Philosophical Perspectives 29 (2015): 40–66
(please cite published version)

Defeatism Defeated

Max Baker-Hytch
University of Notre Dame

Matthew A. Benton
University of Notre Dame

1 Introduction

Many epistemologists are enamored with a defeat condition on knowledge.[1] While the details can differ considerably, these "defeatists" tend to agree that in addition to a belief's being non-accidentally true and justified there is a no-defeater condition (perhaps built into the justification condition) which must be fulfilled for that belief to count as knowledge.

In this paper we present several difficulties for defeatism, understood along either internalist or externalist lines. We then argue that one who accepts the knowledge norm of belief, according to which one ought to believe only what one knows, can explain much of the motivation for defeatism. Maria Lasonen-Aarnio (2010, 2014) explains the plausibility of defeat by appeal to the knower engaging in an "unreasonable" stance.[2] Our approach supplements hers, but goes beyond it by sketching an account based on the knowledge norm of what makes belief "unreasonable"

[1] See pioneering work by Chisholm (1966, 48), Lehrer and Paxson (1969), Klein (1971), and Pollock (1986); more recent advocates include Plantinga (1993, 40–42), Swinburne (2001, 28ff.), Bergmann (2005, 2006), Kvanvig (2007), Lackey (2008, Ch. 2), and Goldberg (2013), among many others.

[2] Lasonen-Aarnio thinks of knowledge as mere safe belief: thus many cases of belief in the face of putative defeat remain knowledge, but they exhibit an unreasonable stance in virtue of manifesting a bad epistemic habit which, in the long run, will result in less knowledge despite being modally reliable in their local environment. We compare our approach with hers in §5.

in some of cases of putative defeat.[3] This is an important result, because on the one hand it respects the plausibility of the intuitions about defeat shared by many in epistemology; but on the other hand, it obviates the need to provide a unified account of defeat that plays well with the most plausible views of how knowledge fits with evidential probability.

2 Distinguishing Defeaters

Defeatists often distinguish between internal or mental state defeaters, and external facts which can defeat even if the subject is unaware of them (Bergmann 2006, ch. 6; Lackey 2008, 44–45; Pryor 2013, 90). Internalist defeatism tends to invoke as a requirement of rationality that one not believe a proposition p when one also gains evidence for, or recognizes that one cannot rule out, the truth of another proposition which indicates that one's belief that p is false or unreliably formed (or sustained).[4] Such defeaters are sometimes called *psychological* or *mental state* defeaters, which do their defeating in virtue of being "had," mentally, by the subject. Internalists of this stripe tend to think that having the relevant psychological state tends to defeat knowledge by defeating one's justification: if one gains information suggesting that one's evidence or reason E for believing that p is not as truth-conducive as one had assumed, one often thereby loses the justification one had from E for believing that p.[5] In §3 we shall examine one such approach, with a focus on what it takes to acquire a defeater.

Externalist defeatists, though they typically permit a kind of mental

[3] Cf. also discussion by Hawthorne and Srinivasan (2013, 21–25).

[4] That the defeating proposition may indicate either falsity or unreliability means that it could be either a "rebutting" or an "undercutting" defeater, respectively: see Pollock (1986) for more.

[5] Though much depends here on how the internalist defeatist thinks of justification: if justification encodes a notion of doxastic permission, and is a threshold conception which sets a different threshold for doxastic justification from that of knowledge-level justification, then in a certain kind of case one might lose knowledge-level justification without losing doxastic justification.

state defeat like that above, tend to allow that a proposition can defeat one's justification whether one is aware of that fact or not. Some externalist defeatists think of these as *normative* defeaters, that is, a doubt or belief which a subject *ought* to have (given the evidence or information available to that subject: Lackey 2008, 45 fn. 21; Goldberg forthcominga); whereas others don't require that one ought to believe them for them to defeat one's knowledge (Klein 1971, 1976). While the above gloss permits these defeaters to be false propositions, since one's evidence might support a falsehood, such defeaters are more often regarded as truths.[6] Thus many externalists about defeat endorse *factual* defeaters: a factual defeater is a true proposition (perhaps that one ought to believe given the information available) which, if added to one's beliefs or one's evidence, would render the belief in question unjustified. Though it is often left unsaid just what this involves, it presumably involves the idea that adding a defeating true proposition D to one's evidence E for p significantly lowers, or would lower, the probability of p: at the very least, $\Pr(p \mid E\&D) < \Pr(p \mid E)$, even if one does not yet believe D.[7] We consider externalist approaches at length in §4. Then in §5 we go on to offer a knowledge-centric approach to epistemic rationality.

3 Internal Difficulties for Defeat

Internalists who think of doxastic justification in terms of basing one's belief on evidence that supports it tend to think of defeat in terms of probability-lowering. On this picture, to acquire a defeater for one's belief that p is simply to add something to one's evidence such that one's evidence no longer supports p. Richard Swinburne is representative of this approach:

> [F]or the internalist, for whom a belief is justified if and only

[6] Indeed, for any case in which one's evidence supports a falsehood which ought to be believed, one might defer to the *fact that* one's evidence supports this as the relevant defeater: in this way, false normative defeaters reduce to factual defeaters after all.

[7] Cf. Klein's (1981, 174–177) "strong characterization" of defeat.

> if it is supported on objective a priori criteria by his evidence, these [defeating] 'circumstances' are just new evidence; both undermining [undercutting] and overriding [rebutting] involve the believer acquiring new evidence that together with the old evidence no longer gives enough support to the original belief to make it justified... An undermining defeater will leave the belief as likely to be true as it had been without the original positive evidence in its support. An overriding defeater makes it likely that the original belief was false. (Swinburne 2001, 28–29)[8]

Suppose then that one believes p on the basis of one's total evidence q and that one's priors are such that $\Pr(p \mid q) > \Pr(p)$. On the above approach, an undermining or undercutting defeater would be a proposition d which is added to one's evidence and which cancels or "screens off" the probabilistic support that q would otherwise lend to p. Formally, $\Pr(p \mid q \,\&\, d) \approx \Pr(p)$.[9] An undercutting defeater thus is supposed to bring it about that typically, the probability of p on one's evidence reverts to (roughly) the prior probability that p (prior, that is, to q's being added to one's evidence).

We want to suggest, however, that this picture of how undercutters can defeat knowledge is seriously flawed, at least given the two most plausible ways of spelling out which proposition is to be added to one's evidence. Consider a paradigm case of undercutting defeat. You are in a factory and stop to observe a conveyor belt carrying red-looking widgets. You have no prior information about the color of the widgets. Given the way they look, you form the belief, and come to know, that the widgets are red. An employee then tells you that the widgets are being irradiated by red lights. According to the picture just sketched about how defeat works,

[8] Compare also BonJour (1985), Pryor (2000), and Kvanvig (2007), though they do not invoke details about probability-lowering.

[9] Cf. Pollock (1987, 484), and Chandler (2013). We use "$\approx$" rather than "$=$" because it is possible for one's updating upon the new evidence including a defeater to revert one to a slightly higher or lower credence for the original p.

the proposition *someone told you that the widgets are being irradiated by red lights* is added to your evidence and is thereby supposed to screen off the part of your evidence that had made it probable for you that the widgets are red, and thus you no longer know. Of course, it matters crucially here how we characterize your initial evidence for believing that the widgets are red. Does the relevant portion of your evidence consist only of the proposition that it *appears* that the widgets are red, or does it include also the proposition that the widgets *are* red? Either way lies trouble.

If one takes the latter route, then it is clear that defeat simply cannot be modelled in terms of the probabilistic framework just sketched. For if one initially knows that the widgets are red and one's evidence comes to include that proposition, adding to one's evidence the proposition that *someone told you that the widgets are being irradiated by red lights* (or that *the widgets are being irradiated by red light*) will do nothing at all to reduce the overall probability on one's evidence that *the widgets are red*—that probability on that evidence will be 1. One prominent account of evidence which has this result is the view that one's evidence is all and only what one knows (E=K) (Williamson 2000, Ch. 9). But it isn't just E=K that allows propositions like *the widgets are red* into one's evidence when known; any view that allows more than just appearance propositions into one's evidence will allow that there can be cases where a subject's evidence includes a non-appearance proposition p and where the subject subsequently comes to believe a proposition d according to which her belief that p was formed in a defective way. In any case of this sort, adding d to the subject's evidence will not by itself do anything to diminish the probability of p on the subject's evidence.[10]

[10]Of course, if upon coming to believe d the subject drops her belief that p, p will—on E=K, at least—thereby drop out of the subject's evidence and consequently p will become less probable on her total evidence. But we take it that the defeatist picture under consideration here is supposed to get the result that p's probability is diminished on the subject's evidence *even if* the subject doesn't drop the belief that p upon coming to believe d. Note also that sometimes one's belief that p gets dropped as a result of adding a defeater to one's evidence, so p gets defeated in this sense, but p might not end up being lowered on one's evidence because one might know other things which entail p (cf. Williamson 2000, 205–206 and 219). We shall set such cases aside.

Because of this, the defeatist is likely to operate with the alternative view that one's relevant evidence includes only appearance propositions.[11] The view we have in mind also contends that:

> [the] justification you get merely by having an experience as of *p* can sometimes suffice to give you knowledge that *p* is the case. (Pryor 2000, 520; cf. also 521)[12]

So, consider again the widgets case. On this appearances-only view of evidence, your evidence by which you can initially come to know the proposition that *the widgets are red* ("*wr*") is just the proposition that *the widgets appear red* ("A*wr*"), where presumably the advocate of this view will want to say that

$$\Pr(wr \mid \mathrm{A}wr) > \Pr(wr)$$

It looks, then, as though the appearances-only view of evidence does allow us to model defeat in terms of acquiring additional evidence that screens off one's initial evidence from supporting the target proposition. For, where "*red light*" stands for the proposition that *the widgets are being irradiated by red lighting*, it is plausible that:

$$\Pr(wr \mid \mathrm{A}wr \,\&\, \mathrm{A}\textit{red light}) \approx \Pr(wr)$$

[11] Note that it won't help to opt for a mixed picture on which one's evidence for some of one's beliefs includes only appearance propositions and for some other of one's beliefs includes propositions not merely reporting appearances. For as long as one's evidence is allowed to include *any* propositions that are not merely about appearances, it is possible for cases to arise in which one's evidence for a non-appearance proposition *p* is itself part of one's evidence for *p*, and in such cases, adding to one's evidence other propositions about how one's belief in *p* was formed will not do anything to diminish the probability of *p* on one's evidence.

[12] Compare also Huemer (2013, §3.d), who characterizes the skeptical threat in terms of knowledge: skeptics argue that we don't *know* much at all about the external world. Huemer then claims that given his phenomenal conservatism (PC), the skeptical worry is largely met (though he does not say PC alone can give one knowledge). Huemer must be assuming something similar to Pryor, that undefeated appearances are sometimes enough not only to give one prima facie justification, but to give one knowledge. Similar views, on a natural reading, are espoused by BonJour (1985, 3–4), Swinburne (2001, 193–99, 220).

That is, even though *the widgets appear red* by itself raises the probability that *the widgets are red*, the conjunction of *the widgets appear red* with *it appears that the widgets are being irradiated by red lighting* does little to raise the probability that *the widgets are red*—the probability that *the widgets are red* simply reverts to (roughly) the prior probability. So far so good. However, a view of evidence that allows subjects to conditionalize only on appearances comes at great cost, as we shall now argue—a cost attaching to the initial claim that conditionalizing on A*wr* could gain one knowledge before encountering the defeater.

Let Ap represent the schematic proposition that *it appears that* p. As has been noted elsewhere,[13] for any proposition p, Ap does not, formally speaking, favour p over a corresponding sceptical hypothesis ($\neg p$ & Ap), even if one's updating on Ap results in p being more probable than $\neg p$ on one's total evidence. On a Likelihoodist account of confirmation, the Law of Likelihood says that a piece of evidence E favours a hypothesis H_1 over a competing hypothesis H_2 just in case:

(LL) $\Pr(E \mid H_1) > \Pr(E \mid H_2)$

And the three most popular accounts of non-relational Bayesian confirmation[14] all have in common the following sufficient condition for favouring, the Weak Law of Likelihood:

(WLL) E favours H_1 over H_2 if:
$\Pr(E \mid H_1) > \Pr(E \mid H_2)$ *and* $\Pr(E \mid \neg H_1) \leq \Pr(E \mid \neg H_2)$

But on either (LL) or (WLL), getting as evidence the appearance proposition Ap does not favour the hypothesis p over the hypothesis that ($\neg p$ & Ap); this is so when one's prior for $\Pr(\text{A}p \mid p) = 1$, and when one's prior $\Pr(\text{A}p \mid p) < 1$.

[13] See, e.g., Williamson (2007, 227–232).

[14] As given by Fitelson (2007, 478ff.): 1. *Difference*: $d(H, E) =_{df} \Pr(H \mid E) - \Pr(H)$
2. *Ratio*: $r(H, E) =_{df} \frac{\Pr(H|E)}{\Pr(H)}$ 3. *Likelihood-Ratio*: $l(H, E) =_{df} \frac{\Pr(E|H)}{\Pr(E|\neg H)}$

First consider the case in which one's prior probability is such that $\Pr(A_p \mid p) = 1$. This makes the left hand side of (LL) equal to 1; but of course, $(\neg p \ \& \ A_p)$ entails A_p, meaning that $\Pr(A_p \mid \neg p \ \& \ A_p) = 1$. So on (LL), where one's prior probability of getting the appearance that p conditional on p is 1, that appearance doesn't favour p over $(\neg p \ \& \ A_p)$, because the probability on each side is 1; and so that appearance is evidentially inert when it comes to favouring p over this skeptical scenario. Because it fails (LL), it fails to meet the sufficiency condition given by (WLL), for its left conjunct is just (LL).

Now we consider the case where p does not *entail* that A_p, and thus one's prior $\Pr(A_p \mid p) < 1$. For instance, the proposition that *the widgets are red* does not entail that *the widgets appear red*, even if it makes it highly probable. For any proposition p whose truth doesn't entail the corresponding appearance proposition A_p, conditionalizing on A_p will, given (LL), actually favour the sceptical hypothesis $(\neg p \ \& \ A_p)$ over p.[15] Recall that for any p, $\Pr(A_p \mid \neg p \ \& \ A_p) = 1$. So, for any p such that $\Pr(A_p \mid p) < 1$, (LL) is true, for:

$$\Pr(A_p \mid \neg p \ \& \ A_p) > \Pr(A_p \mid p)$$

But given that prior, (LL) says that not only does getting A_p fail to favour p over the skeptical hypothesis, but it actually favours the skeptical hypothesis over p.

What about (WLL)? We noted above that where one's prior was 1 for $\Pr(A_p \mid p)$, (LL) fails because the $\Pr(A_p \mid \neg p \ \& \ A_p)$ is also 1. So all the more so will it fail the left conjunct of (WLL) when one's prior $\Pr(A_p \mid p) \leq 1$. For what is needed is that it be greater than $\Pr(A_p \mid \neg p \ \& \ A_p)$, which equals 1. But since it can't be greater than 1, this scenario cannot meet (WLL)'s sufficiency condition for Bayesian favouring.[16]

[15] White (2006) makes such a point, in reply to Pryor (2000).

[16] We grant that its failure to meet this sufficient condition doesn't show that there is *no* Bayesian account of confirmation on which updating on appearances does favour p over its corresponding skeptical hypothesis. But for each of *Difference*, *Ratio*, and *Likelihood-Ratio* from fn. 14, there are hypotheses and priors on which appearances fail to favor the

In sum then, the worst case scenario is that one's appearances confirm the sceptical hypothesis over the proposition said to be appeared-to; the best case scenario, if one contends that $\Pr(A_p \mid p) = 1$, is that the evidence from appearances is evidentially inert in the sense that its appearing to one that p raises the probability of p as much, proportionately, as it does the hypothesis that one is misleadingly appeared-to that p. (Again, these results about evidential confirmation hold even if one's probability for p ends up higher than one's probability for $\neg p$ when updating on A_p.) And the problem is that it seems, in either scenario, that one wouldn't gain *knowledge* that p in the first place purely on the basis of update on an appearance proposition.

Thus internalist defeatists tend to hold two claims which are in tension with each other, namely (i) that the possibility of defeat requires that one conditionalize only upon appearance propositions, and (ii) that conditionalization upon an appearance proposition can suffice for one to *know* the proposition appeared-to (Pryor *op. cit.*). Such internalists owe us a story about how knowledge can be gained in the first place when one's initial evidence confirms a skeptical hypothesis as much or more than the appeared-to, and purportedly known, proposition. (That story will likely appeal to a probability threshold of less than 1 for knowledge. But such a story has serious drawbacks. It will predict that if one's prior probability for p began high enough, one can know p even after updating on evidence which *lowers* p's probability, so long as it is above the threshold. And if so, it is committed to one being able to know even after updating on a defeater.[17])

target hypothesis over its corresponding skeptical hypothesis; and so we are pessimistic that there is an account of Bayesian confirmation which secures the internalist defeatist all that she wants. (For similar pessimism, see Pryor 2013.)

[17]Suppose you begin with a credence well above (say) 0.95 that lottery ticket t will lose, because you've been told that t is part of a 1,000 ticket lottery; on this threshold picture of knowledge, you know t will lose. But if I then tell you it's in fact a 100 ticket lottery, your probability that t will lose goes down, and in that sense you have a defeater; but you're still above the threshold, so you know in the presence of a defeater.

4 Problems for Externalist Defeatists

We suggested in the last section that there are problems with internalist notions of defeat where knowledge is thought to be gained merely by updating on appearances, at least if defeat is modelled on a Bayesian probabilistic framework. That alone might lead one to embrace an externalism on which gaining knowledge merely requires some reliable or safe (etc.) belief-forming process.

For externalists about justification, however, simply adducing an internalist no-defeat condition on knowledge (or justification) is not an attractive option. Doing so results in a gerrymandered picture of knowledge (or justification) that includes an internalist defeat condition bearing no connection with the deep structural feature that they, *qua* externalists, take to be characteristic of knowledge (e.g. safety, sensitivity, track-record reliability, aptness).[18]

But might there be a way to account for the alleged loss of knowledge that occurs in defeat cases by appeal to straightforwardly externalist considerations having to do with a belief's objective connection to the world? If a plausible story could be told about how acquiring defeating evidence always brings it about that the target belief no longer counts as meeting various externalist candidate conditions on knowledge then externalists of various stripes could happily accommodate defeat without any need for co-opting what are basically internalist ways of thinking about defeat. But we now want to try to show that various purely externalist (non-hybrid) approaches might attempt to model defeat are highly problematic.

4.1 Method-switching: externalist method individuation

We begin, in this section and the next, by considering a strategy that might be termed *method-switching*.

Maria Lasonen-Aarnio (2010) suggests that a natural thought about how externalists might try to accommodate defeat is that acquiring de-

[18] See Greco (2010, Chap. 10) for the dilemma faced by externalists here.

feating evidence always brings it about that a subject's belief counts as being causally *sustained* by a method that is unreliable. The basic idea with method-switching is that in a case in which a subject holds a belief that was originally the product of a reliable method, the acquisition of defeating evidence for that belief brings it about that what counts as the salient method that is sustaining the target belief *switches* to a method that is unreliable. The method-switching strategy could in principle be tried with any of the various externalist conditions on knowledge that relativize epistemic reliability to a given method of belief-formation: namely, track-record (i.e. truth-ratio) reliability, sensitivity, and safety. An important decision point for theorists who employ one of these conditions in their account of knowledge is the issue of whether to individuate belief-forming methods internalistically, (that is, as supervening upon the mental states of the subject), or rather, in terms of facts about the method producing and sustaining the subject's belief that go beyond the mental states of the subject. We can call the former approach to method-individuation internalist and the latter externalist. Ultimately we think that method-switching fails given either approach, but for different reasons in each case.

Let's start with externalist method-individuation. As Lasonen-Aarnio argues, the basic difficulty with the method-switching strategy where methods are individuated in terms of the actual facts about how the subject's belief is sustained is that it is simply not plausible that in all defeat cases the subject counts as having her belief sustained by a method that is unreliable. If the belief was initially the product of a reliable method, then how could acquiring misleading evidence that the belief was unreliably formed turn that belief into the product of an *un*reliable method? Presumably, the answer will have to be that the presence of defeating evidence—albeit misleading—is something that should enter into the characterisation of the method via which the subject's belief is sustained. So, for instance, suppose that a subject with normal vision observes a red widget under ordinary lighting conditions and comes to believe that the widget is red. The salient method might initially be something like *normal vision in good lighting at close proximity*. This is a method with a good track record

(a high truth-ratio), and it is also a method that couldn't easily have led the subject to believe in a relevantly similar falsehood in nearby worlds; furthermore, the subject wouldn't have believed that the widget is red by way of that method in the nearest world in which the widget isn't red. The subject's belief formed via this method is sensitive, safe, and reliably formed. Suppose then that the subject acquires misleading evidence that the widgets are being irradiated by red lighting that would make widgets of any colour appear red. The thought, then, would be that in light of the acquisition of this defeating evidence, the salient method now becomes one whose description includes the presence of defeating evidence. We think that there are serious problems with this thought, however.

To begin with, we take it that the characterisation of the salient method that a subject uses in continuing to believe a given proposition ought to reflect the factors that are actually causally operative in sustaining the belief in question.[19] Obviously there are tricky and unresolved issues about how much detail should be included in the description of the salient method in a given case, but still, it seems that we should not include facts that make no causal contribution to the sustenance of the target belief. We shouldn't, for instance, include in the characterisation of the visual method by which I believe that it is raining outside such irrelevances as that it is a Friday or that Barack Obama is currently the US president. But then, why should we include the presence of defeating evidence in the characterisation of the salient method in a defeat case if the subject in that case is so dogmatically disposed that the presence of the defeating evidence has no impact at all upon the psychological processes that continue to sustain her belief? It isn't a satisfactory answer to this question simply to point out that we need to include the presence of defeating evidence in our characterisation of the salient method in such cases in order to explain why the subjects in such cases lose knowledge.

Something the defeatist might try instead is to include *disregarding evidence that one's belief is false/unreliably formed* in the description of the

[19] This is also the approach taken by Becker (2008) and Goldman (1986).

method which sustains the belief of a dogmatist who refuses to heed defeating evidence.[20] Perhaps there is a better case to be made for including this factor in the description of the method than there is for including the mere presence of defeating evidence. At any rate, counterfactual theories of causation[21] will tend to count disregarding evidence of falsity/unreliability as being causally relevant to the sustenance of a belief, because it will (usually) be true that if the subject had not disregarded such evidence then she would not have to continued to hold the target belief. But this thought will not ultimately be of help to the defeatist.

For one thing, it is clear that the mere fact that one disregards evidence of the falsity or unreliability of one's belief is not always sufficient to bring it about that one ought to drop one's belief. Think of cases in which one is told that one is not cold, contrary to how one is feeling. Surely no one will want to claim that one ought to drop one's belief in all such cases. The lesson here is that it will not do to describe the method in so coarse-grained a manner; rather, the method needs to be described in a way that makes reference to the nature of the counterevidence and the circumstances in which one encounters it.

But even still, it is not actually clear how including in the description of the new sustaining method the fact that one disregards counterevidence (of a certain sort and under certain conditions) will succeed in securing the defeatist's desired result: namely that in all such cases the sustaining method is unreliable (or unsafe, etc.). For whether the misleading character of that evidence is included in the description, or whether it is left out of the description, there can be cases where the new (sustaining) method is also safe, sensitive, or reliable as the original belief-formation method—not the result that the defeatist wants.[22] That is, for many safe, sensi-

[20] Thanks to John Hawthorne for suggesting this epicycle.

[21] Cf. Lewis (1973a; 2000), Paul (1998).

[22] We omit discussion of cases in which the defeating evidence is non-misleading for the simple reason that in those cases the method by which the belief was initially formed was indeed unreliable, and so the subject will not count as knowing in the first place. The difficult cases for the reliabilist defeatist are those in which the subject starts out knowing but subsequently encounters misleading evidence suggesting that her belief is

tive, or reliable belief-forming methods M, the corresponding method $M2$, namely *M whilst disregarding defeating evidence pertaining to M*, is about as safe, sensitive, or reliable as M. Beliefs formed (and sustained) by normal vision under conditions of good lighting at close proximity can be just as safe, reliable, and sensitive as those formed (and sustained) by normal vision under conditions of good lighting at close proximity in the presence of evidence that the lighting is not good.

4.2 Method-switching: internalist method individuation

To individuate methods internally, recall, is to individuate them in terms of how it seems to the subject "from the inside," so to speak, that she is forming her beliefs. At a first pass an internalist approach to method individuation might appear to make things go more smoothly for the theorist who wishes to employ the method-switching strategy in order to accommodate defeat within an externalist picture of knowledge.[23] It would be implausible, however, to claim simply that the method that a subject employs in forming and sustaining a belief is to be characterised in terms of how it seems to the subject that she is forming her belief. For this view would incorrectly classify as knowledge a whole range of cases in which the subject is misled into thinking that her beliefs are being formed in a manner that is reliable, safe, and sensitive when in fact the causal process that are responsible for generating and sustaining her beliefs are neither reliable, safe, nor sensitive.[24]

A more promising approach might invoke internalist method individ-

false or unreliably formed, hence our focus on such cases.

[23] Of course, there is the non-trivial matter that individuating methods in an internalist fashion sits uncomfortably with an otherwise externalist approach to epistemology. We will set that aside for now, however.

[24] On this approach it will be extremely hard to diagnose what goes wrong in Gettier cases, since the subjects in Gettier cases by hypothesis take themselves to be employing knowledge-conducive methods (otherwise, we take it, epistemologists would not be inclined to judge, as almost all in fact do, that the subjects in Gettier cases satisfy a range of internalist conditions on knowledge). For more on the difficulties with internalistic method individuation see Williamson (2000, 155–156) and Pritchard (2005, 152–153).

uation when the subject undergoes method-switching: the idea here is that externalist factors themselves individuate the method in general, but once it becomes plausible to a subject that her prior method of belief formation (or of suspending judgment) was different from what she had taken it to be, it comes to matter what method she takes herself to be switching to. So if a subject undergoes a switch in methods, it matters epistemically whether the method to which she thinks she is switching is (un)reliable.[25] The internalist method-switcher contends that it not only matters that the new method be reliable, but also that the subject rightly takes the method used to be reliable. This approach gives a unified treatment to cases where a subject goes from using an unreliable method to using (what they take to be) a reliable one, and cases where one goes from a reliable method to using (what they take to be) an unreliable one.

Take a subject Sue who begins with flawed testimony that it is raining outside, which Sue regards (wrongly) as reliable in her context, and she believes this testimony. But Sue later comes to learn that it is raining through direct observation of the rain, and also to learn that the testimony she had received was mere hearsay; and Sue (rightly) takes her visual method to be reliable. As a result, Sue switches from the method of believing on the basis of (flawed) testimony, to believing on what she takes to be adequate perceptual grounds; Sue seems to gain knowledge. Or take Tom, who forms the belief that a carnival procession is going past his house by looking out of the window and observing the carnival floats going by. Tom believes (or assumes) that his belief is formed under favourable lighting conditions and that he is under the influence of no impairing substances. But Tom's friend Lisa subsequently tells Tom, albeit entirely misleadingly, that she slipped LSD into his tea earlier, and Tom accepts this. Since Tom now thinks that he formed his belief under the influence of LSD, the method that counts as salient is indeed a method whose description includes *vision under the influence of LSD*. Presumably this new method is neither reliable, safe, nor sensitive, and assuming that

[25] Cf. Nozick (1981, 196).

at least one of these conditions is necessary for knowledge, Tom's belief, if he continues to hold it, therefore ceases to count as knowledge.

One drawback of this approach is that it permits cases in which a subject is wrong in significant ways about what new method she has switched to, but right that the new method she takes herself to be using is (un)reliable. Imagine that Sue, in a slight modification of the above case, comes to learn that it is raining not through directly observing the rain out a window, but by instead seeing outside her window a highly realistic video of the rain falling outside. Sue thinks her method is direct observation of the rain, but it is not; nevertheless, the video-mediated route to seeing the rain is as reliable as having seen it directly. Intuitions may be murky here, but we suspect that a defeatist who likes both externalism and an internalist approach to method-switching will intuitively judge Sue not to know, even though their view (as roughly formulated above) predicts that Sue knows. Even in a case where a subject switches from one reliable method to what she takes to be another reliable one—e.g., in Sue's case had her initial testimony been reliable, and then she came to believe that it is raining by seeing the video of the rain while thinking she is directly seeing the rain—we suspect there may be diverging intuitions about whether the subject continues to know. For if one's internal view on one's method matters epistemically, there may be pressure to think that this internal view must be getting things largely correct.

4.3 Defeating evidence as a determinant of which cases are close

One way of spelling out the notion of epistemic safety is in terms of the avoidance of error in all sufficiently close cases.[26] A case is simply an instance of forming, withholding, or continuing to hold a belief. The idea, then, is that one knows in a case α only if there is no case sufficiently close to α in which one believes falsely. The closeness of a pair of cases is

[26] Williamson (2000) often states his safety condition for knowledge in just these terms.

determined by a variety of factors, some of which weigh more heavily than others. Similarity of belief etiology—that is, the similarity of the token causal processes responsible for the generating, sustaining, or dropping of belief in the two cases—is usually thought to rank as the most important determinant of the closeness of a pair of cases alongside modal distance, i.e., the extent to which the world as a whole (not just the aspects of it that are directly relevant to the production of the target belief) differs between the two cases. On this way of construing safety, then, the reason that a case in which one is an envatted brain being fed misleading appearances will *not* count as being sufficiently close to a case in which one is forming perceptual beliefs in the ordinary way is that the token causal processes by which one's beliefs arise differs sharply between the two cases, and moreover, that the world as a whole (including aspects of it not directly relevant to the production of one's beliefs) differs significantly between the two cases.

Hawthorne (2007, 209–210) has suggested that this picture could be augmented so as to accommodate defeat. In addition to the aforementioned determinants of closeness, the closeness of a pair of cases might also, in part, be determined by the similarity of the subject's experiences. Consider a case in which you competently perform a calculation using a fully working calculator and get the correct answer. Plausibly there are no cases of error that are sufficiently close to this case in the respects mentioned earlier, and so your belief counts as safe. Suppose, however, that some time later Devious Dave tells you (albeit entirely misleadingly) that the calculator has a wiring defect such that it churns out mistaken answers half of the time. In line with Hawthorne's suggestion, the fact that you undergo this experience involving Dave's testimony makes a difference to which cases count as close. In particular, other cases in which you undergo the same experience (the one involving Dave's testimony) will now count as sufficiently close to the target case, and some of those other cases, so the thought goes, will be ones in which the calculator's wiring *really is* faulty so that you end up believing falsehoods. Hence, your belief is no longer safe.

A significant worry about this proposal concerns the degree to which it requires that similarity of experiences be privileged above similarity of belief etiology in determining closeness of cases. Suppose you are witnessing the Northern Lights using ordinary unimpaired vision and so form the belief that you are witnessing the Northern Lights. You subsequently receive misleading testimony from a seemingly trustworthy neuroscientist to the effect that you are currently under the influence of a hallucinogenic substance. Now, assuming the defeatist wants to classify this as a case of defeat, the modified safety-theoretic proposal under consideration can explain your (alleged) loss of knowledge only by insisting that your undergoing the experience of being told that you are under the influence of said substance brings sufficiently close a range of cases in which you undergo the same experience (being told that by the neuroscientist) and in which you *really are* under the influence of such a substance, and hence are forming false beliefs. And in order for this to work, very little weight indeed must be given to the great *dis*similarities in belief etiology between the target case (which involves ordinary unimpaired visual processes) and those cases in which your vision is seriously impaired by a hallucinogenic substance. But an approach which so radically prioritises similarity of experiences as the key determinant of closeness threatens unpalatable sceptical results. For if similarity of experiences is given such great weight in determining closeness of cases then it becomes very difficult to explain why radically sceptical scenarios don't threaten the everyday perceptual beliefs of subjects living in non-deceptive worlds. After all, there are possible worlds with brains in vats who undergo exactly the same experiences as we do.

Another concern is that this modified safety proposal will presumably predict that one's knowledge is defeated even when one is told something which is obviously false; recall the case where one is told that one is not cold, contrary to what one knows about how one feels. The modified safety proposal will have to ensure that it defeats knowledge only in cases of intuitive defeat; but as articulated, it overgeneralises this result.

4.4 Alternative reliable processes

According to an externalist approach to modelling defeat presented by Alvin Goldman (1979), S's belief that *p* is defeated if (and only if) S has available to her an alternative reliable process which is such that had she employed that process in addition to the one she actually used, then she would not have continued to believe that *p*. This account has two features: first, the alternative process must itself be reliable at getting one true beliefs (were one to use it), and second, it must be such that were one to employ it after forming a belief that *p*, one would drop the belief that *p*.

A case will help us to get clearer on the proposal here. Suppose that Jane comes to believe that the bank opens on Saturday as a result of Sarah's testimony, but suppose that subsequently Tom confidently tells Jane that Sarah is mistaken and that the bank has very recently changed its opening hours; and that believing on the basis of Tom's testimony is a reliable process such that were Jane to believe him, she'd (normally) have a true belief. Suppose further that Jane is disinclined to listen to Tom and sticks to her belief that the bank is open on Saturday. Many will want to say that Jane has a defeater for her belief. On Goldman's account, the reason Jane gets a defeater for her belief is that she has available to her an alternative reliable process, the one involving Tom's testimony, such that if she had used that process—i.e. if she had listened to Tom—she would have ceased to believe that the bank is open on Saturday.

Now, it is important to note that on Goldman's account an alternative process is *available* to a subject in the relevant sense only if employing that process doesn't involve engaging in any further research. Goldman says:

> [I]t seems implausible to say all "available" processes ought to be used, at least if we include such processes as gathering new evidence. Surely a belief can sometimes be justified even if additional evidence-gathering would yield a different doxastic attitude. What I think we should have in mind here are such additional processes as calling previously acquired evidence to mind, assessing the implications of that evidence, etc. (Gold-

man 1979)

So, suppose that Jane could have consulted a third friend, John, who happens to work at the bank, but that Jane in fact hasn't done so. Since she would have to do more than merely consult her memory in order to ascertain John's answer to the question of whether the bank is open on Saturday, the process involving John's testimony is not available to Jane in the sense Goldman intends. The process involving Tom's testimony, on the other hand, *is* available to Jane in the relevant sense: Jane merely has to consult her memory concerning what Tom said about whether the bank is open on Saturday.

Bob Beddor (2015) has recently presented counterexamples to both the necessity and sufficiency of this alternative reliable process account of defeat. But we see two additional flaws with Goldman's account. The first difficulty arises from cases where intuitively defeating evidence is acquired but is misleading because it is the result of an *un*reliable process.

Consider again the case involving Jane's reliably formed belief about the opening times of the bank. Goldman's explanation for why Jane gets a defeater will be that Tom has made available to Jane an alternative reliable process which is such that if Jane employs that process (in addition to the one she in fact used) then she will cease to believe that the bank opens on Saturday. Now, given that truth-reliability is a matter of yielding true belief *most of the time*, a token process needs to be classified under a salient *type* in order for there to be an answer to the question of whether a given belief was formed in a reliable manner.[27] Even setting aside the question of whether there is some principled solution to the generality problem, we can see that Goldman's way of modelling defeat won't work if we classify the process Tom makes available to Jane under so coarse a type as *testimony*; for the token process involving Sarah's testimony, which Jane actually used, will also get classified under the type *testimony*. And

[27]Hence reliabilism faces the notorious generality problem concerning *which* of the indefinitely many types exemplified by any given token process is the one under which we are to classify that token for the purposes of ascertaining reliability. For a canonical statement of the problem see Conee and Feldman (1998).

so asking what would have happened had Jane used the process type *testimony* will get the result that she would still have believed that the bank is open on Saturday. That is, on the usual Lewis-Stalnaker procedure for evaluating counterfactuals,[28] the way to determine the answer to the question "If Jane had employed the process type *testimony*, would she still have believed that the bank is open on Saturday?" is to consider whether Jane continues to believe that the bank is open on Saturday in the nearest world in which she employs the process type *testimony*. Well, Jane uses that process type in the actual world (i.e., in receiving the initial testimony from Sarah); the actual world is the nearest world to itself; and in the actual world Jane doesn't drop the belief that the bank is open on Saturday.

What is clear, then, is that token processes need to be assigned to narrower types for Goldman's approach to get the right result—namely, that the alternative process would, if used by Jane, have led her to drop the target belief. Let's see how things look when we classify the token process that Tom makes available to Jane under a slightly narrower type. What we need, it seems, is to individuate testimonial process types in a way that distinguishes between particular testifiers. The type to which we assign the process made available to Jane by Tom needs to be something at least as narrow as *Tom's testimony*. If we ask whether Jane continues to believe that the bank is open on Saturday in the nearest world in which Jane employs the narrower process type *Tom's testimony*—rather than the nearest world in which Jane uses *testimony* more generally—the answer clearly seems to be *yes*—the desired result. The problem now, though, is that it is easy to fill out the details of the case in such a way that the type *Tom's testimony* is not itself a reliable one (were Jane to believe his testimony), so that Goldman's (first) condition for defeat is no longer satisfied; yet the intuition that Jane ought to drop her belief persists. We might modify the case to say that, unbeknownst to Jane, Tom is a habitual liar, or is fond of practical jokes and enjoys deliberately misleading people. Under these conditions it is obvious that *Tom's testimony* is not a reliable process type,

[28] See Lewis (1973b) and Stalnaker (1968).

and yet, since we are stipulating that Jane has no reason to suspect such nefarious behaviour on Tom's part, it remains intuitive that Jane ought to drop her belief that the bank opens on Saturday upon receiving Tom's confident testimony to the contrary. We take it that this will be just as much of a problem if one classifies the process that Tom makes available to Jane under any type that is even narrower than *Tom's testimony*—for instance, types such as *Tom's testimony concerning practical matters* or *Tom's testimony concerning bank opening times*. It will be just as easy to fill out the details of the case in such a way that any of these narrower types is unreliable, and so won't count as alternative *reliable* processes available to Jane, and yet the intuition will still persist that Jane ought to drop her belief that the bank opens on Saturday upon receiving Tom's confident testimony to the contrary.

The other difficulty we see with Goldman's account is that it fails to deliver intuitively correct results concerning cases in which subjects do not happen to be disposed to drop the target belief upon employing an alternative reliable process that is available to them. Consider Pollock's red widget case again. On a tour of a factory John observes a red widget and comes to believe that it is red. A worker informs John, however, that the widgets are irradiated by red light as part of a quality control procedure, but John stubbornly sticks to his original belief. Now, suppose that the worker's testimony can plausibly be classified under a highly reliable process type. Still, it is perfectly possible that John is so disposed that even if he had employed the reliable process made available to him by the factory worker—i.e. even if he had accepted the factory worker's testimony and thereby come to believe that the widgets are being irradiated by red light—he would *not* have dropped the belief that the widget is red. As it happens most human beings are cognitively wired up so as to drop beliefs in circumstances in which we come to think that those beliefs were improperly formed, but that is a thoroughly metaphysically contingent matter. Goldman's account of defeat is unable to say what John is doing wrong, since it is not true of John that any alternative reliable process available to him is such that were he to use that process, he would drop the belief that the

widget is red. Since at best it can only get the correct results concerning cases involving subjects with "normal" cognitive dispositions, Goldman's account lacks the normative universalizability that an account of defeat should be able to deliver up.

4.5 Suspension of belief as proper functioning in defeat cases

A proper functionalist about externalist justification (or warrant) might want to appeal to the idea that cases of defeat are ones wherein a properly functioning epistemic agent just would, when becoming apprised of the presence of a defeater, drop the relevant belief. Alvin Plantinga, the foremost proponent of proper functionalism in epistemology, takes this sort of approach to handling defeat:

> You read in a usually reliable guidebook that the University of Aberdeen was founded in 1405 A.D.; you form the belief that it was founded then; that belief has a certain degree of warrant for you. You later read in an authoritative history of Aberdeen that the university was founded in 1495; you now no longer believe that it was founded in 1405, and (as we may put it) the warrant that belief had for you has been defeated. If things are going properly, you will no longer believe the first proposition, and will perhaps not believe the second as firmly as you would have, had you not first believed the first. Plantinga (1993, 40–41)

Proper functionalists might cash out the notion of what is epistemically proper in defeating circumstances in terms of what that agent herself would do, counterfactually speaking, if she were apprised of a defeater; or perhaps it is glossed in terms of what most normal cognitive agents in fact do in similar circumstances, and so proper function is delivered up by the statistical norm for such cognitive populations. Once this idea is in hand, the proper functionalist claims that subjects *should* drop the belief once

apprised of a defeater for it, and this is how a defeater does its defeating work: roughly, because one functions properly in the face of a defeater for one's belief that p only if one, in response to that defeater, drops one's belief that p, it must be that one fails to function properly if one continues to believe that p in the face of a defeater for that belief. And if one fails to function properly in such circumstances (if one does not do as one epistemically should), one thereby lacks epistemic justification.[29]

This approach to handling defeat within a proper function externalist framework comes with several costs. First, if it appeals to proper function in order to generate the notion of a normative defeater, then it owes us an account of what is proper about dropping a belief in the face of a (potential) defeater; and providing such an account cannot, on pain of circularity, appeal to intuitions about defeat cases, for those are the intuitions being explained.[30]

Second, providing an adequate gloss on what proper function is in such cases looks to be a tall order. Take the counterfactual idea sketched above. Suppose it is true of you that, though you now know that p on the basis of E, your dispositions are such that were you to be apprised of a true proposition D which calls into question E's support for p, you would thereby drop your belief and suspend judgment on p. Leaving some details aside,[31] this might seem a nice start to an account of why the truth of D, even if you *don't* learn that D, could prevent you from knowing that p. But it does not predict that some other agent, who is similarly evidentially situated but is *not* counterfactually disposed to drop her belief

[29] E.g., in Bergmann's JPF account (2006, 133), the defeat that would've obtained by way of clause (i), namely taking one's belief to be defeated, is achieved by clause (iia), by (no longer) functioning properly, since one *should* take one's belief to be defeated. For discussion of Plantinga's approach, see Kvanvig 2007.

[30] See Bergmann (2006, 174) for an illustration of this. (In what follows we consider two attempts at the needed proper functionalist account, and reject them both as implausible.)

[31] Such as: won't this be a much too strong notion of defeat? For in a wide range of cases, there might always be available some truth which would lead you to drop your belief (for example, learning that someone else believes differently than you do on the matter). Similar worries apply, *mutatis mutandis*, to Goldman-style appeals to alternative available conditionally reliable processes (see §4.4 above).

that p upon learning that D, has her knowledge defeated. In other words, if defeat is packaged in terms of an individual's cognitive dispositions to respond to new information, it loses normative universalizability: defeat merely turns on one's own cognitive dispositions.

Or take the statistical norm idea, which glosses defeat in terms of what most (suitably similar) cognitive agents in fact do in similar circumstances when apprised of defeating truths.[32] On this approach, proper function is understood as how most such cognitive agents do function. An obvious problem with this way of proceeding is that it makes one's view hostage to empirical fortune: if it turns out that most (suitably similar) cognitive agents who form the belief that the widgets are red continue so to believe even after being told that there are red lights illuminating them, then such a proper function account must allow that such beliefs are not defeated.

4.6 Defeated beliefs as inapt performances

Finally, we turn to consider whether the prospects for an externalist account of defeat might be better relative to a virtue-theoretic version of epistemic externalism. An increasingly prominent externalist account of knowledge is that of so-called virtue reliabilists such as Ernest Sosa (2007; 2009) and John Greco (2010). The basic insight here is that knowledge is in some sense an achievement that is the result of exercising one's cognitive abilities, or intellectual virtues.[33] For example, according to Sosa, belief is

[32] We are simplifying and idealizing, assuming that such an account could provide a plausible way of defining the similarity metrics for cognitive ability and for circumstances.

[33] It is widely held that there are in fact two sorts of virtue epistemologists, who are divided principally by their views on the nature of intellectual virtues. Virtue reliabilists such as Greco and Sosa think of intellectual virtues as including various sub-personal cognitive processes of which the subject may not be reflectively aware and for which she need not be able to take responsibility. Virtue responsibilists such as Zagzebski (1996) and Baehr (2011), by contrast, understand intellectual virtues not in terms of reliable cognitive faculties, but rather, in terms of person-level traits or habits of a certain sort. Examples of such traits might be things like open-mindedness, conscientiousness, intellectual honesty, self-awareness, love of the truth, intellectual courage, intellectual humility, and so on. It seems that this latter construal of intellectual virtues with its emphasis

like other performances in that it can be evaluated along the dimensions of accuracy, adroitness (how competently formed), and aptness: whether the belief (performance) is accurate because adroit. Knowledge, then, is apt belief: belief that is accurate because it is adroitly formed.

But as with other types of externalism adumbrated above, it is unclear how gaining would-be defeating evidence would render an already apt belief inapt. That is, gaining evidence that your performance was inapt is not itself enough to make your performance inapt;[34] likewise gaining evidence that you don't know may not itself be enough to make one not know. One might suppose that sustaining an initially aptly formed belief in the face of defeating evidence amounts to a kind of incompetence; for one might argue that the new performance of maintaining belief in light of defeating evidence renders one's ongoing and overall performance inapt, and that this is so because one is (in Sosa's terminology) in the position of thinking, at the reflective level, that one's animal belief that *p* may have been inapt. Perhaps one lacks reflective knowledge (apt metabelief) even though one may have begun with apt animal belief. But nothing in the virtue framework requires that one have reflective knowledge in order to have animal knowledge; likewise, nothing in the virtue approach renders inapt an apt belief just because one encounters misleading evidence.

5 Knowledge-Centered Epistemic Rationality

The foregoing argued that there are multiple concerns about implementing an account of defeat on either internalist or externalist frameworks.

on intellectual responsibility is committed to viewing knowledge as requiring some kind of reflective access to the processes by which one's belief is formed, and hence, this latter approach is reasonably taken to be offering a broadly internalist account of knowledge. Virtue responsibilists might thus be able to accommodate defeat into their picture of knowledge; but they also might merely have the resources for explaining why knowledge, under putative defeat conditions, counts as "unreasonable" in Lasonen-Aarnio's (2010) sense.

[34] Compare a performer impersonating a well-known figure; the performance might be apt despite the heckling critics who jeer it as an inaccurate portrayal of the celebrity (and those criticisms need not make the impersonator discontinue his performance).

In what follows we sketch a positive proposal about how to accommodate judgments about defeat, on a knowledge-centric picture of epistemic rationality.

5.1 Primary norms and derived norms

Suppose that you take yourself to be bound by a norm stating that

(R) One must: stop one's car if there is a red traffic light ahead

Presumably this is a norm by which all automobile drivers are actually expected to abide. Of course, sometimes drivers fail to obey this norm and yet their violation of it is blameless in a certain sense. Suppose that a tree is obscuring the traffic light column from view so that I fail to notice the red light and sail past obliviously. Despite violating the primary norm there is a sense in which I am not to blame.

We might try to account for that sense by appealing to a second norm, one that is derived from the primary norm:

(Ra) One must: stop one's car if one accepts or believes that there is a red traffic light ahead.

It is because I was prepared to abide by Ra, in attempt to conform to R, that I am in some sense blameless (or less blameworthy) in violating R. But in what sense is Ra *derived* from R? In the following sense: trying to follow a norm whilst simultaneously doing (or continuing to do) something which one takes to be in violation of it is irrational. It follows that trying rationally to follow a norm stating that *one must ϕ only in conditions C* involves trying to refrain from or stop (what one thinks is) ϕ-ing whenever one *thinks* that one is not in C.

Let us now suppose that belief is governed by a primary epistemic norm KNB: *One must: believe that p only if one knows that p*.[35] This is

[35] Williamson (2000, 47, 255–56), Adler (2002), Sutton (2005, 2007), Huemer (2007, 2011), Sosa (2011, 41–53), Littlejohn (2013), among others. See Benton (2014, §3) for discussion. We won't be arguing for KNB here; our aim is to show how KNB can be put to work.

formulation is equivalent to:

> (KNB) One must: not believe that p if one does not know that p.

KNB is the primary norm of epistemic permissibility, from which a guidance norm may be derived:

> (KNBa) One must: refrain from believing p if one comes to believe or accept that one's belief that p is not knowledge.

KNBa is not an iteration principle on which one must judge all one's beliefs to be knowledge (and thus there is no threat of a normative KK principle). Rather, KNBa offers guidance for responsibly conforming to KNB, and is derived from KNB as follows. KNB provides a necessary condition on the epistemic permissibility of belief. One who accepts KNB (implicitly or explicitly) will judge that if any actual or potential belief of theirs is epistemically impermissible by not being knowledge, it should not be believed. So KNBa codifies the idea that one must refrain from doing that which one believes or accepts, given KNB, to be impermissible. This is a fully general result: for any norm on which one should ϕ only in conditions C, accepting that norm and coming to believe that one is not in C gives one sufficient reason to refrain from ϕ-ing.[36]

Given KNB and its derivative KNBa, we can explain why it would seem epistemically irrational or irresponsible for subjects in so-called defeat cases to maintain their beliefs. For paradigmatic cases of defeat are ones in which one comes to believe or accept that one's belief isn't knowledge, and if so, one is in violation of KNBa. Insofar as implicit acceptance of

[36]Cf. Williamson (2000, 245 and 256); for a similar derivation from the norm of assertion, see Benton (forthcomingb). DeRose (2009, 94–95) draws a distinction between primary propriety and secondary propriety. Consider a norm stating that *one must: ϕ in conditions C*. Primary propriety consists in ϕ-ing whenever one is in fact in C. Secondary propriety, on the other hand, consists in ϕ-ing whenever one takes oneself to be in C.

KNB carries with it an implicit commitment to KNBa, that is enough to generate the judgment that there is something problematic about belief in the presence of a defeater. This puts flesh on the bones of the intuitive idea that subjects who continue to believe in cases of putative defeat are genuinely criticizable: not only is it the case that such believers "fail to follow policies that it is rational for someone with the goal of acquiring knowledge... to adopt" (Lasonen-Aarnio 2010, 15). Such believers also *violate* derivative norms of belief which are instrumental means to fulfilling the primary epistemic norm of belief. Thus our approach differs from Lasonen-Aarnio's by not locating what is criticizable or unreasonable about maintaining belief in the face of putative defeat information simply in terms of one's epistemic policies or habits which, in the long run, will yield non-knowledge (including false beliefs); our proposal of epistemic rationality in terms of KNBa locates the source of such criticizability in violating KNBa.

Routinely violating KNBa might manifest a bad epistemic policy or habit which would be bad given the goal of knowledge or even true belief, and if so it would coincide with Lasonen-Aarnio's account. But on our approach, dismissing putative defeating information as evidentially irrelevant will keep one from violating KNBa, even though it might manifest the bad habit which Lasonen-Aarnio thinks of as criticizable. We suspect that whether this in fact manifests a bad habit will depend on how often it leads one to violate KNB itself: for if one is in the long run very good at retaining knowledge in such situations, it might manifest a *good* epistemic habit. Whether one tends to retain knowledge in such situations will figure in whether one is manifesting a good or bad epistemic habit. Significantly though, our explanation (and Lasonen-Aarnio's) of intuitive judgments about cases of defeat is *consistent* with those beliefs continuing in some instances to constitute knowledge, if they were knowledge to begin with: one can violate KNBa without violating KNB itself, because one can be wrong about whether one knows.

Note also that KNB and KNBa together enable an error-theory of why some might deny multi-premise closure. For one who knows each of p and

q and r... and n might also refrain from believing the conjunction (p & q & r... & n), because one believes that one does not know the conjunction; and given KNBa, one does refrain from believing the conjunction. So acceptance of KNBa can explain why one might inclined to deny multi-premise closure, even if multi-premise closure holds.

As we noted earlier, defeatists tend to think of defeaters as coming in several varieties. Doxastic defeaters typically divide into undercutting and rebutting varieties. Let's begin with undercutting defeat.

5.2 Undercutting defeat

In cases of putative undercutting defeat, a subject who acquires such defeating information acquires evidence that their target belief is not an item of knowledge. When one acquires evidence that renders it improbable that you know that p such that one thereby comes to believe or accept that one doesn't know p, then by persisting in believing that p, one is violating the derivative norm KNBa, even though one may in fact continue to know p (supposing one did know p to begin with). So, given KNB and KNBa, there is a very natural explanation of why it can seem that one ought to give up one's belief in putative cases of defeat. It is consistent with this explanation that one's belief that p can remain knowledge, given that one can violate KNBa without violating KNB. And arguably, one's belief can remain knowledge even if it becomes improbable on one's evidence that it is knowledge.[37]

If so, however, this presents a problem for one who, as considered in §3 earlier, accepts a package of views on which both the core of idea of defeat involves probability-lowering and the probability threshold for knowledge is less than 1 (say, 0.95). Suppose one's evidence rightly gives

[37] See Williamson 2011 and 2014. One might worry that KNB and KNBa give conflicting advice in the case in which one knows some p but mistakenly thinks that one does not know it (thanks to Jessica Brown here). But note that KNB does not state that one must (continue to) believe whenever one knows, such that one violates it by dropping one's knowledgable beliefs. Rather, KNB merely states a necessary condition for the propriety of believing: it forbids the combination (Bp & ¬Kp).

one a probability of 0.96 for p, and that on this basis (given that package of views), one knows, because p is true. Still, one might gain evidence which makes the probability of p go up while also making the probability that one knows p go *down*. For suppose one then gets evidence (E) that the truth of p is (or was) subject to a large lottery, where p is (was) 0.9999 likely. The package of views under consideration claims that you initially knew p by being at 0.96. Gaining E makes p's probability go up to 0.9999, but it makes the probability that you know p go down because intuitively, one can't know that a lottery ticket will lose on the basis of mere statistical probability. But if so, the evidence from E defeats your knowledge without it lowering p's probability. So either defeat does not consist solely in probability lowering, or one can indeed continue to know even if it becomes improbable on one's evidence that one knows (and in the presence of misleading evidence, namely E, which makes probable that one does not know).[38]

5.3 Rebutting defeat

We propose to handle putative cases of rebutting defeat in much the same way as putative cases of undercutting defeat, namely, by appeal to the notion that in cases of so-called rebutting defeat, one acquires evidence that one's belief is not knowledge; and if one believes what that evidence suggests, then by KNBa one must surrender one's belief. Rebutting defeat differs from undercutting defeat in that rebutting defeat involves acquiring evidence not merely that one's belief that p lacks a certain connection to the truth that p, but that in fact one's belief is false. But were one to acquire evidence that one's belief is false, one thereby acquires evidence that one's belief is not knowledge; and if acquiring such evidence leads one to believe or accept that one does not know, then by KNBa one ought to stop believing the proposition in question.

[38] Again, our diagnosis based on KNB and KNBa claims that one can continue knowing p even if it is probable that, or one believes or accepts that, one's belief that p isn't knowledge.

Now, one might suppose that in the rebutting scenario, there is a difficulty for our suggestion that such cases really involve *acquiring* evidence that one's belief is false. Acquiring evidence that p is false should make one's probability for p go down.[39] If it is true that E=K—a thesis that would be a natural accompaniment to a knowledge-centric view such as we are seeking to defend—then p will have a probability of 1 on one's evidence just so long as one does in fact know it, and so no evidence that one could possibly acquire will be capable of diminishing p's probability on one's total evidence.

As before, one need not endorse E=K to see this point; *any* view on which known propositions are (or can become) part of one's evidence will face a similar difficulty here. If such a view holds that knowing that p lets p into one's evidence set E, then the probability of $p \mid E = 1$, and adding any other evidence to E will not drop the probability of p from 1.[40] Thus, as before in §3, much turns here on what it is to "acquire" evidence or add a proposition to one's evidence; one's story about that will in turn affect how best to characterize rebutting defeat. But we can agree that *if* one acquires evidence which does lower one's probability of p, one thereby loses knowledge that p. That case will obtain when one drops one's belief that p.

Suppose that one knows that there is an oak tree in a particular quad on the basis of seeing that there is; knowledge is factive, so an oak tree is there, and on a propositional view about evidence, presumably that proposition is now part of one's evidence. One commits it to memory; but one is later presented with information to the effect that there is no oak tree in that quad.[41] What would it take for one to "acquire" a rebutting

[39] If $\Pr(p \mid E\&D) < \Pr(\neg p \mid E\&D)$, then adding D to one's evidence E should reduce the probability of p (unless, of course, the probability of p on one's total evidence is already 1; in that case, the above inequality would not hold).

[40] This is so even if false propositions can be added to one's evidence. Note though that if $\neg p$ is added to an E containing p, one is (if p stays in E) already incoherent in virtue of violating the probability axioms.

[41] Perhaps someone testifies that there is not; or perhaps one returns to the spot where one (thought one) saw it and discovers that one now sees no tree there (and little enough time has passed where it is implausible to suppose it was removed without a trace).

defeater here? One might now believe that there is no oak tree there, and on pain of inconsistency, one drops one's belief that there is an oak tree there; alternatively, one might choose to suspend judgement on whether there is an oak tree there. Either way, on E=K, the relevant proposition (that there is an oak tree there) would no longer be a part of one's evidence. But if one *rejects* the information to the effect that there is no oak tree in that quad and continues on believing that there is, the proposition that there is no oak tree there does not get into one's evidence. And if it is not added to one's evidence, there is no clear way for that proposition to defeat one's knowledge that there is an oak tree there. More generally, there is no way for a proposition which is not added to one's evidence to reduce the probability, on that evidence, of any proposition (within or outside that evidence). (Similar results ensue if we merely take *believed* propositions to be part of one's evidence.)

A defeatist may want to call the above a normative defeater case: being presented with information to the effect that one's belief is false introduces a proposition one *should* believe which defeats even if one doesn't believe it. (Perhaps the candidate proposition is the negation of the believed proposition, or perhaps it is the proposition that one's belief about the oak has been unreliably formed or sustained.) But that approach is implausibly strong: if you're being warmed by a heat fan and simultaneously I tell you that you are cold, you can nevertheless know that you are hot and reject what I am saying.[42] So defeatists who like the normative defeat approach need a principled account of which propositions one should believe, which in turn defeat if one does not believe them.[43] That account must include a non-circular story about why (with rebutters) one should believe the negation of what one knows; and in particular, it owe us a story about what it takes to "acquire" such a defeater even when that

[42] Relatedly, this kind of case calls into question intuitions of defeat based on simultaneously getting two bits of evidence, one which on its own would enable one to know, and the other which defeats it (cf. simultaneous cases considered by Jessica Brown).

[43] Goldberg (forthcominga, forthcomingb) admirably attempts to provide the first full-fledged account of normative defeat in terms of knowledge or evidence one "should have had". For some concerns and criticisms of his approach, see Benton forthcominga.

proposition does not make it into one's evidence because one rejects the defeater.

Opting for our way of handling defeat offers the externalist a way out of the uncomfortable position considered in §4 above, namely it offers an explanation of the motivation for defeatism which also obviates the need to include a defeat condition on knowledge at all. Notably, appealing to the Knowledge Norm of Belief is no *ad hoc* move for an epistemic externalist. Rather, KNB is the most natural accompaniment to epistemic externalism. The externalist thought that it is possible for belief to be unjustified in virtue of environmental factors beyond the subject's purview is best accounted for in terms of KNB: the subject who is unlucky enough to be in an epistemically unfriendly environment is failing to live up to KNB, though she is plausibly excused since she meets the derivative norm KNBa.[44]

6 Conclusion

This paper has argued that there are implementation problems with the notion of defeat in epistemology. §3 revealed these problems on an internalist probabilistic framework in which defeaters do their work by lowering one's probability for a proposition: in particular, there is no uniform way of spelling out what it is for a defeater to be added to one's evidence which enables the defeatist to have both of what they want, namely for the defeater to lower one's probability for some p, but for the evidence (apart from that defeater) to confirm p over skeptical hypotheses concerning p. §4 argued that defeatism sits uneasily with most varieties of externalism about knowledge. §5 offers a new explanation of defeat intuitions, one which fits well with the E=K account of evidence, and which allows that one can retain knowledge in the face of cases of putative defeat. Indeed, the agonies of defeat detailed in §§3–4 might make E=K and anti-defeatism look all the more plausible.

[44] Compare Williamson's (forthcoming) account of justification vs. excuses.

Given the doubts we've raised for the place of defeat in an epistemology which fits knowledge with evidential probability, defeatists should presumably believe they have a defeater for their view. But by their own lights, if they have a defeater for their view, they are thereby unjustified in continuing to hold their view. Defeatists should consider their view defeated.[45]

References

Adler, Jonathan. 2002. *Belief's Own Ethics*. Cambridge: The MIT Press.

Baehr, Jason S. 2011. *The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology*. Oxford: Oxford University Press.

Becker, Kelly. 2008. "Epistemic Luck and the Generality Problem." *Philosophical Studies* 139: 353–366.

Beddor, Bob. 2015. "Process Reliabilism's Troubles with Defeat." *Philosophical Quarterly* 65: 145–159.

Benton, Matthew A. 2014. "Knowledge Norms." *Internet Encyclopedia of Philosophy*, ISSN 2161–0002, http://www.iep.utm.edu/kn–norms/.

——. forthcominga. "Knowledge and Evidence You Should Have Had." *Episteme* .

——. forthcomingb. "Lying, Belief, and Knowledge." In Jörg Meibauer (ed.), *The Oxford Handbook of Lying*. Oxford: Oxford University Press.

[45]Thanks to Jessica Brown, Isaac Choi, Julien Dutant, Yoaav Isaacs, Michael Pace, Jonathan Weisberg, and especially John Hawthorne for helpful feedback. This publication was made possible by support from a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation

Bergmann, Michael. 2005. "Defeaters and Higher-Level Requirements." *Philosophical Quarterly* 55: 419–436.

——. 2006. *Justification without Awareness: A Defense of Epistemic Externalism*. Oxford: Oxford University Press.

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge: Harvard University Press.

Chandler, Jake. 2013. "Defeat Reconsidered." *Analysis* 73: 49–51.

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

Conee, Earl and Feldman, Richard. 1998. "The Generality Problem for Reliabilism." *Philosophical Studies* 89: 1–29.

DeRose, Keith. 2009. *The Case for Contextualism*. Oxford: Clarendon Press.

Fitelson, Branden. 2007. "Likelihoodism, Bayesianism, and Relational Confirmation." *Synthese* 156: 473–489.

Goldberg, Sanford C. 2013. "Disagreement, Defeat, and Assertion." In David Christensen and Jennifer Lackey (eds.), *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press.

——. forthcominga. "On the Epistemic Significance of Evidence You Should Have Had." *Episteme* .

——. forthcomingb. "Should Have Known." *Synthese* .

Goldman, Alvin I. 1979. "What is Justified Belief?" In George Pappas (ed.), *Justification and Knowledge*. Dordrecht: D. Reidel.

——. 1986. *Epistemology and Cognition*. Cambridge: Harvard University Press.

Greco, John. 2010. *Achieving Knowledge*. Cambridge: Cambridge University Press.

Hawthorne, John. 2007. "A Priority and Externalism." In Sanford C. Goldberg (ed.), *Internalism and Externalism in Semantics and Epistemology*, 201–218. Oxford: Oxford University Press.

Hawthorne, John and Srinivasan, Amia. 2013. "Disagreement Without Transparency: Some Bleak Thoughts." In David Christensen and Jennifer Lackey (eds.), *The Epistemology of Disagreement: New Essays*, 9–30. Oxford: Oxford University Press.

Huemer, Michael. 2007. "Moore's Paradox and the Norm of Belief." In *Themes from G.E. Moore: New Essays in Epistemology and Ethics*, 142–57. Oxford: Oxford University Press.

——. 2011. "The Puzzle of Metacoherence." *Philosophy and Phenomenological Research* 82: 1–21.

——. 2013. "Phenomenal Conservatism." *Internet Encyclopedia of Philosophy*, ISSN 2161–0002, http://www.iep.utm.edu/phen–con/.

Klein, Peter. 1971. "A Proposed Definition of Propositional Knowledge." *Journal of Philosophy* 68: 471–482.

——. 1976. "Knowledge, Causality, and Defeasibility." *Journal of Philosophy* 73: 792–812.

——. 1981. *Certainty: A Refutation of Scepticism*. Minneapolis: University of Minnesota Press.

Kvanvig, Jonathan L. 2007. "Two Approaches to Epistemic Defeat." In Deane-Peter Baker (ed.), *Alvin Plantinga*, 107–124. Cambridge: Cambridge University Press.

Lackey, Jennifer. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

Lasonen-Aarnio, Maria. 2010. "Unreasonable Knowledge." *Philosophical Perspectives* 24: 1–21.

——. 2014. "Higher-Order Evidence and the Limits of Defeat." *Philosophy and Phenomenological Research* 88: 314–345.

Lehrer, Keith and Paxson, Jr., Thomas D. 1969. "Knowledge: Undefeated Justified True Belief." *Journal of Philosophy* 66: 225–237.

Lewis, David. 1973a. "Causation." *Journal of Philosophy* 70: 556–567.

——. 1973b. *Counterfactuals*. Malden, Mass.: Blackwell.

——. 2000. "Causation as Influence." *Journal of Philosophy* 97: 182–197.

Littlejohn, Clayton. 2013. "The Russellian Retreat." *Proceedings of the Aristotelian Society* 113: 293–320.

Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge: Harvard University Press.

Paul, L.A. 1998. "Keeping Track of the Time: Emending the Counterfactual Analysis of Causation." *Analysis* 63: 191–198.

Plantinga, Alvin. 1993. *Warrant and Proper Function*. New York: Oxford University Press.

Pollock, John L. 1986. *Contemporary Theories of Knowledge*. Totowa: Rowman & Littlefield, 1st edition. 2nd edn: 1999, with Joseph Cruz.

——. 1987. "Defeasible Reasoning." *Cognitive Science* 11: 481–518.

Pritchard, Duncan. 2005. *Epistemic Luck*. Oxford: Oxford University Press.

Pryor, James. 2000. "The Skeptic and the Dogmatist." *Noûs* 34: 517–549.

——. 2013. "Problems for Credulism." In Chris Tucker (ed.), *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism*, 89–131. Oxford University Press.

Sosa, Ernest. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*, volume 1. Oxford: Clarendon Press.

——. 2009. *Reflective Knowledge: Apt Belief and Reflective Knowledge*, volume 2. Oxford: Clarendon Press.

——. 2011. *Knowing Full Well*. Princeton: Princeton University Press.

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In N. Rescher (ed.), *Studies in Logical Theory*. Malden, Mass.: Blackwell.

Sutton, Jonathan. 2005. "Stick to What You Know." *Noûs* 39: 359–96.

——. 2007. *Without Justification*. Cambridge: MIT Press.

Swinburne, Richard. 2001. *Epistemic Justification*. Oxford: Clarendon Press.

White, Roger. 2006. "Problems for Dogmatism." *Philosophical Studies* 131: 525–557·

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

——. 2007. *The Philosophy of Philosophy*. Malden/Oxford: Blackwell Publishing.

——. 2011. "Improbable Knowing." In Trent Dougherty (ed.), *Evidentialism and its Discontents*, 147–164. Oxford: Oxford University Press.

——. 2014. "Very Improbable Knowing." *Erkenntnis* 79: 971–999.

——. forthcoming. "Justification, Excuses, and Skeptical Scenarios." In Julien Dutant and Fabian Dorsch (eds.), *The New Evil Demon*. Oxford: Oxford University Press.

Zagzebski, Linda. 1996. *Virtues of the Mind*. New York: Cambridge University Press.