Debunking Debunking:

Explanationism, Probabilistic Sensitivity, and Why There is No Specifically Metacognitive Debunking Principle

David Bourget and Angela Mendelovici*

Abstract

On *explanationist* accounts of genealogical debunking, roughly, a belief is debunked when its explanation is not suitably related to its content. We argue that explanationism cannot accommodate cases in which beliefs are explained by factors unrelated to their contents but are nonetheless independently justified. Justification-specific versions of explanationism face an iteration of the problem. The best account of debunking is a probabilistic account according to which subject S's justification J for their belief that P is debunked when S learns that J is no more likely to be true on the hypothesis that P than on the hypothesis that $\neg P$. The probabilistic criterion is fully general, applying not only to cases where the learned undercutting defeater is a proposition about our beliefs or other mental states but to any case of undercutting defeat, providing the grounds for a debunking argument against the existence of a special, metacognitive debunking principle.

Keywords: genealogical defeat, debunking, explanationism, sensitivity, insensitivity, evidence, justification, probabilistic reasoning, Bayesian confirmation

*Forthcoming in 2023 in *Midwest Studies in Philosophy, Volume 47*. This paper is thoroughly co-authored.

1 Introduction

Several well-known arguments have something like the following form:

(1) Subject S learns that their belief that P has feature F.

(2) If subject S learns that their belief that P has feature F, then any justification S has for their belief that P is debunked.

(3) Therefore, any justification S has for their belief that P is debunked.

Arguments of this form aim to show that newly learned information *debunks* the justification for a belief, in the sense of providing an undercutting defeater for the relevant belief, where an *undercutting defeater* for a belief is a proposition learned that "neutralizes" a subject's justification for the belief. The notion of an undercutting defeater can be contrasted with that of a *rebutting defeater*, which is a proposition learned that provides positive reasons for thinking a belief is false. Debunking arguments involve providing undercutting defeaters, which need not also be rebutting defeaters. For our purposes, we will use the terms "undercut" and "debunk" interchangeably.[1]

Let us call arguments of the above form *debunking arguments*. A special case of a debunking argument is that of *genealogical debunking argument*, which is a debunking argument in which the subject learns that their belief that P has feature F by learning about the genealogy of their belief. Even when the information learned does not specifically involve the genealogy of the belief, debunking arguments are generally thought to have a metacognitive flavor in that the information learned pertains to the subject's belief *state* and its properties (such as its reliability) and not just to the proposition believed. While we will say something specifically about genealogical debunking arguments and debunking arguments with a metacognitive flavor in due course, the target of our discussion is the more general category of debunking arguments of roughly the above form.

[1] See Pollock and Cruz 1999 (pp. 196–7) for discussion of the distinction between rebutting and undercutting defeaters.

Debunking arguments have been developed targeting beliefs in moral truths,[2] realism in mathematics,[3] color realism,[4] the existence of God,[5] realism about ordinary objects,[6] and many other beliefs. In all cases, debunking arguments aim to show that the target beliefs lose their justification once their subject learns the relevant information.

In this paper, we are concerned with the conditions under which debunking occurs. Debunking arguments rely on a "linking" premise, represented as premise (2) in the above schema, which specifies sufficient conditions for learned information about a belief to undercut the justification for the belief. While the success of the argument requires only that the relevant conditions be sufficient conditions, insofar as we think (as is plausible) that debunking arguments in various domains rely on the same epistemic principles, these sufficient conditions should be applicable to a wide range of cases.

Two main views of the conditions under which debunking occurs have emerged: explanationism and modalism. According to explanationism, the information that serves as an undercutting defeater for a belief's justification is information about the explanation of the belief.[7] According to modalism, in contrast, the information that serves as an undercutting defeater of a belief's justification is information about the belief's modal features, such as that it is unsafe (roughly, the subject could easily have been wrong) or insensitive (roughly, the subject would have had the belief even if it had been false).[8]

Our view is that neither explanatory features nor modal features such as sensitivity and safety are relevant *per se*: there are no broadly applicable sufficient conditions that capture the conditions for debunking in terms of explanation or

[2]See, for example, Harman 1977, Joyce 2005, and Street 2006.

[3]See, for example, Benacerraf 1973, Field 1989, Clarke-Doane 2012, and Woods 2016.

[4]See, for example, Goldman 1992, Chalmers 2006, Mendelovici 2010, pp. 61–66, and Korman and Locke 2022a.

[5]See, for example, Plantinga 2000 and Barrett 2004.

[6]See Korman 2015.

[7]See Goldman 1967, Faraci 2018, Korman and Locke 2020, Bogardus and Perrin 2022, Korman and Locke 2022b, Korman and Locke 2023, Mendelovici 2010 (pp. 61–6), and Mendelovici and Bourget MS.

[8]For modal accounts of debunking, see Pollock 1987, Kahane 2010, Bogardus 2016, Clarke-Doane 2020, Bedke 2014, and Topey 2020.

the relevant modal features. We argue against modalist attempts to characterize the conditions for debunking in Bourget and Mendelovici MS. In the present paper, we consider explanationist views.

Sections 2 and 3 present versions of explanationism that apply directly to beliefs, as opposed to justifications for beliefs, and argue that they have trouble accounting for certain cases in which subjects have beliefs that are "fixed" (in that they cannot be changed) but are nonetheless well justified. In Section 4, we argue that versions of explanationism that apply to specific justifications for beliefs, rather than to beliefs themselves, can handle these cases but face problems in accounting for cases in which justifications are fixed but nonetheless remain in good standing, as well as other cases. In Section 5, we present and motivate our alternative account of debunking, which can handle all the problem cases. On this account, a justification for a belief is undercut when the subject learns that the justification is just as likely to be true on the hypothesis that the belief is false as on the hypothesis that the belief is true. Unlike explanationist accounts, this account provides sufficient *and necessary* conditions for debunking. Also unlike explanationist accounts, this account is general in that it applies to cases where the undercutting information acquired is not specifically about the origins or other features of a belief. We argue that these properties of our account are features, not bugs; they shed light on the conditions required for debunking, showing why alternative accounts fail, and allow us to account for all cases of debunking without the need for metacognitive principles. Indeed, if we are right, then there is no specifically metacognitive debunking principle and the belief that there is such a principle can itself be debunked (Section 6).

2 Explanationism

The guiding idea behind explanationism is that learning about a lack of appropriate explanatory connection between a belief's content and the facts that explain the belief is sufficient for debunking. An early expression of this idea

can be found in Benacerraf's (1973) remarks to the effect that realism about mathematical entities is on dubious footing if there is no explanation of how we could have reliable beliefs about such entities.

Notions of debunking might apply either to beliefs or to justifications. Most of the discussion of explanationism considers notions of debunking that characterize the debunking of beliefs. Accordingly, we begin by considering explanationist accounts of belief debunking. We consider justification-specific explanationist accounts in Section 4.

A first pass explanationist principle for the debunking of beliefs might look like this:

(Naïve Explanationism) If subject S learns that their belief that P is not at least partly explained by the fact that P, then any justification S has for their belief that P is undercut.

We use the term "learns" as a placeholder for different epistemic attitudes or events that might plausibly be thought to constitute our acquiring a defeater. On some views, for example, a subject needs to justifiably believe a proposition in order for it to serve as a defeater, while on others, a subject need only be such that it is rationally permissible for them to believe the proposition. Our own view is that the epistemic attitude required for having a defeater is that of being in a position to justifiably believe. However, since the differences between different ways of filling out "learns" are subtle and orthogonal to the differences between explanationism and alternative accounts—which pertain to *what* is learned—we set them aside. Importantly, though, our use of "learns" is not factive in that a subject may learn something false.

Although we focus on the debunking of the kinds of mental states that are typically called "beliefs," any kind of mental state that can be justified can also be subject to debunking. If, for example, we think that intuitions or perceptual experiences can be justified (perhaps they are *prima facie* or foundationally justified), then they, too, are susceptible to being debunked. For

ease of exposition, when characterizing explanationism and alternative accounts of debunking, we use the term "belief" as a catch-all for any mental state that is a candidate for justification or for justifying, including experiences with representational contents.

What it is for one (putative) fact to *explain* another can be largely left open for the purposes of most of our discussion. We do, however, want to understand the notion of explanation such that P can explain Q by (partly or fully) causing, realizing, grounding, constituting, and perhaps even being identical to Q.

Naïve Explanationism can account for some cases of debunking. Consider, for example, this case:

(Bob's Grade) Carlos learns that Bob got an A in Alice's class. On that basis, he forms the belief that Bob is a good student. Carlos later finds out that Alice gives As to all her students.

Intuitively, Carlos' belief is debunked once he learns that Alice gives As to all her students. This new information serves as an undercutting defeater for his justification for his belief that Bob is a good student. Naïve Explanationism correctly predicts that this is a case of debunking. Carlos learns that the process that explains his belief's formation does not depend in any way on Bob's being a good student: Bob's being a good student does not cause, realize, or otherwise contribute to Carlos' believing that Bob is a good student. So, by Naïve Explanationism, Carlos' belief is debunked.

While Naïve Explanationism can handle this case, it fails in many others. Perhaps most obviously, it wrongly predicts that beliefs about future events can be easily debunked since future events plausibly do not explain current beliefs (either causally or constitutively).

The problem of future events and other problems have motivated explanationists to explore alternative versions of the view. The general strategy is to broaden the set of facts that may be taken to be explanatorily relevant, taking a belief to be in good standing when it is explained by a fact that is suitably

related to it. Explanationists have suggested various ways of spelling out what constitutes being "suitably related." We can represent these views using the following schema (adapted from Korman and Locke 2023):

(Explanationist Schema) If subject S learns that their belief that P is such that there does not exist some fact Q that at least partly explains their belief that P and $R(Q, P, S)$, then any justification S has for their belief that P is undercut.

Again roughly following Korman and Locke (2023), we will consider four main ways of specifying $R(Q, P, S)$:

(Domain Approach) Q is in the (or a) domain of P. (Korman and Locke 2020)

(Third-Factor Approach) Q is either P or part of the explanation of the fact that P. (c.f. Goldman 1967, Enoch 2009)

(Support Approach) Q supports P. (c.f. Lutz 2018)

(Improved Support Approach) S takes Q to be a reason for believing P, and Q (in fact) supports P. (c.f. Korman and Locke 2023)

Naïve Explanationism can also be expressed using the Explanationist Schema. In the case of Naïve Explanationism, $R(Q, P, S)$ is true iff Q is identical to P. The above-mentioned alternative versions of explanationism, in effect, fill in R in less restrictive ways, allowing Q to be facts that are not P. It is worth pointing out, however, that some restrictions are necessary. Consider a version of explanationism that dropped the R condition (or specified some trivially satisfiable R); call this view *Extreme Explanationism*. On this view, the criterion would not apply to any belief that is at least partly explained by any fact whatsoever. This is clearly too demanding to be helpful in understanding debunking arguments about morality and other topics. For this reason, a non-trivial specification of R is needed to yield an adequate explanationist debunking criterion.

The four ways of specifying R listed above are in need of some clarification. The Domain Approach states that a subject's belief that P is debunked when the subject learns that there is no fact Q that explains the subject's belief that P and is in the same domain as P. There are many possible ways of understanding what it is for P and Q to belong to the *same domain*. If we understand the condition such that P and Q would qualify as being in the same domain when they are both about *something*, then the restriction is trivially satisfied, since all facts and propositions belong to this domain; the view collapses into Extreme Explanationism. If we understand the condition more demandingly such that every proposition pertains only to a single, unique domain, the condition is impossible to satisfy unless P and Q are taken by the subject to be identical, and the view collapses into Naïve Explanationism. Presumably, the intended interpretation employs an intermediate level of demandingness, allowing that the condition is sometimes satisfied and sometimes not satisfied. Since we are concerned with more fundamental difficulties with explanationism, we will grant that some commonsensical way of precisifying the condition can do the job.[9]

Likewise, the Third-Factor Approach needs to be precisified to yield reasonable predictions. The Third-Factor Approach states that a subject's belief B that P is debunked when the subject learns that there is no fact Q that is part of the explanation of both B and P. The challenge here is that there are more and less demanding ways of understanding what it takes for one fact to count as being *part of the explanation* of another. On an overly permissive understanding, the fact that there is something rather than nothing might be taken to count as part of the explanation of any fact. On this understanding, it is too easy for a subject to learn that some Q is part of the explanation of both B and P and the view risks collapsing into Extreme Explanationism. If, instead, we understand what it takes for Q to count as being part of the explanation of P in a way that is overly demanding, the view risks collapsing into Naïve Explanationism. Again,

[9]But see Killoren 2021 and Korman and Locke 2023 for the worry that the domain criterion cannot be specified in a way that is neither too broad nor too narrow.

we assume that there is a reasonable way of understanding the condition.

The notion of *support* invoked by the Support Approach and the Improved Support Approach is also in need of clarification. The relevant notion is that of support relative to a subject's total epistemic state. It is not entailment or any other "internal" relation that propositions might bear to one another independently of what anyone believes. To a first approximation, a proposition P supports a proposition Q (for a subject) when, given the subject's total epistemic state, the truth of P makes Q more likely to be true than it would otherwise be. Support must be understood in this manner for the support approach to yield the intended results (see Korman and Locke 2023).[10]

[10]Our formulation of the Improved Support Approach deviates slightly from Korman and Locke's, though the two formulations are equivalent. Korman and Locke seem to understand "treating as a reason" in such a way that a subject's treating Q as a reason for believing that P entails that the fact Q (at least partly) explains the subject's belief. We don't agree with this claim. It seems to us that, at least on an intuitive, everyday understanding of "treating as a reason," it makes perfect sense for someone to agree that they've treated a falsehood as a reason (hence that what they treated as a reason was neither a fact nor explanatory of their belief). We suspect this is just a terminological difference in our uses of "treating X as a reason": we take it to mean something like *considering X as a justification* for a belief whereas Korman and Locke take it to mean something like *considering X as a justification for a belief and causally responding to X in forming the belief.* On the second interpretation, "treating as a reason" bundles into R the explanatory condition in the Explanationist Schema, which is why Korman and Locke do not explicitly specify such a condition in their own statement of their view. For our purposes, it is useful to understand the notion of treating as a reason such that it does not entail the explanationist element, which is why we formulate the view as we do in the main text.

The motivation for adopting the Improved Support Approach over the Support Approach is that it can account for cases such as the following from Korman and Locke (2023):

> (Switches) Sonya finds herself in an illuminated room. She's not sure if the overhead lights are on or if the room is being lit by the Sun through an open skylight. She looks at the wall and sees two switches. One is labeled 'lights' and is in the 'on' position. The other is labeled 'skylight' and is in the 'closed' position. Because she sees the switches in these positions, she believes that the lights are on and that the skylight is closed. She then learns, however, that these switches control the lights and skylight in some other room, and that the state of the lights and skylight in her room has nothing to do with these switches. Sonya nevertheless continues to believe that the lights are on in her room. (p. 13)

Korman and Locke argue that the Support Approach fails to apply to this case, even though this is a kind of case that should be covered: Sonya, who believes the overhead lights are on and the skylights closed, is in a position to believe that her belief is partly explained by the fact that the overhead lights are on (she wouldn't have been able to see the switches otherwise). So her belief does not meet the sufficient conditions for debunking specified by the Support Approach. But it does meet the conditions specified by the Improved Support Approach because the overhead lights being on isn't something that she takes as a reason to believe the overhead lights are on.

3 Problems for Explanationism

The central intuition behind explanationism, captured in the Explanationist Schema, is that an absence of explanatorily relevant facts bearing the right relation to a belief's content is enough to debunk the belief. This intuition can be shown to be false by considering cases in which a belief is justified but not in a way that is relevant to the explanation of the belief.

Consider this case:

(Gala Apples) When Gabby was a child, before she had even seen or heard of Gala apples, a random cosmic event caused her to acquire the belief that ripe Gala apples are red. The random cosmic event made her such that no future reasoning or experience could either strengthen or weaken her belief. When Gabby grew up, she encountered ripe Gala apples. After enough experiences of particular ripe Gala apples as red, she became justified in believing that ripe Gala apples are red. She eventually learned that her belief was acquired and maintained as a result of the random cosmic event and that her experiences of Gala apples had no impact on her having or maintaining this belief.

Learning about the random cosmic event does not result in Gabby's belief being debunked. This is because Gabby's experiences of ripe Gala apples justify her belief and learning about the cause of her belief does not affect this justification. Her justification, of course, is epiphenomenal, in that it does not cause, constitute, or otherwise play a role in generating or sustaining the belief. Still, in light of this justification, Gabby's belief is not defeated when she learns about the random cosmic event. Her belief remains in good standing.

All of the versions of explanationism discussed in the previous section incorrectly predict that Gala Apples is a case of debunking. The Domain Approach predicts it is a case of debunking because Gabby learns that what explains her belief is the random cosmic event, which she does not take to be in the domain

of the belief's content (it has nothing to do with the study of food, fruits, plants, colors, color vision, etc.). The Third-Factor Approach predicts that Gabby's belief is debunked because she learns that there is no discernible common causal, constitutive, or other explanatorily relevant factor between Gala apples' being red and her belief (she believes that her belief was caused by an unrelated random cosmic event). The Support Approach predicts that Gala Apples is a case of debunking because Gabby learns that what explains her belief (the random cosmic event) does not lend any support to Gala apples being red. *A fortiori*, the Improved Support Approach predicts that the belief is debunked because Gabby learns that no fact that explains her belief supports the belief's content *and* is such that she takes it as a reason for the belief. So, all the versions of explanationism that we've considered incorrectly predict that Gala Apples is a case of debunking.

We've only considered four versions of explanationism, but it is easy to see that there is no other way of filling out the explanationist schema such that it does not incorrectly classify Gala Apples as a case of debunking. What protects Gabby's belief from being debunked is her perceptual evidence. But this evidence is not considered by explanationism because it does not figure in any explanation of Gabby's belief. The fundamental problem is that beliefs can be justified in ways that are independent of their explanations. Since explanationism bases predictions of debunking solely on facts that explain the having of a belief, it is susceptible to mispredicting cases in which subjects have justifications for their beliefs that are independent of their explanations.

Gala Apples is a little far-fetched, involving a highly unlikely (but arguably nomologically possible) random cosmic event. We can easily come up with analogous examples that are more realistic. All it takes is for a subject to have a belief that P that is not causally or metaphysically affected by a good justification that they have for P. Such an "epiphenomenal" justification protects the belief from debunking, but we can easily imagine the subject learning facts

about the belief's origin that make it debunkable by the explanationist's criterion. Consider, for example, these two more realistic scenarios:

(Moral Intuitions) Wilma has an innate belief that human parents are morally obligated to care for their children. This belief was the result of evolution by natural selection, which favored humans who cared for their young, and which was in no way responsive to any moral facts in the area. Wilma's innate belief is solidified by her upbringing, becoming fixed and unchangeable. Later in life, Wilma takes a philosophy course, where she encounters an *a priori* argument for the claim that human parents are morally obligated to care for their children. Wilma considers this argument to be very strong and takes it to justify her antecedent belief. The argument itself, however, does not explain why she has or maintains this belief. Wilma also takes a course on evolutionary biology and comes to learn that her belief was caused by evolutionary pressures that do not in any way track moral facts, which, she came to believe from her philosophy course, are epiphenomenal.

(Ordinary Objects) Cory has an innate, fixed belief that the world contains ordinary objects like tables and chairs. Even though he is able to consider alternative views, such as the view that there are no composite objects, no arguments or experiences could affect his innate belief in ordinary objects. Cory's convictions about ordinary objects eventually collide with others' skepticism, which leads Cory to write an academic book with excellent arguments for the existence of ordinary objects. Cory subsequently learns that his belief in the existence of ordinary objects is caused and sustained by innate psychological mechanisms that carve up reality in useful ways regardless of the metaphysical truths in the area. He considers himself lucky to have had an innate, fixed belief that was causally and constitutively disconnected from the truth but that happened to nonetheless be true and justifiable (as

shown in his book).[11]

As in Gala Apples, these scenarios show that no version of explanationism provides a sufficient condition for debunking. In Moral Intuitions, Wilma learns that there is no fact that explains her belief and is suitably related to it through support, a common causal factor, or a shared domain. In particular, while her justification provides genuine support for her belief (by her lights), she (correctly) believes it plays no role in the explanation of her belief, which she (correctly) believes is fixed independently. So, the belief is debunked according to explanationism. But this is the wrong prediction. In fact, the belief is not debunked, since Wilma has a perfectly good justification for it. Similarly, explanationism incorrectly predicts that in Ordinary Objects, Cory's belief in ordinary objects is debunked since Cory believes it originates in factors that have nothing to do with the metaphysical facts about ordinary objects. But it is not in fact debunked because Cory has independent, unrelated arguments justifying his belief, even though he recognizes that the arguments play no role in causing or sustaining it.

We can imagine more cases following the same pattern: (1) a subject has a fixed belief (e.g., a belief that $1+1=2$, a belief that objects have colors, a belief that there is a God); (2) the subject acquires a good, independent justification for the belief that bears no explanatory connection to their having the belief; and (3) the subject learns that their belief was acquired and sustained in a way that is not discerning of the truth of the matter (and impervious to any further justification). In such cases, in effect, the impact of the independent justification is preempted by the causal factors resulting in the fixed belief. Explanationism predicts that the belief is debunked, but this is the wrong answer—the subject has good, independent justification for it, so the belief is not debunked.

[11] For a defense of ordinary objects, see Korman 2016. For a discussion of debunking arguments about ordinary objects, see Korman 2020, 2014.

4 Justification Explanationism

Earlier, we noted that a debunking criterion might be taken to apply to beliefs or, instead, to particular justifications for beliefs. So, one might wonder whether an amended explanationist account that applies to particular justifications for beliefs has the resources to handle cases like Gala Apples, Moral Intuitions, and Ordinary Objects, since these are cases in which subjects appear to have multiple paths to their beliefs, only some of which are problematic. In Moral Intuitions, for example, one might say that any initial justification Wilma had for her belief is debunked, while the justification she later acquires in her philosophy course is not debunked. In this section, we consider the prospects of such an approach. We argue that it has some advantages over the standard approach of taking debunking to apply in the first instance to beliefs but that it ultimately fails.

A justification-specific criterion for debunking aims to specify conditions that are at least sufficient for a particular justification for a belief to be undercut. Such criteria can serve as specifications of the second premise in justification-specific debunking arguments, which would have something like the following form:

(1) Subject S learns that their justification J for their belief that P has feature F.

(2) If subject S learns that their justification J for their belief that P has feature F, then S's justification J for their belief that P is debunked.

(3) Therefore, S's justification J for their belief that P is debunked.

Justification-specific explanationist accounts of debunking can be constructed using the following schema:

(Justification Explanationism) If subject S learns that their justification J for their belief that P is such that there does not exist some fact Q that at least partly explains S's having of J and $R(Q, P, S)$, then S's justification J for their belief that P is debunked.

We leave it somewhat open what it is for a subject to *have* a justification J for P. At the very least, this requires believing or otherwise mentally committing to the truth of J. Presumably, it also requires taking J to support or be a reason for P. We take justifications themselves to be propositions believed rather than the believing or otherwise representing of propositions. For example, in Bob's grade, Carlos' justification for his belief that Bob is a good student is the proposition *Bob got an A in Alice's class*. Carlos has this justification in that he believes it and takes it to support his belief that Bob is a good student.[12] We take it that when multiple propositions must work together to justify a belief, the resulting justification is the conjunction of all these propositions. A subject who has multiple independent reasons for a belief may have two or more justifications.[13]

Note that Justification Explanationism takes the relevant condition to be that the subject learn the following: it is not the case that there is a Q such that Q is part of the explanation of *their having of* J and $R(Q, P, S)$. The condition is *not* that the subject learn this: it is not the case that there is a Q such that Q is part of the explanation of J *itself* and $R(Q, P, S)$. Since explanationism is concerned with the explanation of mental states, we take this to be the most natural justification-specific version of explanationism.[14]

[12]Readers who wish to use the term "justification" to mean what we mean by "having a justification" can amend the discussion as needed.

[13]Note that whether a subject's justification for a belief counts as debunked or rebutted depends on what exactly is considered to be part of the justification. If Carlos' justification was the more comprehensive proposition *Bob got an A in Alice's class and Alice only gives As to good students*, what Carlos learns would rebut rather than undercut his justification. In practice, what exactly are a subject's reasons for believing a proposition—and hence whether a case is one of debunking or rebutting of justifications—might be difficult to discern and might even be indeterminate.

[14]An alternative construal of Justification Explanationism focused on J rather than on the having of J might look like this:

(Justification Explanationism$_{\text{alt}}$) If a subject S with justification J for their belief that P learns that it is not the case that there exists some fact Q such that Q is part of the explanation of J and $R(Q, P, S)$, then S's justification J for their belief that P is debunked.

This criterion is too easily satisfied because it applies whenever a subject justifies a belief with a fact that they believe they cannot explain. Subjects can, presumably, accept justifications that they take to be inexplicable, as in this scenario:

(Big Bang) Donna learns that there occurred a Big Bang at the beginning of the Universe (J). From this, she infers that the Universe underwent an initial expansion (P). Donna further learns that the Big Bang has no possible explanation, being the beginning

As with the original version of the explanationist schema, there are various options for specifying R. The simplest option, as before, is the relation that requires Q and P to be identical, which results in a justification-specific version of Naïve Explanationism. This view continues to have a problem with future-looking cases. When P is a proposition about future events, the subject can easily learn that P is not part of the explanation of their having of any justification for believing P, so Naïve Explanationism incorrectly predicts that almost any justification for believing a proposition about future events is debunked. The Domain Approach, Third-Factor Approach, and Support Approach avoid this problem. The Improved Support Approach, when suitably construed, collapses into the Support Approach,[15] so we set it aside.

How might Justification Explanationism account for the debunking of foundationally justified beliefs, beliefs that are justified but don't have distinct sources of justification, such as, perhaps, perceptual beliefs or beliefs? For our purposes, we can take foundationally justified beliefs to be justified by the very fact that we have them. Then, what it takes to *have* a justification for a foundationally justified belief B is to believe that one has B and to take this fact to justify B. Whether or not it is correct to think of foundational justification in the specified way, we can stipulate that S's belief B has had its foundational justification debunked just when S *has* B has been debunked as justification for B.

For example, suppose that Alex has foundationally justified mathematical intuitions about numbers, for example, that every natural number has a successor

of both space and time.

This does not seem to be a case of debunking, but all versions of Justification Explanationism$_{\text{alt}}$ predict that it is since Donna learns that there is no Q that explains J.

[15]The Improved Support Approach's specification of R should, presumably, be amended to require the subject to take J (not Q) to be a reason for P:

(Improved Support Approach$_{\text{Justification Explanationism}}$) S takes J to be a reason for believing P, and Q (in fact) supports P.

Put this way, the first conjunct is redundant, since it is already specified in the definition of Justification Explanationism that J is one of the subject's justifications for P. Thus, the Improved Support Approach collapses into the Support Approach. The initial motivation for the Improved Support Approach over the Support Approach was that it could better handle Switches (see fn. 10). However, no improvement is necessary in the case of the Support Approach version of Justification Explanationism (see fn. 16).

(P). Alex's intuition that P can be debunked if they find out that no fact that partly explains their having the intuition is explanatorily related to P. For instance, on the Support Approach version of Justification Explanationism, the justification for their belief based on their intuitions is debunked if they learn that nothing that explains their having of these intuitions supports P.

From this account of the debunking of justifications for beliefs, we can construct an account of the debunking of beliefs themselves. There are multiple ways in which we might attempt to do so, but a first pass at the intuitive idea is that a belief is debunked when learning new information debunks justifications or chains of justification for a belief such that it leaves the belief with no remaining sources of justification. We provide a more precise account of belief debunking in the next section, where we present our preferred approach to debunking, but this rough and schematic account suffices for now.[16]

We are now in a position to see that Justification Explanationism can handle the fixed belief cases from the preceding section. In Gala Apples, what is debunked when Gabby learns about the fixedness of her belief that ripe Gala apples are red is the belief's foundational justification (if any). But her justification from her perceptual experiences is not debunked, since she does not learn that this justification is not appropriately explanatorily connected to the truth of the belief. Since her perceptual experiences plausibly have foundational justification, her belief is not debunked because it continues to have a source of justification.

[16]In fn. 15, we saw that the best understanding of the Improved Support Approach version of Justification Explanationism collapses into the Support Approach and suggested that improvement over the original Support Approach is not needed. We are now in a position to see why an improved approach is not needed: The original motivation for the Improved Support Approach was that it could handle the case of Switches better than the Support Approach (see fn. 10). But the Justification Explanationist version of the Support Approach is fully capable of handling the case. What Sonya takes to justify her belief that the overhead lights are on is the fact that the switch is in the "on" position, not the fact that the overhead lights are on. Assuming that Sonya's belief has no foundational justification, it is debunked (on the Support Approach version of Justification Explanationism) just in case her taking the switch to be in the "on" position is not explained (by Sonya's lights) by any fact that supports her belief. This is the case indeed since what explains her taking the switch to be in the "on" position is that it is in the "on" position, which she learned has nothing to do with the lights in the relevant room. Since the Support Approach can handle the case, there seems to be no need for an improved version of the approach.

Likewise, in the case of Moral Intuitions, any foundational justification that Wilma's innate moral belief might have is debunked, but the justification she acquires from her philosophy course is not debunked. Since her belief continues to have a source of justification, the belief itself is not debunked. Justification Explanationism can account for Ordinary Objects in the same way.

While Justification Explanationism is capable of dealing with the above-mentioned cases of fixed belief, the problem resurfaces at the level of justification. Consider this case:

(Fixed Justification) As in Fixed Belief, Gabby has acquired a fixed belief that ripe Gala apples are red as a result of a random cosmic event, but she also has a perceptual justification for this belief. From her fixed belief, Gabby further infers that ripe Gala apples are not blue.

Justification Explanationism incorrectly predicts that Gabby's justification provided by her belief that ripe Gala apples are red for her belief that ripe Gala apples are not blue is debunked. This is because Gabby has learned that the explanation for her having of this justification is the random cosmic event, which does not support or have anything else to do with ripe Gala apples not being blue.[17]

Justification Explanationism also has trouble handling some cases in which the subject does not *believe* their having of a justification can be explained (whether or not it can in fact be explained), incorrectly predicting that they are cases of debunking:

(The Epiphenomenalist) Frankie is a thoroughgoing epiphenomenalist about the mind: he does not think mental states causally interact with physical

[17]Note that it does not matter that Gabby's belief might plausibly be said to also be justified directly by her perceptual experiences of ripe Gala apples being red. The problem is that her belief that ripe Gala apples are red *also* justifies her belief that ripe Gala apples are not blue and this justification is incorrectly predicted to be debunked by Justification Explanationism.

Note also that it will not do to respond that Gabby's belief that ripe Gala apples are red is unjustified or otherwise in poor standing because of its defective casual origins and hence that it cannot confer justification to other beliefs. As we argued in the previous section, this belief receives justification from Gabby's perceptual experiences, so it is in good standing.

states in any way. He recognizes that there are many correlations between perceptual states and external states, but he believes that they are a lucky accident. For example, Frankie believes he experiences red squares just when there are red squares before him, but he believes this correlation cannot be explained. Frankie believes all this on the basis of sophisticated arguments. At the moment, he is thinking that there is a red square in front of him, and he believes this on the basis of his experience of a red square, which he takes to correlate with the presence of red squares.

Frankie believes that there is a correlation between his red square experiences and the presence of red squares, but he does not believe that the presence of red squares *explains* his experiences of red squares. While Frankie's overall view may appear unlikely to be true, what matters for our purposes is that Frankie *believes* it to be true. Justification Explanationism incorrectly predicts that Frankie's belief that there is a red square before him, which is initially justified by his having of an experience of a red square, is debunked as soon as he learns, thanks to his belief in epiphenomenalism, that there is no explanation of his having the experience. But this is not a case of debunking, since Frankie does think that his red square experience indicates the presence of red squares.[18]

Justification Explanationism only aims to provide sufficient conditions for debunking, but the relevant sufficient conditions are meant to apply broadly and cover all intuitive cases of debunking. This makes the following case problematic because Justification Explanation fails to predict that it is an instance of debunking:

(Barbie Land) Ken falls asleep at a party. When he wakes up, he believes he is in either Barbie Land or Los Angeles, but he has no idea which one. He dimly sees someone wearing a judicial robe. Because of his sexism, he immediately concludes that the person is a man. He then hears

[18]Such cases also make trouble for the original Explanationist Schema.

someone talking to the person and referring to them as "Judge Ken". Ken knows that Judge Ken is the unique male judge in Barbie Land.

Here, P is that the person Ken is looking at is a man. Ken's justification (J) for believing P is that the person he is looking at is wearing a judicial robe. After forming a belief that the person is a man, Ken hears someone referring to them as "Judge Ken." He at once learns both that he is in Barbie Land (where he knows that all judges except one are female, everyone is either male or female, and there is a 50:50 female:male ratio) and that what explains his belief that the person he is looking at is wearing a judicial robe is the presence of Judge Ken, the one male judge. This explanation of his experience, Q, supports P, that the person he is looking at is a man. Q is also plausibly taken to be in the same domain as P and part of the explanation of P. So, Ken's justification (that the person he sees is wearing a judicial robe) is not debunked by any version of Justification Explanationism. But Ken's justification *is* debunked when Ken learns that he is in Barbie Land (of course, his *belief* that the person is a man is not debunked because Ken also acquires a *new* justification for this belief). So, Justification Explanationism fails to provide sufficiently broad sufficient conditions for debunking (it also fails to provide necessary conditions, though, of course, it was not advertised as doing so).

In summary, while a justification-specific version of explanationism has some virtues, including that of being able to handle cases of fixed belief, it faces its own set of problems. In the next section, we outline our preferred alternative approach.

5 Probabilistic Sensitivity

Explanationism has some initial plausibility: learning about the explanation of a belief or justification can cast doubt on its epistemic standing. We want to suggest, however, that the reason learning about the explanation of a belief

or justification can have such epistemological consequences is that from this information we can often infer that our justifications are not appropriately sensitive to the truth of the beliefs they are supposed to justify. We want to suggest that learning about insensitivity is both necessary and sufficient for a belief to be debunked, making any learned explanatory facts irrelevant to debunking except insofar as they allow us to learn facts about insensitivity.

The notions of sensitivity and insensitivity found in the debunking literature are usually modal. For instance, a modal understanding of sensitivity might take proposition Q to be sensitive to proposition P when it is the case that had P been false then Q would have been false. Elsewhere, we argue that modal notions do not yield a correct account of undercutting defeat (Bourget and Mendelovici MS). Instead, we suggest, undercutting defeat should be understood in terms of *probabilistic sensitivity*:

(Probabilistic Sensitivity) Q is *probabilistically sensitive* to P just in case Q is more likely to obtain on the hypothesis that P than on the hypothesis that $\neg P$.

Probabilistic Sensitivity appeals to the notion of likelihood given a hypothesis, which, with some qualifications,[19] we equate with conditional probability (on an

[19]An important qualification is that the probability function should be construed in such a way that it only gives extreme probabilities (0 or 1) to propositions about which the subject is unbudgingly certain, which is arguably the best way for us to capture subjects' expectations using probability functions. So, acquiring new evidence should be modeled using Jeffrey conditionalization (Jeffrey 1965) so as not to assign learned evidence a probability of 1.

Another way of ensuring that the conditionals in Probabilistic Sensitivity can be understood as conditional probabilities is to "bracket" beliefs in P and Q when constructing p—we imagine "unlearning" P and Q (or their negations) if necessary. This accords with our ordinary way of understanding hypotheticals such as those on the two sides of Probabilistic Sensitivity, so this is arguably the best way to formalize the notion of probabilistic insensitivity. Suppose, for example, that Fred knows he has no lottery tickets and that he will not win the lottery. We ask him how likely he is to win the lottery on the hypothesis that he has half of the tickets. It would be a mistake for him to answer that since he will not win (something that he knows), the probability is 0. It would also be a mistake for him to answer that since he *in fact* does not have any tickets (again, something that he knows), the probability is undefined. In answering these ways, Fred would be misunderstanding hypothetical reasoning. To evaluate the relevant hypothetical, Fred must bracket his beliefs in not winning and not having half the tickets to see how the two are probabilistically related.

This bracketing suggestion is related to Howson's (1991) approach to the problem of old evidence. A well-known challenge for this solution to the problem of old evidence is that there is typically no mechanical way of "unlearning" a proposition when the proposition can be inferred from other propositions one has learned. Whether or not this is problematic in the

understanding of probability functions as representing subjective probabilities). Thus, Probabilistic Sensitivity can be expressed as follows:

(Probabilistic Sensitivity*) Q is *probabilistically sensitive* to P just in case $p(Q|P) > p(Q|\neg P)$.

Let us say that Q is *probabilistically insensitive* to P when it is not true that Q is probabilistically sensitive to P.[20] Given this notion of insensitivity, we can offer the following general account of the debunking of justifications:

(Probabilistic Debunking) Any justification J that a subject S has for a belief that P is debunked if and only if S learns that J is probabilistically insensitive to P, i.e., S learns that it is not the case that $p(J|P) > p(J|\neg P)$.

Like Justification Explanationism, this account is justification-specific: it is not an account of when a belief is debunked but rather an account of when a specific justification for a belief is debunked.

For illustration, consider again Bob's Grade. Carlos' justification for believing that Bob is a good student is that Alice gave Bob an A. He subsequently learns that Alice gives As to all her students. From this, Carlos learns that his justification is not more likely to be true on the hypothesis that Bob is a good student than on the hypothesis that Bob is not a good student. So, Carlos' justification for his belief that Bob is a good student is debunked.

An important motivation for Probabilistic Debunking is that it reflects a deep relationship between evidence and justification. On the standard Bayesian conception of evidence, Q provides evidence for P when $p(P|Q) > p(P)$. This is equivalent to saying that Q is probabilistically sensitive to P (i.e., that

context of traditional questions about old evidence, it seems to us that it isn't in the context of hypothetical claims like those made as part of Probabilistic Sensitivity. In this context, we can settle any indeterminacy in how the bracketed probability function should be construed by stipulation based on our interests—like we do when considering counterfactuals (Lewis 1973). That is presumably what we do when we successfully make sense of conditionals such as "on the hypothesis that I have half the ticket, I have a 50% chance of winning the lottery." For our purposes here, it need not be possible to formalize how we evaluate such claims.

[20]We deliberately say "not true" instead of "false" because we want to allow that Q is probabilistically insensitive to P when there is no fact of the matter as to whether Q is sensitive to P because the relevant probabilities are not defined.

$p(Q|P) > p(Q|\neg P)$).[21] Thus, Probabilistic Debunking is equivalent to Evidential Debunking, which explicitly relates debunking and evidence:

(Evidential Debunking) Any justification J that a subject S has for a belief that P is debunked if and only if S learns that it is not the case that J is evidence for P, i.e., S learns that it is not the case that $p(P|J) > p(P)$.

It is independently plausible that a justification J for a belief that P is debunked if and only if a subject learns that J is not evidence for P. Consider first the right-to-left direction of this claim, which states that when a subject learns that J is not evidence for P, J is debunked as justification for P. It is an open question whether a requirement on J being a justification that a subject has for the belief that P is that the subject (perhaps implicitly) take J to be evidence for P; however, if the subject learns that J is *not* evidence for P (if the truth of J does not in any way increase the subjective probability that P is true by their lights), they would be irrational for them to continue believing P on the basis of J. Learning that J is not evidence for P undercuts whatever justification J may have been previously thought to provide for the belief that P. Conversely, if a subject fails to learn that J is not evidence for P, then it does not matter what else the subject learns about J, P, or their belief in J or P. If J antecedently provided the subject justification for P, then their justification is not fully neutralized (though it may be weakened).[22]

[21] To be more precise, $p(P|Q) > p(P)$ is true just in case $p(Q|P) > p(Q|\neg P)$ is true (due to undefined probabilities in some cases, it may not be the case that one side is false just in case the other is). The "is true" equivalence can easily be proven using Bayes' rule, the total probability rule, and the complement rule.

[22] One might question the claim that debunking a justification J for P requires a subject to learn that J provides *no* evidence for P: perhaps it only requires learning that J is at best relatively weak evidence for P. For example, one might claim that J no longer justifies P when J fails to increase the probability of P by a certain relative amount r (i.e., when it is not the case that $p(P|J) \geq rp(P)$). We can accommodate such "threshold" views by modifying Evidential Debunking accordingly, replacing "not the case that J is evidence for P" by "not the case that $p(P|J) \geq rp(P)$". We can then construe Probabilistic Debunking in terms of a corresponding sensitivity threshold as ($p(J|P) \geq vp(J|\neg P)$, where $v = \frac{p(\neg P)}{1/r - p(P)}$), allowing us to preserve the equivalence between Probabilistic Debunking and Evidential Debunking.

In our view, however, a more plausible view is that justification comes in degrees, just like evidence, and that justification is debunked exactly to the degree to which a subject learns it

Elsewhere, we've argued that Probabilistic Debunking correctly handles many cases that are problematic for modal accounts of debunking.[23] Here we want to further motivate Probabilistic Debunking by showing that it correctly handles the cases that are problematic for explanationism.

Before we can apply Probablistic Debunking to the cases discussed in previous sections, we need to see how it can be used to debunk foundational justification and how it can be extended to account for belief debunking.

As in the case of Justification Debunking, it is natural to understand the justification for a foundational belief as being the proposition that one has the belief. We can then say that a belief's *foundational* justification is debunked when the subject learns that their having the belief is insensitive to the truth of the belief.

To define a notion of debunking that applies to beliefs rather than justifications for beliefs, we first need to generalize our notion of justification debunking to chains of justification. Let us call a *justificatory chain* for the belief that P_n a series of propositions $P_1 ... P_n$ where P_i justifies the belief that P_{i+1}. In practice, our complete justifications for our beliefs are (often partially overlapping) justificatory chains involving multiple justifications: we believe P on the basis of Q, which we believe on the basis of R, and so on. One way in which we might consider a justificatory chain debunked is if one of its members is debunked as justification for the belief in the next member. However, that is not the only way. Since sensitivity and evidence are not transitive (Fitelson 2012), it is possible to debunk a justificatory chain without debunking any of its members as justification for the belief in its successor. Consider, for example, the following case:

(Vera's Grade) Ariadne learns that Vera got an A+ in Fred's undergraduate class (A). From this, Ariadne infers that Vera is a good student (B).

is not evidence. If so, then the account of debunking need not be amended. However, a more complete view of debunking will allow for partial debunking. We discuss partial debunking in Mendelovici and Bourget MS.

[23] Bourget and Mendelovici MS.

> From this, Ariadne infers that Vera will get accepted to Excellent University's graduate program in philosophy (C). Ariadne later learns that Fred is involved in university admissions for Excellent University and that he tends to block his best students from being admitted.

When Ariadne learns that Fred tends to block his best students from admission, she learns that Vera's getting an A+ in Fred's class is not sensitive to Vera's being admitted to Excellent University, i.e., that A is not sensitive to C (indeed, she learns that it is sensitive to Vera's *not* being admitted). However, Ariadne does not learn that any proposition in her reasoning is insensitive to the proposition she takes it to support: she does not learn that A is insensitive to B or that B is insensitive to C. Nevertheless, learning that A is insensitive to C is enough to undercut her entire line of reasoning, neutralizing whatever justification A ultimately provides to her belief that C.

This example illustrates, first, that sensitivity is intransitive and, second, that debunking can occur even when every proposition in a chain of reasoning is sensitive to the next one in the chain. In order to capture all the ways that a justificatory chain can be debunked, we can extend our earlier notion of justification debunking as follows:

(Extended Probabilistic Debunking) A subject S's justificatory chain, $P_1 ... P_n$, for their belief that P_n is debunked when for some i and j such that $j > i$, S learns that $\underline{P_i \text{ is probabilistically insensitive to } P_j}$.

Probabilistic Debunking is equivalent to a special case of Extended Probabilistic Debunking in which the justificatory chain under consideration involves only two members, a justification and the proposition justified (note that this does not imply that the relevant belief does not also have a longer justificatory chain, of which the two-member chain is a part). For most purposes, Probabilistic Debunking is sufficient, but Extended Probabilistic Debunking is a more comprehensive principle allowing us to debunk entire chains of justification even

when Probabilistic Debunking does not apply. In the case of a foundationally justified mental state, M, the justificatory chain under consideration has for sole members the proposition that the subject has M and the content of M.

Let us refer to the set of propositions that are part of the complete (non-debunked) justificatory chains a subject has for a belief as the subject's *support network* for the belief. A subject's support network for a belief can shrink due to the debunking of the justificatory chains supporting the belief. We can say that a subject's belief has been debunked when its support network ends up containing no source of justification due to the debunking of some justificatory chains for the belief:

(Belief Debunking) A belief is debunked when the debunking of justificatory chains previously constituting the belief's support network leaves that network without any source of justification.

There are three main *sources of justification* that support networks can enjoy. First, some propositions in the network may be foundationally justified (such as, perhaps, the contents of certain perceptual states). We can consider foundational justification to attach to the metacognitive propositions that are the targets of foundational debunking: when these propositions are removed from a support network, foundational justification is removed from the network. Second, the propositions making up a support network may have coherentist justification stemming from the network's size or coherence. Third, the propositions making up a support network may have infinitist justification arising from a (virtuous) infinite regress within the network. Without settling which of these possible kinds of sources of justification support networks really have, we can say that a belief whose support network contains no source of justification is in bad standing.

Consider again Bob's Grade for illustration. Once Carlos' unique justification for his belief that Bob is a good student has been debunked, his support network for this belief contains no more justifications. *A fortiori*, it contains no sources

of justification. So it is debunked.

Unlike the explanationist views we've considered, Probabilistic Debunking and Belief Debunking specify necessary and sufficient conditions for the debunking of justifications and beliefs, respectively. Indeed, we can see the necessary condition specified by Probabilistic Debunking as explaining what goes wrong with the explanationist views we've considered in cases where they incorrectly predict debunking: they incorrectly predict debunking in cases where a justification is probabilistically sensitive to a belief's truth. Let us see what the probabilistic view has to say about cases that are problematic for explanationism.

In Gala Apples, Gabby first acquires a fixed belief about Gala apples and later acquires a justification for this belief. Perhaps the fixed belief has some *prima facie* foundational justification. Probabilistic Debunking predicts that any such foundational justification is debunked when Gabby learns about the origin of her belief, since Gabby learns that she is just as likely to have the belief on the hypothesis that it is true as on the hypothesis that it is false, i.e., she learns that her belief is insensitive to ripe Gala apples being red. But her perceptual justification for her belief is not debunked. We can understand her perceptual justification as the proposition that each of the experienced ripe Gala apples is red. This proposition is sensitive to the truth of the generalization that ripe Gala apples are red, i.e., it is more likely to be true on the hypothesis that ripe Gala apples are red than on the hypothesis that it is not the case that ripe Gala apples are red.[24] Since Gabby does *not* learn that her perceptual justification is insensitive to the content of her belief, Probabilistic Debunking correctly predicts that her perceptual justification is not debunked. And since her support network for her belief continues to contain a proposition with foundational justification, and hence continues to have a source of justification, the account correctly

[24]We can also understand Gabby's perceptual justification as the conjunction of the contents of all her perceptual experiences. Those who deny that perceptual experiences are representational can instead take Gabby's perceptual evidence to be the *fact* that she has perceptual experiences as of red, ripe Gala apples. Since Gabby does not learn that her having perceptual experiences as of ripe, red Gala apples is insensitive to the truth of her belief, her perceptual justification is not debunked and neither is her belief.

predicts that the belief is not itself debunked. The cases of Moral Intuitions and Ordinary Objects are handled in analogous ways.

In Fixed Justification, Gabby forms the belief that Gala apples are not blue (P) based on her fixed belief that Gala apples are red (J). The probabilistic account correctly predicts that this justification is not debunked since J is more likely to be true on the hypothesis that P than on the hypothesis that $\neg P$. Since Gabby's belief continues to have a source of justification, the belief is not debunked.

In The Epiphenomenalist, Frankie forms the belief that there is a red square in front of him (P) on the basis of the fact that he has an experience of a red square (J). He does not believe the experience can be explained, but he nonetheless believes that J is more likely on the hypothesis that P than on the hypothesis that $\neg P$. Since he does not learn that his justification is insensitive to the content of his belief, Probabilistic Debunking correctly predicts that his justification is not debunked. By Belief Debunking, the belief is not debunked, either.

In Barbie Land, Ken forms the belief that the person he sees is a man (P) on the basis of the person's wearing of a judicial robe (J). Ken then learns both that the person is Judge Ken, which provides new justification for P, and that he (Ken) is in Barbie Land, from which he learns that J is insensitive to P (in Barbie Land, it is more likely that a person is wearing a judicial robe on the hypothesis that they are not a man than on the hypothesis that they are man). So, Probabilistic Debunking correctly predicts that Ken's justification is debunked, and Belief Debunking correctly predicts that his belief is not debunked.

In short, Probabilistic Debunking and Belief Debunking can account for the cases that are problematic for explanationism, which supports our approach over explanationist alternatives.

6 There is no specifically metacognitive debunking principle

Debunking arguments are generally thought to invoke the learning of information pertaining to the genealogy of a belief or, more generally, information about the relevant belief state, such as information about its modal features or what explains it, rather than information merely about the proposition it represents. By extension, a justification-specific version of a debunking argument would seem to involve the learning of information about the having of our justifications. In short, we can say that debunking arguments are generally thought to be *metacognitive*: something that we learn about our mental states, and not just about the propositions our mental states represent, provides an undercutting defeater for our beliefs.

Unlike explanationist and modalist principles, Probabilistic Debunking is not metacognitive. On the probabilistic account, a subject S's justification J for their belief that P is debunked when S learns that J is probabilistically insensitive to P. Here J is the *proposition* playing the justificatory role, not the subject's state of *having* the justification, the latter of which might consist in a mental state of representing J and taking J to support P. Likewise, P is the proposition believed, not the subject's state of believing the proposition. What matters for debunking is learning about relations between the propositions J and P, namely learning that J is probabilistically insensitive to P. Metacognitive facts, which pertain specifically to the *having* of J or the *believing* of P, are at most indirectly relevant to debunking insofar as they allow us to learn facts about the relations between the propositions J and P themselves.

If our proposal is correct, there is no true, natural, widely applicable, metacognitive debunking principle: all modal and explanationist principles mispredict some cases because they are not equivalent to Probabilistic Debunking. Our examples in this paper highlight precisely the ways that explanationist principles

fail by diverging from Probabilistic Debunking, and our examples in Bourget and Mendelovici MS highlight the ways that modalist principles fail by diverging from Probabilistic Debunking.

Further, if we are right, then no metacognitive principle is *needed* in order to make a debunking argument. Probabilistic Debunking, a general principle of undercutting defeat, can account for all the relevant cases but does not invoke the learning of specifically metacognitive information. Of course, information about the genealogy of a belief or the having of a justification can provide information about probabilistic sensitivity, but it is only to the extent to which it does so that it is (indirectly) relevant to debunking.

A special case of debunking in which metacognitive information *is* relevant to debunking is that of the debunking of foundational justification. We saw that the best way to construe the justification of foundational beliefs is by taking the subject's justification, J, to be the proposition that the subject has the relevant belief. So, for example, in a case of moral intuitions that are taken to have foundational justification, the justification of an intuition is the proposition that the subject has the intuition. In such cases, our justifications are propositions *about* mental states, so Probabilistic Debunking tells us that we must engage in metacognitive reasoning to determine whether the states are debunked. This does not, however, make Probabilistic Debunking a metacognitive principle, since it is still the relation between the relevant J and P that matters for debunking. It just so happens that in such cases J is a proposition about mental states.

Interestingly, in practice, many significant and widely discussed debunking arguments in the literature appear to be aimed at foundational beliefs, which have no justification other than their foundational justification. In the moral case, our moral intuitions might be thought to have a foundational justification that is debunked by evolutionary considerations; in the case of colors, our color experiences seem to have a foundational justification that is debunked by color science; and in the case of ordinary objects, our experiences of ordinary objects

seem to have a foundational justification that is debunked by cognitive science. In each of these cases, the belief enjoys little justification other than its foundational justification. It is no coincidence that many debunking arguments target foundational beliefs: it is precisely when our justification is purely foundational that it is most vulnerable to debunking, since it is in such cases that we have no further propositions to support our beliefs. It also seems likely that the focus on foundational beliefs has contributed to a tendency to ignore the distinction between justification debunking and belief debunking: since foundational beliefs have no justification other than their foundational justification, justification debunking and belief debunking converge in these cases.

The fact that many "naturally occurring" debunking arguments target foundational beliefs gives rise to the illusion that there is something specifically metacognitive about such debunking arguments, that they set themselves apart from other instances of justifications being undercut, forming a natural category in need of specific debunking principles. Early attempts to elucidate such principles invoke notions of "off-track" explanations or "coincidence," while explanationist and modalist accounts are attempts to provide refined principles along similar lines. It might seem that the fact that typical debunking arguments invoke metacognitive facts supports the claim that there is a specifically metacognitive debunking principle. But we are now in a position to run a debunking argument against this justification for this claim. The argument goes as follows, where J is the proposition that there is a group of interesting, paradigm cases of debunking that involve learning about our mental states and P is the proposition that there exists a natural, specifically metacognitive debunking principle that is sufficiently broad to cover all the paradigm cases of debunking:

(1) You have learned that J is probabilistically insensitive to P.

(2) J is debunked as justification for your belief that P if and only if you learn that J is probabilistically insensitive to P. (From Probabilistic Debunking)

(3) Therefore, your justification J for your belief that P is debunked.

If the foregoing discussion is correct, then the justification for your belief that there is a special debunking principle is not probabilistically sensitive to the proposition that there is a special debunking principle. J is at least as likely to be true on the hypothesis that there is no special metacognitive debunking principle as it is on the hypothesis that there is such a principle. This undercuts any justification provided by J for your belief that there is a special metacognitive debunking principle. If there is no other justification for P, then the belief in the existence of a special metacognitive debunking principle has been debunked.

7 Conclusion

We have argued that explanationism fails to provide broadly applicable sufficient conditions for debunking. We've suggested an alternative account of debunking on which a justification for a belief is debunked when the subject learns that their justification is probabilistically insensitive to the truth of the belief. We've shown how this account can be applied to the debunking of foundationally justified beliefs and how it can be extended to construct an account of belief debunking.

Our account is not specifically metacognitive. That the account is general, along with the fact that it can be motivated *a priori* through its connection to the notion of evidence, is a virtue—it would be surprising if brand new epistemic principles took effect when it came to undercutting defeaters involving information about our mental states!

In summary, undercutting defeat involves nothing more or less than learning about the failure of evidential support.[25]

[25]A distant ancestor of this paper was presented at an invited symposium on debunking arguments against color at the 2021 Meeting of the Pacific Division of the American Philosophical Association (initially scheduled for 2020 but postponed due to COVID-19). Thanks to Dan Korman for organizing this session, Justin Clarke-Doane for commenting on the paper presented there, and the audience for excellent comments. Thanks also to Dan Korman and Justin Clarke-Doane for reading earlier drafts of the ancestral paper and for much discussion on colors, debunking, and related topics, and to two anonymous reviewers whose incompatible but fair reports motivated many improvements to the views defended in this paper and its framing and presentation.

References

Barrett, J. L. (2004). *Why Would Anyone Believe in God?* AltaMira Press, Lanham MD.

Bedke, M. S. (2014). No coincidence? *Oxford Studies in Metaethics*, 9:102–125.

Benacerraf, P. (1973). Mathematical truth. *Journal of Philosophy*, 70(19):661–679.

Bogardus, T. (2016). Only all naturalists should worry about only one evolutionary debunking argument. *Ethics*, 126(3):636–661.

Bogardus, T. and Perrin, W. (2022). Knowledge is believing something because it's true. *Episteme*, 19(2):178–196.

Bourget, D. and Mendelovici, A. (MS). Debunking, sensitivity, and evidence.

Chalmers, D. J. (2006). Perception and the fall from eden. In Gendler, T. S. and Hawthorne, J., editors, *Perceptual Experience*, pages 49–125. Oxford University Press, Oxford.

Clarke-Doane, J. (2012). Morality and mathematics: The evolutionary challenge. *Ethics*, 122(2):313–340.

Clarke-Doane, J. (2020). *Morality and Mathematics*. Oxford University Press.

Enoch, D. (2009). The epistemological challenge to metanormative realism: How best to understand it, and how to cope with it. *Philosophical Studies*, 148(3):413–438.

Faraci, D. (2018). Explanation in ethics and mathematics: Debunking and dispensability. *Analysis*, 78(2):377–381.

Field, H. H. (1989). *Realism, Mathematics & Modality*. Blackwell.

Fitelson, B. (2012). Evidence of evidence is not (necessarily) evidence. *Analysis*, 72(1):85–88.

Goldman, A. (1992). *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. Cambridge: Mass.: Mit Press.

Goldman, A. I. (1967). A causal theory of knowing. *Journal of Philosophy*, 64(12):357–372.

Harman, G. (1977). *The Nature of Morality: An Introduction to Ethics*. Oxford University Press.

Howson, C. (1991). The 'old evidence'problem. *The British Journal for the Philosophy of Science*, 42(4):547–555.

Jeffrey, R. C. (1965). *The Logic of Decision*. University of Chicago Press, New York, NY, USA.

Joyce, R. (2005). *The Evolution of Morality*. Bradford.

Kahane, G. (2010). Evolutionary debunking arguments. *Noûs*, 45(1):103–125.

Killoren, D. (2021). An occasionalist response to korman and locke. *Journal of Ethics and Social Philosophy*, 19(3).

Korman, D. Z. (2014). Debunking perceptual beliefs about ordinary objects. *Philosophers' Imprint*, 14.

Korman, D. Z. (2015). *Objects: Nothing Out of the Ordinary*. Oxford University Press UK, New York, NY.

Korman, D. Z. (2016). *Objects: Nothing Out of the Ordinary*. Oxford University Press.

Korman, D. Z. (2020). Ordinary objects. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition.

Korman, D. Z. and Locke, D. (2020). Against minimalist responses to moral debunking arguments. *Oxford Studies in Metaethics*, 15:309–332.

Korman, D. Z. and Locke, D. (2022a). On debunking color realism. In Machuca, D. E., editor, *Evolutionary Debunking Arguments: Ethics, Philosophy of Religion, Philosophy of Mathematics, Metaphysics, and Epistemology*, pages 257–277. Routledge.

Korman, D. Z. and Locke, D. (2022b). On debunking color realism. In Machuca, D. E., editor, *Evolutionary Debunking Arguments: Ethics, Philosophy of Religion, Philosophy of Mathematics, Metaphysics, and Epistemology*, pages 257–277. Routledge.

Korman, D. Z. and Locke, D. (2023). An explanationist account of genealogical defeat. *Philosophy and Phenomenological Research*, 106(1):176–195.

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

Lutz, M. (2018). What makes evolution a defeater? *Erkenntnis*, 83(6):1105–1126.

Mendelovici, A. (2010). *Mental Representation and Closely Conflated Topics*. PhD thesis, Princeton University.

Mendelovici, A. and Bourget, D. (MS). Debunking perceptual experience: A probabilistic framework and its application to the case of color experience.

Plantinga, A. (2000). *Warranted Christian Belief*. Oxford University Press USA, New York, US.

Pollock, J. and Cruz, J. (1999). *Contemporary Theories of Knowledge, 2Nd Edition*. Rowman & Littlefield.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, 11(4):481–518.

Street, S. (2006). A darwinian dilemma for realist theories of value. *Philosophical Studies*, 127(1):109–166.

Topey, B. (2020). Realism, reliability, and epistemic possibility: on modally interpreting the benacerraf–field challenge. *Synthese*.

Woods, J. (2016). Mathematics, morality, and self-effacement. *Noûs*.