992, 1994, 1995, 2001). For a useful overview of the development of Pollock's framework, see Prakken and Horty (2012).

reliability. Simplifying somewhat, I propose identifying an agent's reasons for belief with the states that serve as potential inputs to their reliable processes. I argue that the resulting package preserves the best of both frameworks. In particular, it provides an elegant treatment of defeat while remaining faithful to reliabilism's reductive project.

Of course, I am not the first to advocate an 'impure' or 'hybrid' version of reliabilism. Recently various authors have argued for a synthesis of reliabilism and evidentialism.[5] However, I show that extant evidentialist-reliabilist hybrids lack one of the chief advantages of the view advocated here: namely, its ability to explain defeat. Much like the 'pure' Reasons First framework, extant hybrids struggle to provide a theory that is both reductive and predictive. They also problematically single out a privileged class of beliefs—those entailed by the agent's evidence—as immune to defeat. Reasons First Reliabilism fares better on both these fronts.

2 The Classic Reliabilist Account of Defeat

2.1 Why Reliabilists Need an Account of Defeat

In order to introduce reliabilism's difficulties with defeat, it will be helpful to start with a simple version of reliabilism:

Simple Reliabilism An agent's belief is justified iff it is formed by a reliable belief-forming process.

Next, consider a stock example of defeat:

Seeing Red Lori is gazing at a wall, which appears red. Consequently, she comes to believe RED: *The wall is red*. Just then, a generally reliable acquaintance, Sal, mentions to Lori that there are hidden red lights angled towards the wall.[6]

According to Simple Reliabilism, Lori's belief in RED remains justified even after she receives Sal's testimony. After all, her belief is formed via vision, and we can stipulate that she has excellent eyesight. But, intuitively, Sal's testimony defeats Lori's justification for believing RED.[7]

[5]See Alston (1988); Henderson et al. (2007); Comesaña (2010, 2018); Goldman (2011); Tang (2016); Pettigrew (2018); Miller (forthcoming). Most of these syntheses have been motivated by considerations other than defeat, though Miller (forthcoming) is an important exception.

[6]For discussion of this sort of case, see a.o., Chisholm (1966); Pollock (1995); Lasonen-Aarnio (2010a).

[7]Perhaps, some may suggest, once Lori receives Sal's testimony, her belief in RED is no longer the result of vision alone. Rather, it's the result of a complex process: *using vision while disregarding testimony that vision is locally unreliable*. Given this way of typing Lori's belief-forming process, the process is arguably unreliable. However, I think there is reason to be skeptical of this 'typing maneuver.' First, in order to for this maneuver to work in full generality, we would need to take on board a substantive commitment: *In every case of defeat, there is some way w of typing the agent's belief-forming process on which it comes out unreliable*. And in order to ensure that this approach is not *ad hoc*, we'd need to go further: we'd

In view of such cases, most reliabilists conclude that Simple Reliabilism is at best an adequate account of *prima facie* justification. In order for a belief to be *ultima facie* justified (that is, justified *full-stop*) it is not enough for it to be reliably formed. It also needs to satisfy a 'No Defeaters' condition.[8]

However, introducing a 'No Defeaters' condition raises a difficult question. *Defeat* is clearly an epistemic notion. In order to fulfill their reductive ambitions, reliabilists need to explain this notion in non-epistemic terms. Can this be done?

2.2 The Alternative Reliable Process Account of Defeat

Reliabilists have not left this question unanswered. The standard reliabilist strategy is to explain defeat in terms of counterfactuals about what the agent would have believed, were they to have used some alternative reliable process.[9] More precisely:

Alternative Reliable Process Account (ARP) An agent A's belief B is defeated iff there is some alternative reliable (or conditionally reliable) process available to A which, if it had been used in addition to the process actually used, would have resulted in A's not holding B.

At first blush, ARP provides a promising account of defeat. It appears to explain defeat in entirely naturalistic terms. It also seems to deliver the right verdicts in many cases. Take **Seeing Red**: since Sal is stipulated to be generally reliable, *deferring to Sal's testimony* is a reliable process. And if Lori had used this process in addition to vision, she wouldn't have continued to believe RED.

These advantages nothwithstanding, ARP faces three serious challenges.

3 Difficulties for the Classic Reliabilist Account

3.1 Defeater Defeaters

An initial difficulty for ARP—noted in passing by Lyons (2009): 124—is that it yields the wrong results when defeaters are themselves defeated.An example:

Two Testimony Seeing Red As before, Lori believes the wall is red, based on its appearance. And as before, Sal comes along and mentions that the wall is illuminated by red lights. But now another reliable acquaintance, Anne, comes along and provides compelling—though ultimately misleading—testimony that Sal is a compulsive liar.

need to give some independent motivation for typing the belief-forming process using w. For further development of this concern, see Beddor (2015a): 147-148, where I argue that this typing maneuver stands in tension with some of the most promising solutions to the generality problem. For related criticisms of the typing maneuver, see Lasonen-Aarnio (2010a): 4-7; Baker-Hytch and Benton (2015): 45-47.

[8] See e.g., Goldman (1979); Lyons (2009).

[9] This proposal dates back to Goldman (1979), and has been recently defended by Lyons (2009, 2016). Close cousins are defended in Grundmann (2009) and Bedke (2010).

According to ARP, Lori's belief in RED is defeated, even after receiving the evidence of Sal's mendacity. After all, Anne's testimony is misleading, and so *trusting Sal's testimony* continues to be a reliable process. And this process remains available to Lori. (Lori could, after all, simply disregard Anne's testimony.) But this is the wrong verdict. Intuitively, Anne's testimony defeats the defeater provided by Sal's testimony. In doing so, it reinstates Lori's justification for believing RED.

3.2 Hidden Circularity

A second concern for ARP was raised in Fumerton (1988), but has not received much attention in the subsequent literature. The worry is that ARP, when properly unpacked, smuggles the notion of *ultima facie* justification into the analysis of defeat. As a result, the reliabilist account of justification fails to be reductive; worse still, it is circular. Fleshing out this worry requires some stage-setting.

Many reliabilists opt for a theory that distinguishes between inferential and non-inferential beliefs. To motivate this complication, consider an inferential process such as deducing the consequences of what one already believes. This process is not reliable or unreliable *simpliciter*; it is only conditionally reliable.

In order to handle beliefs formed through conditionally reliable processes, Goldman (1979) officially formulates reliabilism as a recursive theory:

Recursive Reliablism

Base Clause If (i) A's belief *B* results from a belief-independent process that is unconditionally reliable, and (ii) *B* is undefeated, then *B* is *ultima facie* justified.

Recursive Clause If (i) A's belief *B* results from a belief-dependent process that is conditionally reliable, (ii) the inputs to this process were *ultima facie* justified, and (iii) *B* is undefeated, then *B* is *ultima facie* justified.[10]

The recursive clause requires that the conditionally reliable processes operate on *ultima facie* justified beliefs. Despite this, the theory is still reductive. After all, we can use the theory to explain what it is for these input beliefs to be *ultima facie* justified. Either these inputs are themselves inferential or they are not. If they are, we appeal once again to the recursive clause; if not, we appeal to the base clause. Either way, we eventually arrive at some foundational beliefs whose justificatory status can be explained using the base clause. Now, the only epistemic notion that appears in the base clause is the notion of being *undefeated*. Assuming ARP provides a reductive account of this notion, Recursive Reliabilism is reductive.

Where, then, lies the problem? Fumerton observes that, according to ARP, conditionally reliable processes sometimes function as defeaters: a belief is defeated if there is some conditionally reliable process available to the agent which, had it been used, would have resulted in the agent no longer holding the belief. But a conditionally

[10]See also Lyons (2013), who argues that distinguishing between inferential from non-inferential beliefs can help reliabilists handle the new evil demon problem (Cohen 1984).

reliable process cannot lead someone to abandon a belief all on its own; it can only do so if it is fed certain inputs. This raises the question: what epistemic status do these input beliefs need to have? Presumably, Fumerton suggests, just as Recursive Reliabilism required that the inputs be *ultima facie* justified, so too should ARP. And so ARP really amounts to the following:

ARP Unpacked A's belief B is defeated iff either:

1. There is some reliable belief-independent process that A could have used, which would have resulted in A not holding B, or
2. There is some conditionally reliable belief-dependent process that A could have used to process *ultima facie* justified inputs, which would have resulted in A not holding B.

But if this is the proper way of understanding ARP, then reliabilism's reductive project is in trouble. After all, Recursive Reliabilism qualified as reductive because the base clause purported to tell us what it takes for a foundational belief to be *ultima facie* justified without using any epistemic terms in the *analysans*. But if we use ARP Unpacked to explain what it is for a belief to be undefeated, the base clause will itself rely on the notion of *ultima facie* justification.[11]

3.3 Alternative Processes that One Should Not Use

According to ARP, an agent's belief is defeated whenever they have an available reliable process that meets a certain counterfactual condition—namely, that if they were to use it, they would abandon their belief. But it seems that an agent can have an available reliable process that meets this condition without having any good reason to use it. When this happens, the mere availability of the process does not seem to defeat the belief.

Here's a case I offered in an earlier paper (Beddor 2015a: 149-150) that illustrates this point:

Thinking About Unger Harry sees a tree in front of him; he consequently believes TREE: *There is a tree in front of me*. Now, Harry happens to be very good at

[11] Fumerton's objection hinges on the assumption that the inputs to the conditionally reliable process need to be *ultima facie* justified. Could proponents of ARP simply reject this assumption? This is certainly a coherent option. Indeed, some authors have defended the idea that unjustified beliefs can function as defeaters (e.g. Lackey 1999; Bergmann 2006), and Goldman himself flirts with this view in places (1986: 62, 111). However, this response arguably yields counterintuitive results. Imagine a variant of **Seeing Red** where Lori unjustifiably believes that the wall is illuminated by red lights, but continues to believe RED anyway. Intuitively, her overall *set* of beliefs is epistemically defective. But does this defect render her belief in RED unjustified? To answer this, it will help to consider a slightly different question: should Lori abandon her belief in RED? I think not. After all, she has no *good* reason to abandon it. Instead, it's her belief about the lighting that should get the boot. But if Lori should retain her belief in RED, it becomes hard to maintain that this belief is defeated. (If it were defeated, she should presumably abandon it.) This line of reasoning supports Fumerton's claim that only justified beliefs can serve as defeaters.

> forming beliefs about what Peter Unger's skeptical 1975 time-slice would advise him to believe in any situation. Call this process his 'Unger Predictor': in any situation, Harry's Unger Predictor spits out an accurate belief about what doxastic attitudes Unger's 1975 time-slice would advise an agent to adopt in that situation. Moreover, Harry has a high opinion of Unger's 1975 time-slice. Were he to realize that Unger would advise him to suspend judgment on some proposition, this would lead him to suspend judgment on that claim. So if Harry had used his Unger Predictor, he would have come to believe SUSPEND: *Unger would advise me (Harry) to suspend judgment regarding* TREE. This would, in turn, have caused Harry to suspend judgment regarding TREE.

According to ARP, Harry's belief in TREE is defeated. After all, there is a reliable process available to him (his Unger Predictor) that would have resulted in him no longer believing TREE, had it been used. But this seems wrong. Harry is not, as a matter of fact, using his Unger Predictor; perhaps he hasn't used it in many years. As it stands, he has an excellent reason to believe TREE (the testimony of his senses). The mere availability of his Unger Predictor does not seem undermine this reason.

Defenders of ARP may suggest there's an easy fix. A natural reaction to **Thinking About Unger** is that the example exploits a *subject matter mismatch*. Harry's belief is about trees. The Unger Predictor does not produce doxastic attitudes about trees, but only about Unger's advice. This suggests a simple patch to ARP: simply require that in order for an alternative process to defeat an agent's belief in p, that process must produce beliefs about p-related matters.

However, it is doubtful whether this simple fix suffices. Suppose Harry had used his Unger Predictor. Then some process would have led him to suspend judgment on TREE. It's just that this would have been a two-stage process. This first stage would have been his Unger Predictor; the second stage would have been a process that implements Unger's predicted advice. Call this two-stage process his 'Unger Emulator'. Now, this Unger Emulator produces doxastic attitudes towards all sorts of subjects, including the presence of trees. Assuming that this process is reliable, then ARP—even once amended—still delivers the wrong result.

Of course, some might question whether this Unger Emulator process is reliable. Sure, it avoids all errors, but only at the cost of avoiding all truths! Perhaps this is too high a cost; perhaps the right conception of reliability will classify this process as unreliable.[12] But even if we concede the point, we can simply tweak the example (Beddor 2015a: 153-154). Meet *Shmunger*, whose skepticism is much more modest. Shmunger has lots of true beliefs about all sorts of subjects; she is only a skeptic when it comes to trees. We can then run the case using Shmunger instead of Unger. Simply stipulate that Harry has an extremely reliable Shmunger Predictor, which is part of a Schmunger Emulator: were Harry to reflect on what Shmunger would advise, he

[12]The formal epistemology literature offers a natural place to look for a measure of reliability along these lines. See the discussions of credal scoring rules in Joyce (1998); Moss (2011); Pettigrew (2016), among many others.

would come to believe *Schmunger would advise me to suspend judgment on whether there is a tree in front of me*, which would in turn cause him to suspend judgment on TREE. Note that here proponents of ARP cannot plead that Harry's Schmunger Emulator is unreliable. After all, it systematically produces true beliefs on a wide array of subjects; it only leads to suspension of judgment on arboreal matters.[13]

3.4 Looking Forward

Taken together, these problems show that ARP will not do. But it would be premature for reliabilists to admit defeat. The rest of this paper develops a more promising approach, which draws on the resources of the Reasons First tradition.

Here is the plan for what follows. I start (§4) by outlining the most well-developed version of the Reasons First framework to date, which is due to John Pollock. While Pollock offers a promising formal framework for understanding the structure of justification and defeat, I argue that it should not supplant reliabilism (§5). Rather, we should seek a theory that combines the structural features of Pollock's framework with the core reliabilist strategy for reducing the epistemic to the non-epistemic. §6 develops such a theory; §7 advertises its advantages; and §8 compares it to other hybrid views.

4 Pollock's Reasons First Framework

According to the Reasons First tradition, justification is intimately connected with *reasons*. This idea has considerable intuitive appeal. 'A justified belief is supported by reasons' has the ring of a platitude; 'A belief can be justified, even though all the reasons count against it' has the ring of a contradiction.

Moreover, the Reasons First approach offers a promising treatment of defeat. The basic idea is simple: for a belief to be *prima facie* justified is for it to be based on *prima facie* reasons that support it. Defeat occurs when the agent acquires reasons that

[13] In my earlier paper, I also offered a counterexample to the necessity of ARP for defeat. In the proposed counterexample, Clarence reliably forms a belief that p; a reliable interlocutor later tells him $\neg p$. But Clarence irrationally disbelieves everything this interlocutor says. Moreover, no amount of reflection or counseling would ever uproot this deep-seated mistrust. Intuitively, Clarence's belief is defeated. But ARP seem unable to deliver this result: there is no process available to Clarence which would lead him to trust his interlocutor; hence there is no process available to him which, if used, would lead him to abandon his belief in p. (Cf. Baker-Hytch and Benton 2015: 53.)

However, in this case—unlike **Thinking About Unger**—it now seems to me that a modification of ARP will suffice. All that's needed is to reformulate ARP in terms of dispositions rather than counterfactuals: A's belief that p is defeated iff there is some alternative reliable process available to A which, when fed A's current states as input, is *disposed* to lead A to cease believing p. Defenders of ARP could then propose that Clarence has a general *testimony-believer* process available to him. This is a process that, for any testifier (or, perhaps, any testifier that he has no good reason to distrust), produces a relatively high credence in their testimony. This process is generally reliable. Moreover, it is *disposed* to lead Clarence to cease believing p, when fed the experience of receiving his interlocutor's testimony as input. It's just that this disposition is masked by Clarence's mistrust of his interlocutor.

either count against the belief itself, or against the support provided by the reasons on which it is based.

This way of understanding defeat has been given a systematic development by John Pollock in a series of papers spanning over thirty years.[14] Pollock's framework has proven influential beyond epistemology, laying the groundwork for much research in computer science and AI—an influence that attests to the explanatory fruitfulness of its core ideas. In this section, I offer a streamlined overview of Pollock's framework. (Readers uninterested in the formal nuts-and-bolts should feel free to skim.)

Pollock's key piece of formal machinery is an *inference graph*: a labeled directed graph providing an abstract representation of all the reasons that bear on the justificatory status of an agent's beliefs. The agent's reasons—as well as the conclusions they support—are represented by nodes. Support and defeat relations are represented by directed edges. An agent need not actually perform all of the inferences encoded in their inference graph. Rather, the inference graph represents all the inferences that are rationally available to them.

For illustration, Fig. 1 provides an inference graph for **Seeing Red**. Dashed arrows represent support relations; solid arrows represent defeat relations. Here Lori's visual experience provides a *prima facie* reason in support of the node, RED. And Sal's testimony provides a *prima facie* reason in support of the node, *The wall is illuminated by red lights* (RL). This in turn supports the node, *Lori's visual experience doesn't reliably indicate the truth of* RED (U), which defeats RED.

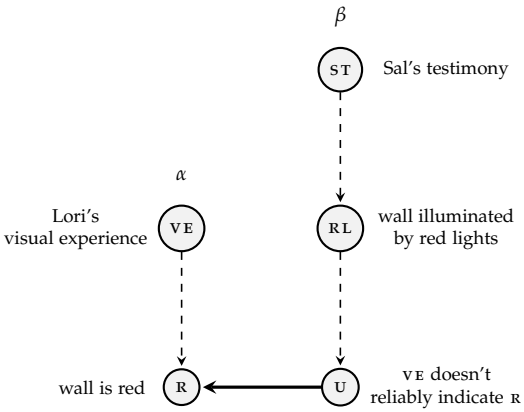

Figure 1: Seeing Red

Inferential support often involves multiple steps. Following Pollock, we can represent multi-step arguments with *inference branches*. An inference branch is an ordered sequence of nodes, each of which is the immediate ancestor of the next. For example, in Fig. 1 branch α is the directed path from Lori's visual experience to RED. Branch β is the directed path originating in the experience of receiving Sal's testimony, leading through RL, and terminating in U.

[14] See the references in fn. 4.

Using these resources, we can now flesh out the animating idea behind the Reasons First framework as follows:

Justified Belief as Undefeated Reasoning An agent's belief is *ultima facie* justified iff it is the result of an ultimately undefeated inference branch.

This formulation invites two questions. First, what does it mean for a belief to be the result of an inference branch? For starters, the belief must be supported by the sequence of reasons represented by the inference branch. But this is not sufficient: the agent must actually have gone through this reasoning, and hold the belief on this basis.

Second, what does it mean for an inference branch to be ultimately undefeated? This question is harder, and Pollock's answer proceeds in stages.

The first stage is to give an account of what it means for one reason to be defeated by another. According to Pollock, there are two species of defeat: *rebutting defeaters* and *undercutting defeaters*. A rebutting defeater for a node n is a *prima facie* reason to think that n is false. By contrast, an undercutting defeater for n targets the inferential connection between n and the reasons that support it. **Seeing Red** is like this: Sal's testimony that the wall is illuminated by red lights is not itself a reason to believe that the building is not red, but it is a reason for thinking that the building's appearance does not give good grounds for thinking that the building is red. While there are different ways of fleshing out the notion of an undercutting defeater, for our purposes we can define an undercutting defeater for n as a *prima facie* reason for thinking that the considerations that support n do not reliably indicate its truth in the agent's present circumstances.[15]

If we follow Pollock in assuming that these are the only two species of defeat,[16] we can venture the following disjunctive definition of when one inference branch defeats another (cf. Pollock 1992):

Branch Defeat An inference branch ψ defeats an inference branch χ iff a node of ψ defeats a node of χ, where a node n defeats a node n' iff n either rebuts or undercuts n',

—i.e., either n is a *prima facie* reason to believe $\neg n'$, or n is a *prima facie* reason to believe that the immediate ancestors of n' do not reliable indicate the truth of n' in the agent's present circumstances.

[15]The 'in the present circumstances' qualification is important: after all, Sal's testimony does not provide a reason for thinking that reddish wall appearances do not in general indicate the presence of a red wall.

[16]This is a controversial assumption; some have suggested that cases of higher order evidence constitute a distinct species of defeat (e.g., Christensen 2010). My own inclination is to think that when higher order evidence functions as a genuine defeater, it does so by indicating that the agent's actual (or believed) grounds for their belief do not reliably indicate the truth of this belief. If this is right, then defeat by higher order evidence is really just a type of undercutting defeat. But this is a debate that will need to be deferred to another occasion.

This gives us a definition of when one inference branch defeats another. But what we really want is a definition of when an inference branch is ultimately undefeated. The simplest option would be to say that an inference branch is ultimately undefeated just in case there is no inference branch that defeats it. But this delivers the wrong results in cases of defeater defeat.

Recall **Two Testimony Seeing Red**, in which Anne testifies that Sal is a compulsive liar. As in **Seeing Red**, Sal's testimony supports the node, *The wall is illuminated by red lights* (RL), which supports the node, *Lori's visual experience doesn't reliably indicate the truth of* RED, which undercuts RED. Hence the simple account predicts that branch α is ultimately defeated, and hence that Lori's belief in RED is unjustified. But this is wrong. After all, Anne's testimony provides a *prima facie* reason to believe that Sal is a liar, which provides a *prima facie* reason to believe that Sal's testimony does not reliably indicate RL, which undercuts RL. As noted in §3.1, it thereby reinstates Lori's justification for believing RED. (See Fig. 2.)

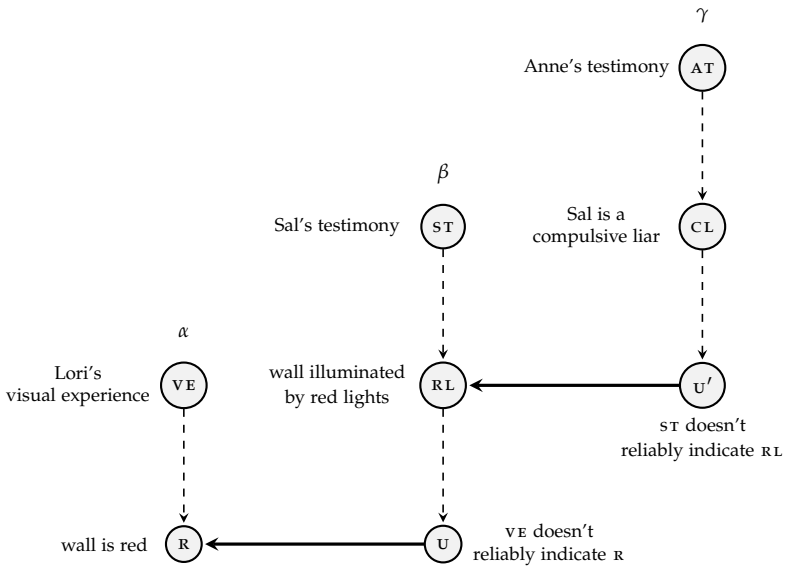

Figure 2: Two Testimony Seeing Red

For this reason, Pollock opts for a somewhat more complicated account of what it takes for an inference branch to be ultimately undefeated. Here we will follow the treatment in Pollock (1987), who introduces a technical notion of *being in at a level*, defined recursively as follows:

In At A Level

1. All inference branches are in at level 0.
2. An inference branch ψ is in at a level $n+1$ iff ψ is not defeated by any inference branch that is in at level n; otherwise, ψ is out at level $n+1$.

Next, we use the notion of being in at a level to characterize what it is for an inference branch to be ultimately undefeated, as follows:

Undefeated Inference Branch An inference branch ψ is ultimately undefeated iff there is an m such that for every $n \geq m$, ψ is in at level n.

To get a feel for this proposal, let's walk through how it applies to **Two Testimony Seeing Red**. While all three inference branches depicted in Fig. 2 are in at level 0, only γ is in at every level, since only it lacks a defeater. Since γ defeats β, the latter is out at every level ≥ 1. And so while α is out at level 1 (since it is defeated by a branch that is in at level 0), it is back in at level 2, and and remains in at every level thereafter. (See Table 1.) So α is ultimately undefeated. Hence Lori's belief in RED qualifies as *ultima facie* justified, as desired.

Level	α	β	γ
0	in	in	in
1	out	out	in
2	in	out	in
$n > 2$	in	out	in

Table 1: Computing defeat.

5 Reason to Want More

Pollock's framework offers a promising way of handling some of the cases that created trouble for ARP—in particular, cases of defeater defeat. Why not just jettison reliabilism in favor of the Reasons First program?

There are a number of reasons why one might be dissatisfied with Pollock's theory as it stands. Some of these are issues of detail which Pollock himself sought to address in his later work. For example, Pollock ultimately opts for a slightly more complicated characterization of what it takes for an inference branch to be ultimately undefeated—complications that are mainly driven by a desire to handle self-defeating inferences. He also complicates the account by adding additional structure in order to represent the *strengths* of reasons for belief.[17] Other concerns focus on specific applications of Pollock's framework to various philosophical puzzles, such as the lottery paradox.[18] I will set these worries aside, since they are not directly relevant to our purposes. Rather, I want to raise two more fundamental concerns.

One concern is that Pollock's account does not accommodate the intuitions and impulses that motivate reliabilism. First, it does not capture the intuition that there is an important connection between justification and truth. Suppose a belief is based on undefeated reasons that support it. Why should we expect the belief to be connected with the truth in any interesting way? Pollock's framework provides no answer. Second, and more critically, Pollock's framework does not satisfy the reductive impulse behind reliabilism. Pollock explains *ultima facie* justification and defeat in terms of the notion of a *prima facie* reason for believing. But this is surely an

[17] See Pollock (1994, 1995) for his refined treatment of self-defeating inference branches. See Pollock (2001) for discussion of strengths of reasons.

[18] See Lasonen-Aarnio (2010b).

epistemic notion. Those who want an account of justification in entirely non-epistemic terms will be left empty-handed.

Of course, some Reasons Firsters might retort that we should never have hoped for a reductive analysis in the first place. Reductive analyses of other epistemic phenomena have a fairly spotty track record—the analysis of knowledge being a case in point. Why think that the prospects for a reductive analysis of justification will be any better?

But this brings me to my second concern, which that even if we renounce the dream of a fully reductive account, it is natural to expect *some* account—reductive or not—of *prima facie* reasons. After all, without some account, the Reasons First framework will not offer a predictive theory at all. Unless we have some independent grip on *prima facie* reasons for belief, we will not be able to apply Pollock's framework to particular cases in order to make predictions about whether a belief is justified or defeated.

Pollock was sensitive to this concern, and in various places he offers remarks intended to fill this lacuna. For example, he states that perceptual appearances provide *prima facie* reasons to believe; so does memory; so does statistical syllogism; so does deduction and induction (Pollock 1987: 486-490). Arguably, these remarks go some distance towards giving us an independent grip on the notion of *prima facie* reasons.

However, I think there are still grounds for dissatisfaction. As it stands, Pollock's remarks look more like a *list* of various sources of *prima facie* reasons than a genuine theory thereof. A genuine theory should be explanatorily satisfying: it should tell us what perception, memory, and induction have in common, in virtue of which they furnish *prima facie* reasons for belief, whereas, say, wishful thinking and counterinduction do not. By comparison, reliabilism offers a much more unified and theoretically satisfying account of the ultimate grounds of justification. According to reliabilism, all the ultimate sources of justification have one property in common: their reliability.

For these reasons, we should not simply replace reliabilism with Pollock's framework. Instead, we should seek a synthesis that preserves the chief virtues of both approaches. It is to this task which I now turn.

6 Reliabilism about Reasons

The basic idea behind my proposal is simple: reliabilists should identify reasons for belief with the inputs to reliable or conditionally reliable belief-forming processes.

A more careful statement proceeds recursively. The base clause gives an account of an agent's foundational reasons:

Reliablist Reasons (Base Clause) If s is a non-doxastic state of an agent A, and there is a reliable process available to A which, when given s as input, is disposed to produce a belief in p, then s is a *prima facie* reason for A to believe p.

What sort of states play this role? The clearest candidates are perceptual experiences. For example, Lori's visual experience of a red-looking wall is a *prima facie* reason to believe RED. Why? Because she has a reliable process that takes the contents of her perceptual experiences as input and produces a belief in those contents as output. And this process is disposed to produce a belief in RED, when applied to her visual experience of red-looking wall.

Do states other than perceptual experiences also fit the bill? Perhaps—depending on one's views, rational intuitions and seemings may also serve this foundational role. Perhaps even non-experiential states could play the part. For our purposes, there is no need to take a stand on this issue.

Next, we add a recursive clause, which gives an account of an agent's derivative reasons:

Reliablist Reasons (Recursive Clause) If A has a *prima facie* reason to believe p, and there is some conditionally reliable process available to A which, given a belief in p as input, is disposed to produce a belief in q, then p is a *prima facie* reason for A to believe q.

To illustrate, suppose Lori is capable of inferring, *At least one thing in my vicinity is red* from RED. Since this inferential process is conditionally reliable, the recursive clause tells us that RED provides a *prima facie* reason for Lori to hold this inferential belief.

Round everything out with the customary closure clause (nothing else is a *prima facie* reason for A to believe p), and you have a complete reliabilist theory of reasons. Of course, this theory could—and perhaps should—be complicated in various ways. For example, in §5 I mentioned that Pollock's final inference graphs include representations of the strengths of an agent's reasons. A natural way of modeling this in a reliabilist framework is to take the strength of an agent's reasons to correspond to the degrees of reliability (and conditional reliability) of the relevant processes. Thus if there is an extremely reliable process that is disposed to produce a belief in p, given state s_1 as input, but there is only a somewhat reliable process that is disposed to produce a belief in p, given state s_2 as input, then s_1 is a stronger reason to believe p than s_2.

Equipped with Reliabilist Reasons, the reliabilist can embrace Pollock's framework. In particular, she can accept Justified Belief as Undefeated Reasoning: for a belief to be *ultima facie* justified is for it to be the result of an ultimately undefeated inference branch. But what *this* means is now given a reliabilist interpretation. Let's take this step by step.

Recall that in order for a belief B to be the result of an inference branch ψ, B must be supported by the reasoning in ψ, and must be held as a causal consequence of this reasoning. On the reliabilist interpretation advocated here, an inference branch is just a chain of reliable or conditionally reliable processes. If B is a foundational belief, then this chain consists of a single reliable process applied to some non-doxastic state. If B is an inferential belief, then the first link in this chain is a conditionally

reliable process applied to some further belief B', which is itself the result of a chain of reliable or conditionally reliable processes.

Next, what does it mean for an inference branch to be ultimately undefeated, on a reliabilist picture? Here too, reliabilists can accept Pollock's account. Following Pollock, reliabilists can explain this in terms of the notion of *being in at a level*, which is in turn characterized in terms of when one inference branch defeats another. And they can go on to analyze what it is for one inference branch to defeat another in terms of *prima facie* reasons. But what it takes for there to be such a reason is now explained in reliabilist terms.

Call the synthesis of Reliabilist Reasons with Pollock's framework, 'Reasons First Reliabilism.' I now argue that this synthesis preserves the primary advantages of both traditions.

7 Problems Solved

7.1 A More Satisfactory Reasons First Approach

By giving the Reasons First approach a reliabilist twist, we avoid the concerns for Pollock's framework voiced in §5.

Unlike the 'pure' version of Pollock's framework (§4), Reasons First Reliabilism preserves the main selling points of reliabilism. First, it captures the intuition that there is an important connection between justification and truth. On the approach defended here, for an agent to have a reason to believe p is for them to have a reliable (hence truth-conducive) process that is disposed to produce a belief in p.

Second, Reasons First Reliabilism remains faithful to reliabilism's reductive ambitions. Rather than resting content with a non-reductive analysis of justification in terms of reasons to believe, Reasons First Reliabilism shows how this notion can itself be reduced to non-epistemic notions. It thus fulfills the reliabilist goal of providing a fully naturalistic account of justification.

Moreover, the account of reasons that emerges is explanatorily satisfying. It does not just give a disjunctive list of all of the potential sources of *prima facie* reasons (e.g., you have a *prima facie* reason to believe p if it perceptually appears to you that p, or you seem to remember that p, or...). Instead, it tells us what all of these sources have in common. According to Reliabilist Reasons, what all reasons have in common is that they serve as the inputs to reliable (or conditionally reliable) processes.

7.2 A More Satisfactory Treatment of Defeat

I'll now argue that Reasons First Reliabilism also preserves the main advantage of Pollock's framework: specifically, its superior treatment of defeat. To make this point, let us revisit the difficulties facing ARP and see how Reasons First Reliabilism avoids them.

7.2.1 Defeater Defeat

The first difficulty for ARP was that it has trouble with defeater defeat. As we have seen, Pollock's definition of an undefeated inference branch is tailor-made to handle such cases. Since Reasons First Reliabilism makes use of Pollock's definition of an undefeated inference branch, it can enjoy the fruits of Pollock's labors.

To illustrate, recall **Two Testimony Seeing Red**. Reasons First Reliabilists can accept the inference graph we sketched for this case (Fig. 2). And they can say all the things that we said earlier about this inference graph. In particular, they can say that the inference branch responsible for Lori's belief in RED is ultimately undefeated, since the only inference branch that defeats it is out at every level ≥ 1.

However, Reasons First Reliabilists do not stop there. They supplement this formal representation of Lori's reasons with a reliabilist account of where the nodes come from, and why the various support and defeat links hold. According to Reliabilist Reasons, the reason why the experience of receiving Anne's testimony is a *prima facie* reason to believe that Sal is a compulsive liar is that there is a reliable process (*believing reliable interlocutors*) available to Lori that, when fed this experience as input, is disposed to produce a belief that Sal is a compulsive liar. Similar remarks apply, *mutatis mutandis*, to the other nodes depicted in the graph.

7.2.2 Circularity Worries

The second difficulty was that ARP turns out to be circular. To recap: the worry was that the proper way of unpacking ARP will rely on the notion of *ultima facie* justification. But ARP is used to articulate the conditions under which a belief is undefeated—a concept that occurs in the base clause of Recursive Reliabilism.

Reasons First Reliabilism avoids this worry. The 'Reasons First' part of the framework gives us a way of defining *ultima facie* justification in terms of the notion of a *prima facie* reason for believing. And the reliabilist part of the framework gives us a recursive definition of this primitive in terms of the inputs to various belief-forming processes. Crucially, the base clause of this definition (Reliabilist Reasons) does not itself rely on the notion of defeat, or any other epistemic notion for that matter.[19]

[19] At the same time, Reasons First Reliabilism respects Fumerton's claim that a conditionally reliable process can only serve as a defeater if it is applied to *ultima facie* justified inputs. *Proof:* Suppose A's inference graph contains some node n, and suppose that A unjustifiably believes some proposition q that, when fed into a conditionally reliable process, is disposed to produce a belief in d, where d is either of the form, $\neg n$ or *n's immediate ancestors do not reliably indicate n*. Since A's belief that q is unjustified, it follows (from Justified Belief as Undefeated Reasoning) that either (i) A's belief in q is not the result of any inference branch, or (ii) it is the result of some inference branch, but one of the branch's nodes is ultimately defeated. If (i), then A doesn't even have a *prima facie* reason to believe q, and so (by Reliabilist Reasons) q is not a *prima facie* reason to believe d. If (ii), then q is a *prima facie* reason to believe d, but q is ultimately defeated. Either way, n is not ultimately defeated.

7.2.3 Alternative Processes that One Should Not Use

The final difficulty for ARP came from cases where an agent has an alternative reliable process that they have no good reason to use. In **Thinking About Unger**, ARP predicts that Harry's belief in TREE is defeated merely in virtue of the fact that he has an available reliable process (his Unger Predictor) which, were he to use it, would lead him to suspend judgment on whether there's a tree in front of him.

Perhaps, some may suggest, the problem re-emerges if we amend the case. Meet Elijah, the eliminativist. Elijah thinks trees do not exist, and he wants others to share this belief. Suppose that Harry has a highly reliable Elijah Predictor, which is part of an Elijah Emulator: if he were to predict that Elijah would advise him to believe p in his current situation, this would in turn lead him to believe p. And so if Harry were to apply his current experiential states to his Elijah Predictor, he would be led to believe $\neg$TREE. Does Reasons First Reliabilism predict that this gives Harry a rebutting defeater for his belief in TREE?

An initial point: given this way of describing the case, it is doubtful whether Harry's Elijah Emulator is reliable. After all, it systematically misleads him about the presence of trees, which hardly bodes well for its reliability! However, the objector might try to circumvent this point by tempering Elijah's eliminativism. Just stipulate that, much like Shmunger before him, Elijah has has entirely correct beliefs about all sorts of topics—astronomy, geography, physics, whatever. It is only when it comes to trees that Elijah is an eliminativist. And so Harry's Elijah Emulator is overall reliable, even though it is unreliable on arboreal matters.

Suppose we grant all of this. Then Reasons First Reliabilism does indeed predict that Harry has a rebutting defeater for his belief in TREE. However, this need not worry us, provided that this defeater is itself defeated. Consider: why, exactly, would Harry be unjustified in using his Elijah Emulator in his current situation? Presumably, because he has good reason to think that his visual experience of trees reliably correlates with the presence of trees. Where does this reason come from? Presumably from his past experiences, which support the generalization that having a visual experience representing x is a reliable indicator of the presence of x. Plausibly, there is a reliable process available to him that, given these past experiences as input, is disposed to produce a belief that his current tree-like experiences reliably indicate TREE rather than $\neg$TREE. If this is correct, then these past experiences constitute an undercutting defeater for his rebutting defeater for believing TREE. (See Fig. 3.) And so his belief in TREE counts as *ultima facie* justified, as desired.

Could proponents of ARP co-opt this response? No. After all, the key move here is to diagnose this variant scenario as a case of defeater defeat. But as we've seen, ARP lacks an adequate story about how defeater defeat works.

7.3 Capturing the Role of Reasons in Justification

The primary payoff of recasting reliabilism as a Reasons First theory is that it provides a satisfactory treatment of defeat. However, a further benefit is also worth noting.

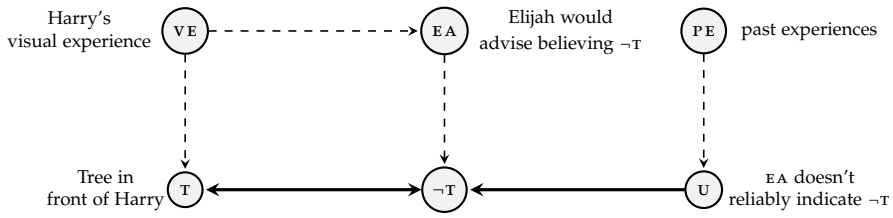

Figure 3: Thinking about Eliminativism

Historically, reliabilism has had little to say about reasons for belief. While it has not *denied* their existence, it has maintained a conspicuous silence about their nature.But clearly there are such reasons, and any complete epistemology should have something to say about them. Moreover, the notion of *reasons for belief* seems to be closely connected to the notion of *justification*. Intuitively, whether a belief is supported by reasons makes a difference to its justificatory status—indeed, this is precisely the intuition that provided the impetus for the Reasons First approach. It would be nice if reliabilists had some way of accommodating this thought.

Reasons First Reliabilism offers a natural strategy for doing so. It brings the notion of *reasons for belief* into the reliabilist fold. And it captures the intuition that the justificatory status of a belief depends on the agent's reasons.

8 Comparison with Evidentialist Hybrids

8.1 The Two Component View

I've advocated integrating reliabilism with a Reasons First framework—a framework that has traditionally been viewed as a rival to reliabilism. In doing so, I may appear to be joining my voice to a rising chorus. In recent years a wave of authors have suggested integrating reliabilism with evidentialism—a view that has also long been viewed as a competitor to reliabilism.[20] How does Reasons First Reliabilism differ from more familiar evidentialist-reliabilist hybrids?

To answer this, it will be helpful to look in some detail at how evidentialist-reliabilists handle defeat. On a standard evidentialist view, justification and defeat are explained in terms of *evidential support*. A belief is *prima facie* justified when it is supported by some initial body of evidence e_1. Defeat occurs when the agent acquires further evidence e_2 which, when combined with e_1, no longer supports the belief.

There are a number of ways of trying to integrate this approach with reliabilism. A particularly straightforward strategy is suggested by Goldman (2011), who proposes that *ultima facie* justification involves two components: a reliable process condition and an evidential support condition. That is:

Two Component View A's belief that p is *ultima facie* justified iff both:

[20] See the references in fn. 5.

Reliable Process Condition A's belief that p is the result of a reliable belief-forming process,
Evidential Support Condition A's total evidence supports believing p.[21]

It is easy to see in broad brushstrokes how the Evidential Support Condition helps with defeat. Take **Seeing Red**. When Lori first has the visual experience of a red-appearing wall, her total evidence supports believing RED, hence her belief is *ultima facie* justified. But once she receives Sal's testimony, her total body of evidence expands. This more inclusive body of evidence no longer supports believing RED.

As Miller (forthcoming) notes, a view along these lines also avoids many of the problems facing ARP. Take defeater defeat: when Lori acquires Anne's testimony in **Two Testimony Seeing Red**, her total body of evidence changes once again, and RED regains its former level of support. Or take **Thinking About Unger**: arguably, Harry's total evidence supports believing RED, despite the availability of his Unger Predictor.

Given these virtues, is there any reason to prefer Reasons First Reliabilism to the Two Component View? While a full adjudication of this issue will need to be left to another occasion, I want to briefly raise two reasons for thinking that the answer is 'yes.'

8.2 First Advantage: Reductive and Predictive

As it stands, the Two Component View is not reductive. After all, the Evidential Support Condition packages together two unreduced epistemic notions:

(i) The notion of an agent's total *evidence*,[22]
(ii) the notion of a body of evidence *supporting* a belief.

Perhaps, some might suggest, this just shows that we need supplement the Two Component View with a reductive analysis of these notions. To do so, proponents of the Two Component View could try taking a page from the Reasons First Reliabilist. According to Reliabilist Reasons, *prima facie* reasons are the inputs to reliable and conditionally reliable processes. Why not say the same about evidence? An agent's total evidence, on this view, is the total set of states of the agent that can serve as potential inputs to reliable processes of the agent.[23]

This would give us a reductive analysis of (i). What about (ii)? According to one common approach, evidential support should be understood in probabilistic terms: a

[21] See Tang (2016); Comesaña (2018); and Miller (forthcoming) for other hybrid views that impose an evidential support condition.

[22] Note that the challenge here is not just to give an account of *evidence* in non-epistemic terms. It's to give an account of what it is for an agent to *possess* evidence in non-epistemic terms. See Beddor (2015b) for discussion of some difficulties on this front.

[23] Things get complicated if we allow the agent's evidence to also include doxastic states which also serve as inputs to conditionally reliable processes. After all, we'd need some way of excluding unjustified beliefs from counting as evidence. One option would be to define evidence recursively, along the lines of Reliabilist Reasons. For now, I'll set this complication aside.

body of evidence e supports believing p just in case the probability of p given e is sufficiently high. Putting these two suggestions together, we get:

Evidential Support Condition (Unpacked) A's total evidence supports believing p iff $Pr(p \mid \text{A is in states } s_1 \ldots s_n) > t$, where

- $s_1 \ldots s_n$ are all the states of A that can potentially serve as inputs to A's reliable belief-forming processes,
- t is some threshold.

What sort of probability is at issue here? One option would be to define Pr in epistemic terms: for example, we could say that Pr reflects the credences that are justified by the evidence. But clearly this is to give up any reductive ambitions.[24]

Perhaps, then, we should follow Tang (2016) in taking Pr to reflect *objective probabilities*. This would indeed result in a reductive theory. But it gives rise to a further concern: Is the theory predictive? And do the predictions vindicate our pretheoretic judgments about defeat?

To flesh out this concern, go back to **Seeing Red**. According to the view under consideration, to determine whether Lori's belief is defeated by Sal's testimony, we check the objective probability of RED conditional on Lori's total post-Sal-testimony evidence. But how do we check this? Perhaps via intuition, but it is questionable whether we have clear-cut intuitions about such probabilities. And things only get worse when consider defeater defeat. In **Two Testimony Seeing Red**, how do we determine the objective probability of RED conditional on Lori's post-Sal+Anne-testimony evidence? The worry, then, is that while this version of the Two Component View may be *consistent* with our intuitions about defeat, it does not yet *predict* these intuitions.[25]

This is hardly the final word on the matter. But at the very least it highlights the hurdles that arise when we try to develop the Two Component View in a way that is both fully reductive and predictive.

8.3 Second Advantage: No Immunity to Defeat

The second advantage of Reasons First Reliabilism stems from a structural difference between the two approaches. One hallmark of Pollock's framework is that nothing is in principle exempt from defeat: for any proposition p, it's possible in principle to

[24] See Comesaña (2018), who embraces this consequence.

[25] Some might suggest that we can mitigate this worry by further supplementing the Two Component View with a substantive theory of objective probabilities. But will any of the substantive theories on offer help? Suppose, for example, we define objective probabilities in terms of hypothetical frequencies: the objective probability of p at w is the proportion of some relevant class of worlds—say, worlds with the same initial conditions and physical laws as those that obtain at w—where p obtains. There is a worry that this just pushes the trouble back a step. Do we have clear judgments about the ratio of worlds where Lori is in her post-Sal+Anne-testimony state and RED is true to worlds where Lori is in this state and RED is false?

generate a defeater for p—just imagine a usually reliable source who tells you $\neg p$, or that your reasons do not support p. By contrast, the Evidential Support Condition singles out a class of propositions that get a free pass from defeat. Let me explain.

It's a familiar observation that whenever some part of an agent's evidence entails p, their total evidence also entails p, hence the probability of p conditional on their total evidence is 1. So probabilistic approaches to defeat predict that one cannot have a defeater for something entailed by any part of one's current evidence:

Limited Indefeasibility If some subset of A's evidence entails p at t, then A's total evidence supports p at t.

At least two sorts of cases suggest that Limited Indefeasibility runs contrary to intuition. The first comes from cases where an agent has a defeater for a belief in a necessary truth. Consider:

Logical Luck Tom comes up with a sound proof of a particular logical theorem L. Sometime later, he is told by his highly accomplished logic professor that his proof contains a mistake. Tom nonetheless disregards her testimony, continuing to believe L on the basis of his proof.

Intuitively, the professor's testimony provides an undercutting defeater for Tom's belief in L. But any view that validates Limited Indefeasibility cannot account for this intuition: since L is a necessary truth, Tom's total evidence trivially entails L.

Some might regard necessary truths as a special case, to be dealt with via independent means.[26] Still, a second class of counterexamples remains: cases where one has a defeater for some proposition that is itself part of one's evidence. According to the view of evidence under consideration, one's evidence consists in various states that serve as inputs to reliable processes. But why think that these states enjoy some special exemption from defeat? Even if we enjoy some sort of privileged access to these states, it doesn't seem that this access is indefeasible.[27] Consider:

Emotional Introspection Kilian is happy for his brother, who recently received a promotion. By introspection, Kilian comes to justifiably believe, *I am happy for my brother* (HAPPY). Later that day, he has a therapy session with an extremely well-credentialed psychiatrist, who tells him that he is mistaken: Kilian is actually jealous of his brother; he is simply unwilling to acknowledge this. While the psychiatrist mounts a compelling argument, Kilian ignores her, continuing to believe HAPPY.

[26] For example, some might propose taking a page from Stalnaker's (1999) strategy for handling the problem of logical omniscience. According to this proposal, our intuitions about **Logical Luck** not really tracking Tom's justification for believing L, but rather Tom's justification for believing some contingent proposition associated with L—for example, the proposition: S_L *is true*, where S_L is some sentence that expresses L. However, even if this strategy can be made to work, it is a mark in favor of Reasons First Reliabilism that it has no need of such maneuvers.

[27] See Armstrong (1963) and Schwitzgebel (2008) for related arguments.

Kilian's total evidence includes his happiness for his brother, since this state serves as the input to a reliable process (introspection). And this experience entails HAPPY.[28] Limited Indefeasibility thus predicts that his justification for this belief is undefeated. But this seems wrong. Even though the psychiatrist is mistaken, her testimony still provides a rebutting defeater for his belief.

Reasons First Reliabilism fares better here, since it is not committed to Limited Indefeasibility. Even if p is entailed by one of your reasons, you could still have a reliable process that is disposed to deliver either a belief that $\neg p$, or a belief that your basis for believing p does not reliably indicate its truth. In **Logical Luck** there is a reliable process available to Tom (*trusting the testimony of experts*) that, when applied to the experience of receiving his professor's testimony, is disposed to produce a belief that his proof is not a reliable guide to the truth about L. Similarly, in **Emotional Introspection** the same reliable process is available to Kilian. When applied to the experience of receiving his psychiatrist's testimony, this process is disposed to produce a belief in $\neg$HAPPY. For the Reasons First Reliabilist, no belief enjoys a principled immunity to defeat.

8.4 Taking Stock

Reasons First Reliabilism has certain affinities with extant hybrids of evidentialism and reliabilism: both are attempts to meld reliabilism with theoretical frameworks that are usually associated with internalism. However, there are important theoretical differences between the two approaches. While a fully detailed comparison would require a paper in its own right, I've given some reason to think that these differences speaker in favor of Reasons First Reliabilism. In particular, Reasons First Reliabilism is both reductive and and predictive, whereas it proves difficult to develop the Two Component View in a way that enjoys both these virtues. Second, the Two Component View—and probabilistic approaches to defeat more generally—grant a certain class of propositions a principled exemption from defeat. Reasons First Reliabilism bestows no such favors.

Of course, given the affinities between the two approaches, some may be inclined to classify Reasons First Reliabilism as a type of hybrid view. Should they do so, I would raise no objection. The important point is that if it is a hybrid view, it is the most promising one to date, at least when it comes to handling defeat.[29]

[28] It entails it both in the sense that its content (trivially) entails HAPPY, and in the sense that the fact that Killian has this experience entails HAPPY. So regardless of whether we understand evidential support in terms of probabilities conditional on the contents of an agent's states or in terms of probabilities conditional on the fact that the agent is in these states, the counterexample goes through.

[29] Another view that bears some resemblance to Reasons First Reliabilism is the reasons first virtue epistemology developed in Sylvan and Sosa (2018). According to Sylvan and Sosa, facts about what an agent is justified in believing are determined by facts about what she has sufficient epistemic reason to believe, which are in turn determined by facts about her competent attractions to assent to various propositions. While there are a number of similiarities between the two approaches, there is also a crucial difference. Sylvan and Sosa do not offer their reasons first brand of virtue epistemology as a reductive approach. Rather, they take the notion of a 'competent attraction to assent' to be a normatively

9 Conclusion

For most of their history, the reliabilist tradition and the Reasons First tradition have been developed in isolation from each other. In this paper, I've argued that an integration of the two approaches proves mutually beneficial. In particular, the sort of Reasons First Reliabilism developed here avoids reliabilism's difficulties with defeat, while still preserving the explanatory advantages that make reliabilism attractive.

While I have focused on justification, my conclusions also have implications for the study of knowledge. Recently, some authors sympathetic to externalism have argued for the surprising conclusion that knowledge is indefeasible.[30] One argument for this bold conclusion is that we have no satisfactory externalist story about how knowledge defeat works.[31] The conclusions of this paper show one way of providing such a story. As long as justification is a necessary condition on knowledge, then one can appeal to the reliabilist treatment of justification offered here to explain how knowledge is likewise subject to a defeat condition.

References

William Alston. An internalist externalism. *Synthese*, 74:265–283, 1988.

D.M. Armstrong. Is introspective knowledge incorrigible? *Philosophical Review*, 72:417–432, 1963.

Max Baker-Hytch and Matthew Benton. Defeatism defeated. *Philosophical Perspectives*, 29: 40–66, 2015.

Bob Beddor. Process reliabilism's troubles with defeat. *The Philosophical Quarterly*, 65(259): 145–159, 2015a.

Bob Beddor. Evidentialism, circularity, and grounding. *Philosophical Studies*, 172:1847–1868, 2015b.

Matthew Bedke. Developmental process reliabilism: on justification, defeat, and evidence. *Erkenntnis*, 73(1):1–17, 2010.

Michael Bergmann. *Justification without Awareness*. Oxford University Press, Oxford, 2006.

Roderick Chisholm. *Theory of Knowledge*. Prentice Hall, Englewood Cliffs, NJ, 1966.

David Christensen. Higher-order evidence. *Philosophy and Phenomenological Research*, 81(1): 185–215, 2010.

Stewart Cohen. Justification and truth. *Philosophical Studies*, 46:279–296, 1984.

Juan Comesaña. Evidentialist reliabilism. *Noûs*, 44(4):571–600, 2010.

Juan Comesaña. Whither evidentialist reliabilism? In Kevin McCain, editor, *Believing in Accordance with the Evidence*, pages 307–325. Springer, 2018.

Richard Fumerton. Foundationalism, conceptual regress, and reliabilism. *Analysis*, 48(4): 178–184, 1988.

Alvin Goldman. What is justified belief? In Pappas, editor, *Justification and Knowledge*. Reidel, Dordrecht, 1979.

Alvin Goldman. *Epistemology and Cognition*. Harvard University Press, Cambridge, MA, 1986.

loaded notion—a notion that cannot be reduced to talk of reliable processes.

[30] See e.g., Lasonen-Aarnio (2010a) and Baker-Hytch and Benton (2015).

[31] See, in particular, Baker-Hytch and Benton (2015), who use the failure of ARP as a premise in an argument for the indefeasibility of knowledge.

Alvin Goldman. Towards a synthesis of reliabilism and evidentialism. In Dougherty, editor, *Evidentialism and its Discontents*, pages 254–280. Oxford University Press, New York, 2011.
Alvin Goldman. *Reliabilism and Contemporary Epistemology: Essays*. Oxford University Press, New York, 2012.
Alvin Goldman and Bob Beddor. Reliabilist Epistemology, 2015. URL `https://plato.stanford.edu/archives/win2016/entries/reliabilism/`.
Thomas Grundmann. Reliabilism and the problem of defeaters. *Grazer Philosophische Studien*, 79(1):65–76, 2009.
David Henderson, Terry Horgan, and Matjaž Potrč. Transglobal evidentialism-reliabilism. *Acta Analytica*, 22:281–300, 2007.
James Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65(4):575–603, 1998.
Jaegwon Kim. What is 'naturalized epistemology'? *Philosophical Perspectives*, 2:381–405, 1988.
Hilary Kornblith. *Knowledge and its Place in Nature*. Clarendon Press, Oxford, 2002.
Jennifer Lackey. Testimonial knowledge and transmission. *The Philosophical Quarterly*, 49(197): 471–490, 1999.
Maria Lasonen-Aarnio. Unreasonable knowledge. *Philosophical Perspectives*, 24:1–21, 2010a.
Maria Lasonen-Aarnio. Is there a viable account of well-founded belief? *Erkenntnis*, 72: 205–231, 2010b.
Jack Lyons. *Perception and Basic Beliefs*. Oxford University Press, Oxford, 2009.
Jack Lyons. Should reliabilists be worried about demon worlds? *Philosophy and Phenomenological Research*, 86:1–40, 2013.
Jack Lyons. Goldman on evidence and reliability. In Kornblith and McLaughlin, editors, *Goldman and his Critics*. Blackwell, Oxford, 2016.
Emilia Miller. Liars, tigers, and bearers of bad news, oh my! Towards a reasons account of defeat. *Philosophical Quarterly*, forthcoming.
Sarah Moss. Scoring rules and epistemic compromise. *Mind*, 120(480):1053–1069, 2011.
Richard Pettigrew. *Accuracy and the Laws of Credence*. Oxford University Press, Oxford, 2016.
Richard Pettigrew. What is justified credence? *Episteme*, pages 1–18, 2018.
John Pollock. Defeasible reasoning. *Cognitive Science*, 11:481–518, 1987.
John Pollock. How to reason defeasibly. *Artificial Intelligence*, 57:1–42, 1992.
John Pollock. Justification and defeat. *Artificial Intelligence*, 67:377–408, 1994.
John Pollock. *Cognitive Carpentry*. MIT Press, Cambridge, MA, 1995.
John Pollock. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133(2):233–282, 2001.
Henry Prakken and John Horty. An appreciation of John Pollock's work on the computational study of argument. *Argumentation and Computation*, 3:1–19, 2012.
Eric Schwitzgebel. The unreliability of naive introspection. *Philosophical Review*, 117:245–273, 2008.
Robert Stalnaker. *Context and Content*. Oxford University Press, New York, 1999.
Kurt Sylvan and Ernest Sosa. The place of reasons in epistemology. In Star, editor, *The Oxford Handbook of Reasons and Normativity*. Oxford University Press, Oxford, 2018.
Weng Hong Tang. Reliability theories of justified credence. *Mind*, 125(497):63–94, 2016.