

A Formal Account of Epistemic Defeat

Matthew Kotzen

Final Draft of July 31, 2008

1 Introduction

It is often the case that we have evidence for a hypothesis. Sometimes, that evidence remains in place, continuing to have the same epistemic effect that it had when we first acquired it. But sometimes, that evidence is *defeated*; in one way or another, something makes us reconsider the epistemic attitude that we were at first inclined to take toward the hypothesis on the basis of our evidence.

Epistemic defeat is an important concept in epistemology and various uses are made of it in the epistemological literature. To the extent that we wish to have a general understanding of what evidence is and how it has the epistemic force that it has, we need to understand how that force can be defeated.

There are also several more specific controversies in epistemology and the philosophy of science that focus of the notion of defeat. Here is a (very) partial list:

1. Fallibilism about knowledge has come to be the orthodoxy in epistemology, where that view is understood as the thesis that a subject can know a proposition on the basis of defeasible evidence—i.e., evidence that is such that the agent could, in principle, acquire a defeater for that evidence. Similarly, it's now common to distinguish, for example, a priori evidence from indefeasible evidence; most epistemologists hold that just because a subject believes or even knows a proposition on the basis of a priori evidence, it doesn't follow that he couldn't acquire an (a priori or a posteriori) defeater for that evidence. It's also common to distinguish immediate justification—i.e., justification that doesn't rest on antecedent justification for any other propositions—from indefeasible justification. Jim Pryor, for instance, thinks that much perceptual justification is immediate, but he doesn't think that that entails that the justification is question is indefeasible.¹
2. In *Warrant and Proper Function*,² after developing a general theory of defeat, Alvin Plantinga argues that belief in the conjunction of Darwinism and Naturalism is “self-defeating,” in the sense that those who believe

¹See, e.g., Pryor 2000.

²Plantinga 1993 a.

the conjunction rationally ought to stop believing the conjunction. More specifically, he argues that since false but evolutionarily adaptive empirical beliefs are just as likely to evolve as true adaptive empirical beliefs, any reason to believe in the conjunction of Darwinism and Naturalism is a defeater for the reliability of our empirical beliefs, and hence for the (empirical) belief in that conjunction itself.

3. In “Fine-Tuning and Multiple Universes,”³ Roger White argues that, though data about the “fine-tuning” of our universe for life is no evidence that there are multiple other universes, the existence of other universes would defeat the evidence that the fine-tuning data provides for the hypothesis that our universe was intentionally designed so as to allow for life.
4. Other debates in the philosophy of science surround whether the fact that a particular theory “accommodated,” rather than predicted, some body of data defeats the evidence that that data provides for the theory in question. Some philosophers argue for the view that all that matters is the fit between the theory and the data, and not the causal history of the design of that theory, whereas other philosophers argue that theories that accommodate the data are too easy to come by to earn genuine support from the data they accommodate.⁴

Again, this list is far from complete. And I won’t be explicitly taking a stand on any of these issues here. The point is, rather, that the notion of epistemic defeat is a central one in the discussion of many interesting issues in philosophy. For this reason, it would be useful to have an account of precisely what it is that’s at issue in these debates, and that seems to require an account of epistemic defeat.

Moreover, the distinction between different *kinds* of defeat plays an important role in epistemology. It’s standard to distinguish between two types of defeaters of S ’s evidence for her belief that q —“undercutting” defeaters and “opposing”⁵ defeaters. Very roughly, an undercutting defeater undermines the effect that the evidence has on p , without directly providing evidence against q . An opposing defeater, on the other hand, serves as “direct” evidence against q .

Suppose, for instance, that John tells me that it is raining outside and that I believe on that basis that it is raining outside. If I were to learn that John in fact has no idea what the weather is like right now but likes to make up claims about the weather and confidently share them with others, that would serve as an undercutting defeater of my evidence that it’s raining. This new information about John isn’t directly evidence that it’s not raining outside; rather, it merely undermines the evidential force that I originally took his testimony to have. If, however, I were to go outside and see for myself that it’s sunny, that experience

³White 2000.

⁴See, e.g., Achinstein 1994, Collins 1994, Harker 2006, Horwich 1982, Maher 1988, Schlesinger 1987, van Fraassen 1980, and White 2003.

⁵This latter class is more commonly referred to as “rebutting” defeaters; for reasons that will become clear later, I prefer the term “opposing.” The term is due to Jim Pryor.

would serve as an opposing defeater of my evidence that it's raining; here, my new experience does serve as direct evidence *against* the proposition that it's raining.

One thing we might wonder about is whether we can draw this distinction in a more principled and precise way—and, if so, how we should go about doing it. This is particularly important because some philosophers—for instance, Jim Pryor⁶—have expressed doubt about whether this distinction can be captured using the standard Bayesian probabilistic model of partial belief. If this doubt is justified, then this is *prima facie* evidence for some richer formal model that can capture the distinction; after all, we want a formal model of belief to capture all those distinctions that play a crucial role in epistemology.

In this paper, I will give a formal account of epistemic defeat generally, and also of the distinction between undercutting and opposing defeaters, using the standard Bayesian apparatus. I will also introduce a third kind of defeat that seems to have gone unnoticed in the epistemological literature, which I call “bi-directional” defeat. My account will also handle cases of “partial” defeat, where the evidential effect under consideration is only partly interfered with, as well as cases of “hybrid” defeat, where the same defeater plays both an undercutting and an opposing role (or an undercutting and a supporting role).

2 A First Stab

Consider again the case where I take John's testimony that it's raining to be some evidence that it's raining.

Let E be John's testimony that it's raining out and let H be the proposition that it is raining out. Let's assume that, given the background information that I have at the time that I hear John's testimony, E is evidence for H . In Bayesian terms, this amounts to the assumption that $p(H|E) > p(H)$.

Call my credence in H before collecting E , $p(H)$, my **prior credence in H** . Call my credence in H after collecting E but before acquiring any defeaters, $p(H|E)$, my **evidential credence in H** . Call my credence in H after collecting E and after acquiring defeater D , $p(H|E \wedge D)$, my **defeated credence in H** .

In my rough characterization above of the difference between undercutting and opposing defeaters, I suggested that while opposing defeaters constitute direct evidence against the relevant proposition, undercutting defeaters do not. Undercutting defeaters merely undermine the evidential force of some other evidence for the proposition, and therefore do not constitute direct evidence *against* the proposition. One might think, then, that we can characterize opposing defeaters as those propositions which, if learned, lower our credence in H (assuming E):

OPPOSING1: D is an opposing defeater for the evidence that E provides for H just in case $p(H|E \wedge D) < p(H|E)$.

⁶See Pryor ms b.

But this clearly will not work. However we should understand the distinction between the “directness” of opposing defeaters and the “indirectness” of undercutting defeaters, both kinds of defeaters are defeaters, and both can therefore lower our credence in H . Take, for instance, the information that John is a highly unreliable testifier about the weather—a paradigm case of an undercutting defeater (of John’s testimonial evidence that it’s raining). Once I learn this, it seems clear, I ought to become less confident that it’s raining than I was before finding out how unreliable John is. Thus, in this case, it’s true that $p(H|E \wedge D) < p(H|E)$ even though the D in question is an undercutting defeater. Therefore, OPPOSING1 can’t be used to uniquely characterize opposing defeaters.

Still, it’s somewhat plausible that the condition that $p(H|E \wedge D) < p(H|E)$ characterizes defeaters in general, even if it doesn’t distinguish between opposing and undercutting defeaters. It’s plausible that, when E ’s evidential effect on H is defeated, the agent’s defeated credence in H is lower than her evidential credence. In other words, when an agent acquires some defeater of E , it’s plausible that she ought to become less confident in H than she was before she acquired that defeater. So let’s provisionally accept:

DEFEATER1: D is a defeater for the evidence that E provides for H just in case $p(H|E \wedge D) < p(H|E)$.

One major advantage of DEFEATER1 over several extant theories is that it accounts for cases of *partial* defeat, where our defeated credence in H is only somewhat lower than our evidential credence in H . Most other theories of defeat are developed in a framework of binary belief, and so are insensitive to this kind of partial defeat. For example, in *Contemporary Theories of Knowledge*,⁷ Pollock and Cruz give the following characterization of defeaters:

If E is a reason for S to believe H , D is a defeater for this reason if and only if D is logically consistent with E , and $E \& D$ is not a reason for S to believe H .

Chisholm provides an account that is similar in spirit to Pollock and Cruz’s,⁸ though Chisholm sets things up in terms of defeat of some evidence’s “tendency to make a hypothesis evident”:

D defeats E ’s tendency to make H evident =_{df} E tends to make H evident; and $D \& E$ does not tend to make H evident.

A shortcoming of both of these accounts is that if E is strong evidence for H , D can be a partial defeater of that evidence even if the agent’s defeated credence in H is still above the threshold for belief. Suppose, for instance, that an agent’s prior credence in H is .1, that E justifies an evidential credence in H of .99,

⁷Pollock and Cruz 1999, p. 37. Here and throughout, I have changed the notation of the theories I’m discussing for the sake of consistency.

⁸Chisholm 1989, p. 53.

and that D justifies a defeated credence in H of .98. We'd still want to call D a (partial) defeater of the evidence that H provides for E , even if $E \wedge D$ is “a reason to believe” H (to use Pollock and Cruz’s terminology) or “makes H evident” (to use Chisholm’s). DEFEATER1 addresses this phenomenon of partial defeat, because the condition that $p(H|E \wedge D) < p(H|E)$ is perfectly consistent with $p(H|E \wedge D)$ being greater than any nonzero threshold value that we might choose.

Still, though, we have the problem of distinguishing opposing defeaters from undercutting defeaters.

One natural idea is that, since undercutting defeaters merely undermine the evidential effect that E has on H , the net effect of acquiring some undercutting defeater should be to push the agent’s credence in H back to the value it had before the agent acquired E , so that her defeated credence in H equals her prior credence in H . Thus:

UNDERCUTTING1: D is an undercutting defeater for the evidence that the E provides for H just in case $p(H|E \wedge D) = p(H)$.

By contrast, since an opposing defeater actually gives us independent evidence against H , it has the capacity to push the agent’s credence in H below the value it had before the agent acquired E , leading to a defeated credence that is lower than the prior credence. Thus, perhaps we should accept:

OPPOSING2: D is an opposing defeater for the evidence that E provides for H just in case $p(H|E \wedge D) < p(H)$.

Notice that each of UNDERCUTTING1 and OPPOSING2 individually entail DEFEATER1 (since we’re assuming that $p(H) < p(H|E)$),⁹ which is desired. Obviously, if D is either an undercutting defeater or an opposing defeater, it’s a defeater. Thus, we should expect our characterizations both of undercutting defeaters and of opposing defeaters to (individually) entail our characterization of defeaters simpliciter.

However, there are two problems here.

First, and fairly obviously, just because some opposing defeaters *can* push the agent’s credence below what it was before acquiring E , it clearly doesn’t follow that *all* opposing defeaters have that effect. The opposing defeater in the rain example above—my going outside and seeing for myself that it’s sunny out—is a particularly strong opposing defeater. Barring skeptical scenarios and elaborate tricks, the experience I have when I walk outside and seem to see sunny and cloudless skies makes me virtually certain that it’s not raining out, regardless of whose testimony about the weather I’ve previously heard. Thus, assuming that I found it at least moderately credible that it was raining out before hearing John’s

⁹The proofs are trivial: 1) Suppose UNDERCUTTING1’s condition is true. So $p(H|E \wedge D) = p(H)$. Since $p(H) < p(H|E)$, it follows that $p(H|E \wedge D) < p(H|E)$. So DEFEATER1’s condition is true. 2) Suppose OPPOSING2’s condition is true. So $p(H|E \wedge D) < p(H)$. Since $p(H) < p(H|E)$, it again follows that $p(H|E \wedge D) < p(H|E)$ by transitivity. So DEFEATER1’s condition is true.

testimony, the combined effect of John’s testimony and the opposing defeater of that testimony (i.e., my sunny experience) will be to make me less confident that it’s raining out than I originally was, satisfying OPPOSING2’s condition. But there surely are weaker opposing defeaters possible. For example, suppose that instead of going outside and observing the weather for myself, I had instead come across Jill, someone in whom I have the same confidence that I do in John. If Jill were to tell me that it’s not raining out, it’s fairly plausible that this would serve as a (weaker) opposing defeater of John’s testimony. But since I have equal confidence in John and Jill, it’s plausible that my defeated credence in H would (approximately) equal my prior credence. After all, there’s a natural sense in which these two pieces of testimonial evidence “cancel each other out.”¹⁰ Thus, for some opposing defeaters, $p(H|E \wedge D) = p(H)$ (at least approximately). Thus, we have a counterexample both to the necessity of OPPOSING2 and to the sufficiency of UNDERCUTTING1. Moreover, we can vary the example a bit by supposing that I have slightly more confidence in John than I do in Jill, so that Jill’s testimony doesn’t fully neutralize the effect of John’s. In that case, it seems that my defeated credence that it’s raining should be slightly higher than my prior credence, though obviously lower than my evidential credence. (After all, I still have some trust in Jill’s testimony, just not as much as I do in John’s.) Thus, in this case, $p(H|E \wedge D) > p(H)$, which isn’t even provided for by UNDERCUTTING1 or OPPOSING2. This is another counterexample to the necessity of OPPOSING2.

Second, and relatedly, just as an opposing defeater might only partially oppose our evidence for H (as in the case where I have slightly less confidence in Jill’s testimony than in John’s), it’s also possible for an undercutting defeater to only partially undercut our evidence for H . I considered a case above where I learned that John has no idea what the weather is like and is a highly unreliable testifier about the weather. It’s plausible that this undercutting defeater might totally neutralize the evidential impact of John’s testimony on the proposition that it’s raining out; in this case, my defeated credence that it’s raining out is exactly the same as my prior credence that it’s raining out, so UNDERCUTTING1’s condition is satisfied. But it’s clearly possible for John’s testimonial evidence that it’s raining out to be only partially undercut—say, by the information that John *sometimes* makes up claims about the weather with no evidence, or that John observed the weather over an hour ago and thus *might* not be correctly reporting the current weather. It’s plausible that these claims would serve as (partial) undercutting defeaters of John’s testimonial evidence that it’s raining out, but that they wouldn’t completely undermine the confirmatory effect that John’s testimony has on H . In other words, acquiring one of these undercutting

¹⁰Compare: I pick a ball at random out of an urn with 99 white marbles and one red marble. My prior credence that the selected ball is red is quite low (.01). Two fairly reliable people take a look at the selected ball; one tells me that it’s red and the other tells me that it’s white. Conflicting reports such as these are just as likely to occur when the selected ball is red as when it’s white, so $p(\text{REPORTS}|\text{WHITE}) = p(\text{REPORTS}|\text{RED})$. From this, it follows that that the net effect of these reports is to leave my prior credence that the ball is red unchanged. This result is completely independent of the value of my prior credence that the ball is red.

defeaters would force a defeated credence lower than my evidential credence, but still higher than my prior credence. Thus, $p(H) < p(H|E \wedge D) < p(H|E)$. This is a counterexample to the necessity of UNDERCUTTING1.

So far, nothing that I have said rules out the sufficiency of OPPOSING2. Even if we grant that not every opposing defeater generates a defeated credence in H lower than the subject's prior credence in H , still it might be true that only opposing defeaters can do so. And indeed, it's fairly plausible that OPPOSING2 does specify a sufficient condition for opposing defeat. After all, if D merely undermines the effect of E , how could D leave us with a defeated credence in H that is lower than our prior credence in H ? It seems that in order to end up with a defeated credence lower than our prior credence, we'd need some independent reason to disbelieve H . A reason merely to doubt the evidential force of E (i.e., an undercutting defeater) seems like it couldn't be such an independent reason to disbelieve H . Thus, only opposing defeaters seem to be capable of generating a defeated credence which is lower than the prior credence. So OPPOSING2's condition looks at least sufficient for opposing defeat.

3 Developing the Account

Here's where we are:

DEFEATER1: D is a defeater for the evidence that E provides for H just in case $p(H|E \wedge D) < p(H|E)$.

OPPOSING2' (only a sufficient condition): D is an opposing defeater for the evidence that E provides for H if $p(H|E \wedge D) < p(H)$.

Obviously, we'd like to say more. We'd like a necessary condition for opposing defeat, as well as necessary and sufficient conditions for undercutting defeat. What should we say about cases where $p(H) < p(H|E \wedge D) < p(H|E)$? These cases satisfy DEFEATER1's condition, so D will count as a defeater. But such cases don't satisfy the sufficient condition specified in OPPOSING2', so we have no guarantee that D will count as an opposing defeater. So, is D an opposing defeater or an undercutting defeater in such cases? As argued above, it might be either. We just don't yet have the means to determine which.

Here's how Pollock and Cruz characterize opposing defeaters:

If E is a defeasible reason for S to believe H , D is a [n opposing] defeater for this reason if and only if D is a defeater (for E as a reason for S to believe H) and D is a reason for S to believe $\neg H$.¹¹

The idea behind Pollock and Cruz's account of opposing defeaters can't be that opposing defeaters are defeaters that motivate us to lower our credence in H (i.e., to have a defeated credence that is lower than our evidential credence); as already discussed, undercutting defeaters do that too. Nor can it be that

¹¹Pollock and Cruz 1999, p. 37.

opposing defeaters leave us with a low credence in H ; if your prior credence in H was low, then a fairly strong undercutting defeater can do that too. Rather, a more plausible reading of their account is that opposing defeaters are reasons to lower our credence in H *even when we ignore E* .

In other words, a natural thought here (regardless of whether it is Pollock and Cruz's thought or not) is that the core difference between an opposing defeater and an undercutting defeater is that, since an undercutting defeater merely undermines (perhaps only partially) the evidential effect of E on H , an undercutting defeater would have no effect on H once E is removed from consideration. By contrast, an opposing defeater provides some general reason to think that H is false; thus, its effect on H isn't mediated by E . In other words, an undercutting defeater reduces our credence in H only on the assumption of E , while an opposing defeater reduces our credence in H regardless of whether we assume E or not. To distinguish the two types of defeaters, then, we need to take a look at the effect of the defeater on H when we haven't already taken E into account.

This motivates the following:

UNDERCUTTING2: D is an undercutting defeater for the evidence that the E provides for H just in case $[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) = p(H)]$.

OPPOSING3: D is an opposing defeater for the evidence that E provides for H if and only if $[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) < p(H)]$.

Again, each of UNDERCUTTING2's condition and OPPOSING3's condition entails DEFEATER1's condition, which is desirable. Here, that's because DEFEATER1's condition just is the first conjunct of each of UNDERCUTTING2's condition and OPPOSING3's condition. Moreover, notice that UNDERCUTTING2's condition and OPPOSING3's condition are (as UNDERCUTTING1's condition and OPPOSING2's condition were) mutually exclusive. I'll say more about this below.

It would be noteworthy if the condition specified in OPPOSING2' entailed the condition specified in OPPOSING3, given our background assumptions.¹² As already argued, it's plausible that OPPOSING2' specifies a sufficient condition for opposing defeat. And if OPPOSING3 is true and the condition specified in OPPOSING2' entails the condition specified in OPPOSING3, then it follows that OPPOSING2' is true too. This would be a *prima facie* mark in OPPOSING3's favor.

In fact, the condition specified in OPPOSING2' comes close to entailing the condition specified in OPPOSING3. I've already argued that OPPOSING2's condition entails DEFEATER1's condition, which is the first conjunct of OPPOSING3's condition. (OPPOSING2's condition says that $p(H|E \wedge D) < p(H)$, and we're assuming throughout this discussion that $p(H) < p(H|E)$. By transitivity, $p(H|E \wedge D) < p(H|E)$, which is the first conjunct of OPPOSING3's condition.) What about the second conjunct?

¹²Namely, that $p(H|E) > p(H)$.

It might seem as though OPPOSING2's condition does entail the second conjunct of OPPOSING3's condition. After all, OPPOSING2's condition says that $p(H|E \wedge D) < p(H)$ —i.e., that $E \wedge D$ lowers the agent's credence in H . But we're assuming that $p(H|E) > p(H)$ —i.e., that E alone raises the agent's credence in H . So if E alone raises the agent's credence in H , and the conjunction $E \wedge D$ lowers the agent's credence in H , isn't the only explanation for this that D alone must lower the agent's credence in H (and by more than E raises it)? After all, if D raised (or was neutral to) the agent's credence in H , how could the combined effect of E and D be to lower the agent's credence in H , given that E alone raises it?

The above argument, however, is invalid. The somewhat plausible-sounding principle

CONJUNCTIONS OF CONFIRMERS ARE CONFIRMERS (CCC): If A is evidence for H , and B is evidence for H , then $A \wedge B$ is evidence for H .

is false. Here's a counterexample: Suppose that someone is applying for a job in your department, and that you don't know which graduate program she's coming from. You know that Graduate Program X (hereafter, "X") produces above-average percentages of Metaphysics students, and also above-average percentages of Logic students. However, suppose that you also know that X never produces students who do both Metaphysics and Logic; students are forced to choose between these areas in their first year, and aren't permitted to work in both. However, this policy is unique to X; though they are somewhat rare, other programs do produce students who work in both Metaphysics and Logic. Suppose first that you find out only that the applicant does Metaphysics (it's left open when she works in other areas too). Since you know that X produces an above-average number of Metaphysics students, this is some evidence that the applicant comes from X. If, instead of finding out that she does Metaphysics, you had instead found out that she does Logic, that too would have been some evidence that she comes from X. But if you learn both that she does Metaphysics and that she does Logic, this is clearly conclusive evidence that she is not from X, since X doesn't produce anyone who does both Metaphysics and Logic. Thus, the fact that the applicant does Metaphysics is a confirmer that she's from X, and the fact that she does Logic is a confirmer, but the joint effect of these confirmers is to disconfirm. Hence CCC is false.

Suppose that, in this case, you learn first that the applicant does Metaphysics. As argued above, this is evidence that she's from X. When you learn that she also does Logic, you become certain that she's not from X. It's quite natural, then, to characterize the information that she does Logic as a defeater of your evidence that she's from X. But there's something a bit strange going on here. If you hadn't already learned that she does Metaphysics, then the information that she does Logic would have been evidence *for* the hypothesis that she comes from X. Thus, whether the information that she does Logic is a confirmer or a defeater of the hypothesis that she's from X seems to turn on the order in which I learn her specialties.

I don't know how to answer the question whether, in the case above, the information that the applicant does Logic is an undercutting defeater or an opposing defeater. I'm hesitant to classify it as an undercutting defeater, since (given that we also know that she does Metaphysics) the defeated credence is lower than the prior credence. And I'm hesitant to classify it as an opposing defeater, since it actually confirms H if we assume that we don't know that she does Metaphysics. I don't think that our pre-theoretic notions or intuitions are clear enough to deliver an unambiguous verdict here. I propose that we classify this case as involving a new type of defeater, which I call a bidirectional defeater, characterized as follows:

BIDIRECTIONAL1: D is a bidirectional defeater for the evidence that E provides for H if and only if $[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) > p(H)]$.

So where are we now? We have the following tripartite distinction:

UNDERCUTTING2: D is an undercutting defeater for the evidence that E provides for H just in case $[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) = p(H)]$.

OPPOSING3: D is an opposing defeater for the evidence that E provides for H if and only if $[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) < p(H)]$.

BIDIRECTIONAL1: D is a bidirectional defeater for the evidence that E provides for H if and only if $[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) > p(H)]$.

We retain

DEFEATER1: D is a defeater for the evidence that E provides for H just in case $p(H|E \wedge D) < p(H|E)$.

as a general characterization of defeaters, since the three conditions above individually entail it (and it clearly entails the disjunction of the three conditions).

Now let's see whether this gets the cases that we started with right.

Take the case where John's testimony that it's raining out is undercut by the information that John is an unreliable testifier about the weather. Does this case satisfy UNDERCUTTING2's condition? Clearly, $p(H|E \wedge D) < p(H|E)$ here. When I first heard John's testimony, I became more confident that it was raining out. When I learned that he was unreliable, I became less confident that it was raining out. Notice that this inequality holds regardless of whether the defeater is that John is highly unreliable or that he is only somewhat unreliable. Either way, it's plausible that the defeater induces a defeated credence lower than my evidential credence. What about the second conjunct? Before hearing John's testimony, I had my prior credence that it's raining out, $p(H)$. How would this credence be affected by learning only that John is unreliable? Intuitively, not at all. After all, I haven't heard John's testimony yet, so I don't even know what his testimony is, and I certainly haven't taken it into account yet. I might not have even heard of John, so the information that he tends to get weather reports

wrong shouldn't seem particularly relevant to me. Thus, my $p(H) = p(H|D)$, so both conjuncts of UNDERCUTTING2's condition are satisfied.

Now take our paradigm cases of opposing defeaters, where John's testimony that it's raining is defeated either by my own sunny experience outside or by Jill's testimony that it's not raining. Is OPPOSING3's condition satisfied? Again, it's fairly clear that $p(H|E \wedge D) < p(H|E)$; both of these defeaters give me reason to lower my credence in H on the assumption that I've already taken E into account. Moreover, it's fairly clear that the second conjunct of OPPOSING3's condition is also satisfied. Since each of these defeaters intuitively constitutes direct evidence against H , it should follow that acquiring the defeaters would force me to lower my credence in H , even if E hasn't already been taken into account. Even if I haven't heard John's testimony or don't even know who John is, still a sunny experience or Jill's testimony that it's not raining constitute evidence that it's not raining. Accordingly, $p(H|D) < p(H)$, so OPPOSING3's condition is satisfied.

4 Hybrids

I think that DEFEATER1, UNDERCUTTING2, OPPOSING3, and BIDIRECTIONAL1 do a fairly good job characterizing ordinary types of epistemic defeat. However, these principles do less well when we consider hybrid cases of defeat. Consider the following case, due to Jim Pryor:¹³

You suspect that ghosts might exist and you also suspect that elves might exist, though you've never seen either before. However, you're certain that it's not the case that both ghosts and elves exist. You know that, if there are ghosts, they have the following three properties: (a) They can pass through you without your noticing; (b) When they pass through you, they cause you to hallucinate an object¹⁴ in front of you; and (c) When they pass through you, they leave a chalk mark on your back.

Now, suppose that you have a visual experience as of an elf in front of you. Though certainly not conclusive (since someone could be playing a trick on you, or you could be hallucinating, etc.), this seems to be some evidence that there is an elf in front of you. Immediately afterwards, you look at your back in a mirror, and you see a chalk mark. Intuitively, two reactions seem appropriate here. First, you ought to (at least partially) discount the confirming effect that that you took your visual experience as of an elf to have on the proposition that there are elves. Since you've acquired some evidence (in the form of the chalk mark¹⁵) that the visual experience you just had was a hallucination, you ought

¹³Personal correspondence.

¹⁴Suppose that this is a random object. Ghosts aren't particularly likely to make you hallucinate elves or ghosts, though they sometimes may.

¹⁵Of course, chalk marks aren't always due to ghosts, so it's possible that your experience as of an elf was veridical, but that (say) you recently backed into a chalk board. Still, the point is that the chalk mark is some evidence that a ghost just passed through you, thus some evidence that your visual experience was a hallucination, and thus gives you some reason to discount that visual evidence. If you prefer, we could change the example so that ghosts leave

to regard that visual experience as a far less reliable indicator of the actual presence of elves than you previously thought. Second, it seems, you ought to decrease your credence that there are elves. After all, the chalk mark on your back is some evidence that there are ghosts; since you know that there aren't both ghosts and elves, any evidence that there are ghosts is also evidence that there are no elves.¹⁶

Of course, I'm not suggesting that you should do these two things in order, if that suggestion even makes any sense. More realistically, you just see the chalk mark on your back, and your credence does whatever it does, all in one step. But it does seem natural in this case to separate out two different components of your epistemic reaction to this information. On the one hand, the information that you have a chalk mark on your back seems to undercut your visual evidence that there are elves, and on the other, it seems to directly oppose the proposition that there are elves. Therefore, this looks to be a hybrid case of undercutting and opposing defeat.

The trouble here is that UNDERCUTTING2's condition and OPPOSING3's condition are incompatible, since UNDERCUTTING2's condition entails that $p(H|D) = p(H)$ and OPPOSING3's condition entails that $p(H|D) < p(H)$. Thus, if these principles are correct, then hybrid cases of undercutting and opposing defeat are impossible.

In the case under consideration, it's plausible that $p(H|D) < p(H)$; even before having a visual experience as of an elf, the information that there's a chalk mark on your back is some evidence against the proposition that there are elves. So, OPPOSING3 counts this as a case of opposing defeat. But that's only part of the story; intuitively, this case is also a case of undercutting defeat, and UNDERCUTTING2's condition isn't satisfied (indeed, necessarily, it can't be satisfied if OPPOSING3's condition is). If this is right, then UNDERCUTTING2's condition can't be necessary for undercutting defeat.

I think that this sort of case motivates reconsidering our story about undercutting defeat. Though I think that it is false, Pollock and Cruz's account is instructive here. Here's how they characterize undercutting defeaters:

If believing E is a defeasible reason for S to believe H , D is an *undercutting* defeater for this reason if and only if D is a defeater (for believing E as a reason for S to believe H) and D is a reason for S to doubt or deny that E would not be true unless H were true.¹⁷

This characterization is a bit odd. First, there are notorious problems with counterfactual analyses, and I suspect that it would be a relatively straightforward matter to generate problems to do with overdetermination, pre-emption, etc. in this case. But more importantly, suppose again that John tells me that it's

a special chalk mark on your back that can't occur in any other way.

¹⁶Actually, the principle that "if E confirms $H1$, and $H1$ entails $H2$, then E confirms $H2$ " is false. But it's easy to fill in the details of the case so that this particular application of the principle is OK. For further details, see my Kotzen ms.

¹⁷Pollock and Cruz 1999, p. 37.

raining out (E), and that I believe on that basis that it is raining out (H). According to Pollock and Cruz, D is an undercutting defeater of John's testimony (for me) if only if D gives me a reason to doubt or deny that: *John wouldn't say it's raining unless it were raining* (and D is a defeater). In some cases, this account works just fine: a paradigm undercutting defeater like information that there is no correlation whatsoever between John's weather reports and the actual weather would indeed give me a reason to deny that *John wouldn't say it's raining unless it were raining*.

However, suppose that I knew in advance that John is very unlikely to say that it's raining when it's not in fact raining. Now, suppose I learn that John is also very unlikely to say that it's raining when it *is* raining; in fact, we can even suppose that he's equally as unlikely to say that it's raining regardless of whether it's raining or not. This information should clearly be a defeater for John's testimony about the weather, since it entails that John is just as likely to say that it's raining when it is raining as when it's not. But this information does *not* give us any reason to doubt our deny that *John wouldn't say it's raining unless it were raining*; we knew all along that John is very unlikely to say that it's raining when it's not raining, and the information in question doesn't change our attitude about that.

How should we fix things up? We need to weaken UNDERCUTTING2's condition, since it's not necessary for undercutting defeat. And we need the weakened condition to be compatible with OPPOSING3 (or whatever condition we formulate for opposing defeat), so that we can get hybrid cases of undercutting and opposing defeat. Pollock and Cruz's characterization of undercutting defeat seemed to get at the idea that undercutting defeaters are such that, when they are assumed to be true, the evidential connection between E and H is interfered with; however, as argued above, their formulation of that idea is flawed.

Intuitively, in the ghost case above, the reason that the chalk mark serves as an undercutting defeater is that it's some evidence that your visual experience as of an elf was a hallucination. Thus, once you find out that there's a chalk mark on your back, your visual experience as of an elf no longer confirms the hypothesis that there are elves as much as that experience did before you knew that there was a chalk mark on your back. In other words, the degree of confirmation that E confers on H is lower once we assume D .

Let $dc(E, H, K)$ be a real-valued function of three variables E , H , and K , which quantifies the degree of confirmation that E confers on H relative to background information K . The above discussion motivates:

UNDERCUTTING3: D is an undercutting defeater for the evidence that E provides for H (relative to background information K) just in case $dc(E, H, K) > dc(E, H, K \wedge D)$.

In other words, an undercutting defeater is such that the amount of confirmation that the evidence confers on the hypothesis is lower when you assume the truth of the undercutting defeater than when you don't.

It's controversial how best to measure degree of confirmation.¹⁸ Here are a few candidates:

$$dc1(E, H, K) = p(H|E \wedge K) - p(H|K)$$

$$dc2(E, H, K) = \log \left(\frac{p(H|E \wedge K)}{p(H|K)} \right)^{19}$$

$$dc3(E, H, K) = \log \left(\frac{p(H|E \wedge K)/p(\neg H|E \wedge K)}{p(H|K)/p(\neg H|K)} \right) = \log \left(\frac{p(E|H \wedge K)}{p(E|\neg H \wedge K)} \right)$$

(The former is the log of the ‘‘Bayes Factor’’ and the latter is the log of the ‘‘likelihood ratio.’’)

For reasons that are beyond the scope of this paper to address, I like $dc3(E, H, K)$ as a measure of degree of confirmation.²⁰ But for my purposes here, I wish to remain agnostic about which measure of confirmation is best. Whichever account is right, we can plug that account into UNDERCUTTING3 to yield a precisified account of undercutting defeat.²¹

For example, if you also like $dc3(E, H, K)$, then UNDERCUTTING3 becomes:

UNDERCUTTING3*: D is an undercutting defeater for the evidence that E provides for H (relative to background information K) just in case

$$\log \left(\frac{p(H|E \wedge K)/p(H|\neg E \wedge K)}{p(H|K)/p(\neg H|K)} \right) > \log \left(\frac{p(H|E \wedge D \wedge K)/p(\neg H|E \wedge D \wedge K)}{p(H|D \wedge K)/p(\neg H|D \wedge K)} \right).$$

This looks complicated, but in fact it just expresses the simple idea that an undercutting defeater reduces the extent to which the evidence confirms the hypothesis. In what follows, I'll focus on UNDERCUTTING3* rather than UNDERCUTTING3, for the sake of concreteness. But everything that I say should apply equally well to any any other precisification of UNDERCUTTING3.

Two points are important here.

First, notice that UNDERCUTTING2's condition entails UNDERCUTTING3*'s condition, given the assumptions that I've been making. UNDERCUTTING2's condition entails that $p(H|D) = p(H)$, which entails that $p(\neg H|D) = p(\neg H)$, so the denominators of the expressions on both sides of the inequality can be

¹⁸See Fitelson 1999 and Eells and Fitelson 2000 for good surveys of various candidates.

¹⁹One purpose of taking the log of these quantities is so that the measure counts as a so-called ‘‘relevance measure,’’ where the measure is positive if E confirms H , negative if E disconfirms H , and 0 if E is neutral to H . Another purpose is to ensure scale-invariance. For our current purposes, the log can be ignored. Since log is a monotone increasing function, it will follow from the fact that $A > B$ that $\log A > \log B$ (and conversely). So if we want to compare two degrees of confirmation, all we need to do is to compare the argument of the log.

²⁰For some reasons to accept $dc2(E, H, K)$, see Milne 1996 (though see also Pollard 1999). For some reasons to accept $dc3(E, H, K)$, see Eells and Fitelson 2000.

²¹Of course, if there is no ‘‘one true measure’’ of degree of confirmation but rather just a plurality of different measures, then my account entails that there will be many different notions of undercutting defeat—one relative to each of the confirmation measures. But I think that this is precisely the right result; if there is no one privileged way to measure evidence, then I don't think that there can be one privileged way to measure undercutting of evidence either.

ignored. And we're supposing that $p(H|E \wedge D) < p(H)$, since we're supposing that D is a defeater, which entails that $p(\neg H|E \wedge D) > p(\neg H)$. It follows that UNDERCUTTING3*'s condition is satisfied. So, UNDERCUTTING3* entails that UNDERCUTTING2's condition is a sufficient condition for undercutting defeat, which is desirable. Moreover, this result doesn't essentially depend on my selection of $dc3$ as the measure of degree of confirmation. Any reasonable measure of degree of confirmation is going to have the result that, given the relevant assumptions, UNDERCUTTING2's condition entails UNDERCUTTING3's condition, suitably precisified. Take, for example, $dc1$. Assume that UNDERCUTTING2's condition holds. Since UNDERCUTTING2's condition entails that $p(H|D) = p(H)$, UNDERCUTTING3 will hold just in case $p(H|E) > p(H|E \wedge D)$. But that's just our condition on defeat, which we're supposing to hold (and which UNDERCUTTING2's condition entails anyway).²²

Second, and more importantly, notice that UNDERCUTTING3*'s condition and OPPOSING3's condition are compatible. A defeater D can simultaneously reduce the amount of confirmation that E confers on H and serve as evidence against H in the absence of E . Indeed, the information that there's a chalk mark on your back in the ghost example seems to do precisely this. Since the chalk mark is some evidence that you're hallucinating, it reduces the degree of confirmation that your visual experience as of an elf confers on the hypothesis that there are elves. And since the chalk mark is some evidence that there are ghosts (and since you know that there can't be both ghosts and elves), it's also direct evidence against the claim that there are elves.

Let E be your visual experience as of an elf, let H be the hypothesis that there are elves, and let D be the chalk mark on your back. Clearly, $p(H|E) > p(H)$ and $p(H|E \wedge D) < p(H|E)$. If we ignore the visual experience, the chalk mark is still evidence against the hypothesis that there are elves, so $p(H|D) < p(H)$, so OPPOSING3's condition is satisfied. It's a little trickier to see that UNDERCUTTING3*'s condition is satisfied, but it is. Let's fill in the details of the case a bit to see that. Suppose that, before having the visual experience, your credence that there are elves is .2 (so $p(H) = .2$) and your credence that there are ghosts is .2 (and your credence that both exist is 0). After having a visual experience as of an elf, your credence that there are elves goes up to .9 (so $p(H|E) = .9$). If you had just seen the chalk on your back without having the visual experience as of an elf, your credence that there are elves would have gone down to .1 (so $p(H|D) = .1$). Seeing the chalk on your back after having the visual experience as of an elf doesn't have as dramatic an effect (since the chalk mark might not have been caused by a ghost, and so doesn't guarantee that your experience as of an elf was hallucinatory); let's suppose that it pushes your credence that there are elves down to .15 (so $p(H|E \wedge D) = .15$). UNDERCUTTING3* will then hold just in case $\log\left(\frac{.9 \cdot .1}{.2 \cdot .8}\right) > \log\left(\frac{.15 \cdot .85}{.1 \cdot .9}\right)$, which will hold just in case $36 > 1.588$. So UNDERCUTTING3*'s condition holds. Clearly, this argument would have gone through even if I had used slightly different numbers.

²²This will hold for any relevance measure.

5 A Second Taxonomy

Now that we've abandoned UNDERCUTTING2 in favor of UNDERCUTTING3, should we modify OPPOSING3, BIDIRECTIONAL1, or DEFEATER1 as well? I see no reason yet to modify OPPOSING3's condition; as argued above, it is consistent with UNDERCUTTING3's condition, which is desired, and it seems to capture our intuitions about opposing defeat quite naturally.

But once we accept UNDERCUTTING3, it's easy to see that its condition doesn't entail DEFEATER1's condition. D can lower the extent to which E confirms H , and yet can fail to lower a rational agent's evidential credence (so $p(H|E \wedge D) \geq p(H|E)$). In other words, just as there are hybrid cases of undercutting and opposing defeat, there are also hybrid cases of undercutting defeat and *evidential support*.

To see such a case, all we need to do is make a minor adjustment to the ghost/elf case above. This time, assume that before you see the chalk mark on your back, you have a visual experience as of a ghost, rather than an elf. As before, after you have the visual experience, you then look at your back in a mirror and see a chalk mark.

Intuitively, just as in the first case, seeing the chalk mark (at least partially) undercuts your visual evidence (this time, that there are ghosts), since the chalk mark is evidence that the experience you just had as of a ghost was in fact a hallucination. However, simultaneously, seeing the chalk mark is evidence that there are ghosts, since you know that ghosts (if they exist) make chalk marks on people's backs when they pass through them.

Let H be the proposition that there are ghosts, let E be your experience as of a ghost in front of you, and let D be the chalk mark on your back. Clearly, $p(H|E) > p(H)$, since your experience as of a ghost is evidence that there are ghosts. It's also plausible that $p(H|D) > p(H)$, since just seeing a chalk mark on your back (even without having just seemed to see a ghost) is also evidence that there are ghosts.

The case as I presented it is perhaps a bit underdescribed, but we could imagine filling it out in such a way that $p(H|E \wedge D) > p(H|E)$.²³ To simplify a bit,²⁴ all we really need to suppose is that the chalk mark is better evidence for the existence of ghosts than your visual experience as of a ghost is; that way, even if the evidential impact of the visual experience as of a ghost is completely undercut, the evidential impact of the chalk mark more than makes up for this.

Still, even if it's true that $p(H|E \wedge D) > p(H|E)$, it doesn't follow that UNDERCUTTING3's condition is violated. In the case under consideration, it's still plausible that the information that there's a chalk mark on your back lessens

²³Of course, it would be enough to violate DEFEATER1 if $p(H|E \wedge D) = p(H|E)$ for some defeater D of E , but I think it's clearer to focus on cases where $p(H|E \wedge D) > p(H|E)$.

²⁴The reason that this is a bit of a simplification comes from my discussion above of the CCC principle. Just because E disconfirms H by a little and D confirms H by a lot, it doesn't follow that $E \wedge D$ confirms H ; since CCC is false, it wouldn't even follow from that fact that E and D both confirm H that $E \wedge D$ confirms H . But if we assume that the relevant effects here are reasonably independent of each other (as I'm free to do—after all, I only need one counterexample), what I say in the main text will do.

the degree to which the visual experience as of a ghost confirms the hypothesis that there's a ghost. This isn't at all threatened by the fact that $p(H|E \wedge D) > p(H|E)$. If this were a case where $p(H|D) = p(H)$, then it would follow from the fact that $p(H|E \wedge D) > p(H|E)$ that UNDERCUTTING3's condition is violated. But, as already remarked, the chalk mark is direct evidence that there are ghosts, so $p(H|D) > p(H)$. In short, even though $E \wedge D$ confirms H more than E alone does, the fact that $p(H|D) > p(H)$ leaves it open that E might confirm H less on the assumption of D than it does without assuming D .

Again, let's try to fill in the details of the case a bit in order to test it out. Suppose as before that, before having the visual experience, your credence that there are elves is .2, your credence that there are ghosts is .2 (so $p(H) = .2$), and that your credence that there are both is 0. After having the visual experience as of a ghost, your credence that there are ghosts goes up to .95²⁵ (so $p(H|E) = .95$). If you had just seen the chalk mark on your back without having the visual experience as of a ghost, your credence that there are ghosts would have gone up—say, to .97 (so $p(H|D) = .97$). Seeing the chalk on your back after having the visual experience as of a ghost should confirm H to an even higher degree (since the chalk mark might not have been caused by a ghost, and so doesn't guarantee that your experience as of a ghost was hallucinatory); let's suppose that it pushes your credence that there are ghosts up to .98 (so $p(H|E \wedge D) = .98$). UNDERCUTTING3*'s condition will then hold just in case $\left(\frac{.95/.05}{.2/.8}\right) > \left(\frac{.98/.02}{.97/.03}\right)$, which will hold just in case $76 > 1.515$. So UNDERCUTTING3*'s condition holds.

So, in some cases D is an undercutting defeater for the evidence that E provides for H , and yet $p(H|E \wedge D) \geq p(H|E)$, so DEFEATER1 is false.

What about BIDIRECTIONAL1? Again, here the issue is partly stipulative; as far as I know, "bidirectional" defeaters aren't discussed anywhere in the literature on epistemic defeat, and I certainly don't think that we have clear intuitions about when a given D is serving as a bidirectional defeater. But it is worthy of note that BIDIRECTIONAL1 was a natural characterization of a third kind of defeat in the context of the acceptance of all of DEFEATER1, UNDERCUTTING2, and OPPOSING3. And once we abandon UNDERCUTTING2, it's no longer clear that BIDIRECTIONAL1 is the right way to characterize this third kind of defeat. Moreover, it turns out that it's straightforward to construct cases quite similar to the job applicant case above where $p(H|D) \leq p(D)$ (say, by changing the case so that X produces an *average* number of Logic students). I propose that the crucial feature of this "third kind" of defeat isn't, as BIDIRECTIONAL1 claims, that D goes from confirming H to disconfirming H when we assume E , but rather that E goes from confirming H to disconfirming H when we assume D . This happens in the original job applicant case, where the fact that the candidate does Metaphysics goes from confirming that she's from X to disconfirming

²⁵I assumed that the evidential credence in the existence of elves was .9; the reason it's higher here is that a ghost-induced hallucination of a ghost entails that there are ghosts, whereas a ghost-induced hallucination of an elf does not entail that there are elves (in fact, it entails that there are not elves).

that she's from X when we assume that she also does Logic. And this also happens in a modified job applicant case where X produces an average, rather than an above average, number of Logic students. So I propose to characterize bidirectional defeat as follows:

BIDIRECTIONAL2: D is a bidirectional defeater for the evidence that E provides for H just in case $p(H|D) > p(H|E \wedge D)$.

Since we're assuming that $p(H|E) > p(H)$, the satisfaction of BIDIRECTIONAL2's condition entails that E 's positive relevance to H turns into negative relevance to H once D is assumed as background information.

As it turns out, BIDIRECTIONAL2's condition entails UNDERCUTTING3's condition,²⁶ so this taxonomy has it that all bidirectional defeaters are undercutting defeaters (but not vice versa). Again, bidirectional defeat is a partly stipulative matter, so I'm perfectly happy with this result. On my taxonomy, undercutting defeaters lower the amount that E confirms H when they're assumed as background. Bidirectional defeaters, then, are just the special subclass of undercutting defeaters that lower the amount that E confirms H so much that that quantity becomes negative; E *disconfirms* H when a bidirectional defeater is assumed as background.

How should we amend DEFEATER1 as a general characterization of defeat? I see no obvious or natural characteristic that UNDERCUTTING3's condition and OPPOSING3's condition have in common. So perhaps undercutting and opposing defeat are more dissimilar than they at first seemed. Both undercutting and opposing defeaters "count against" a hypothesis in *some* sense, but there might not be any natural way to formalize this sense. So I think that all that we can usefully say is that D is a defeater of the evidence that E provides for H just in case it's either an opposing or an undercutting defeater (we don't need to include bidirectional defeat here, since all bidirectional defeaters are undercutting defeaters, on my account). Thus:

DEFEATER2: D is a defeater for the evidence that E provides for H (relative to background information K) just in case $\{[p(H|E \wedge D) < p(H|E)] \wedge [p(H|D) < p(H)] \vee [dc(E, H, K) > dc(E, H, K \wedge D)]\}$.

One might worry that, just as there are hybrids of *undercutting* defeat and evidential support, so too might there be hybrids of *opposing* defeat and evidential support, which would make problems for DEFEATER2 and OPPOSING3. But I don't think that there really are such hybrids. Take a putative hybrid of opposing defeat and evidential support, such as the information that John says it's raining but Mary says it's not. We're inclined to count this is a hybrid because it's a conjunction of a proposition that supports the hypothesis that it's raining (John says it's raining) and a proposition that opposes that hypothesis (Mary says it's not raining). But if that's a sufficient condition for such a hybrid, then we will end up with far too many hybrids. Consider the information

²⁶For all relevance measures.

that John said only that it's raining. That's just plain evidential support for the proposition that it's raining, and certainly not a hybrid opposing *defeater* for rain. But *John said only that it's raining* is a conjunction of *John said it's raining or John said that the barometric pressure is falling* and *John didn't say that the barometric pressure is falling*. Since the former conjunct is evidential support for rain, and the latter conjunct is opposing defeat for rain, we end up with the result that *John said only that it's raining* is a hybrid opposing defeater for rain, which is unacceptable.²⁷ I take this to show that there can't be a non-trivial account of hybrids of opposing defeat and evidential support. Thus, I think it's welcome result that OPPOSING3 prevents *D* from providing net support for *H*, and that DEFEATER2 doesn't allow for pure²⁸ hybrids of evidential support and opposing defeat.

6 Redundancy and Undercutting

Unfortunately, there is a serious problem with UNDERCUTTING3. UNDERCUTTING3 says that *D* is an undercutting defeater just in case it lowers the degree to which *E* confirms *H*. The problem is not with the necessity of this condition; as far as I can tell, everything that we would intuitively count as an undercutting defeater does satisfy the condition. Rather, the problem is with the sufficiency; there is a broad class of propositions that do lower the degree to which *E* confirms *H*, and yet don't seem to deserve to be called undercutting defeaters.

The class that I have in mind is the class of propositions that are somehow *redundant* of *E*. Suppose, for example, that I know that my friend Rex is a fairly reliable predictor of the weather. Each night, Rex communicates to me his predictions for the weather the following day. Since Rex wants to make sure that I receive his predictions, he *both* sends me an email and *also* leaves me a voicemail with his predictions (the predictions are always identical in the email and the voicemail).

One night, I receive Rex's email with a prediction of rain the following day. Since I take Rex to be reliable about these matters, the email is evidence that it is going to rain. Next, I listen to Rex's voicemail with (of course) the exact same prediction. Intuitively, the voicemail is *redundant* or perhaps *irrelevant* once I've already read the email, but I don't think it's natural to call it an undercutting defeater of the evidence that the email provided for rain tomorrow. After all, the voicemail doesn't cast any doubt on the reliability of the emailed report; it's not like the voicemail said something like "The email that I sent you earlier was mistaken."

But UNDERCUTTING3 entails that the voicemail is an undercutting defeater for the evidence that the email provides for rain. When we ignore the voicemail, the email is good evidence for rain, since I know that Rex is reliable about the

²⁷I'm assuming here that if *P* and *Q* are logically equivalent, then *P* is a hybrid opposing defeater iff *Q* is a hybrid opposing defeater, but I think that's overwhelmingly plausible here.

²⁸By "pure," I mean to refer to defeaters that aren't *also* hybrid undercutting defeaters

weather. Suppose that the email justifies a credence in rain of .9. But when we assume the voicemail as background information, the email is no longer good evidence for rain; if I've already heard the voicemail and thus already increased my credence in rain to .9, the email is completely redundant and thus doesn't do anything to further confirm the hypothesis that it's going to rain. Let E be the email, let H be the hypothesis that it's going to rain tomorrow, and let D be the voicemail. Then, $dc(E, H, K)$ is high, since the email is good evidence for rain relative to my background information K . But $dc(E, H, K \wedge D)$ is 0; once I've already taken D into consideration, E does nothing further to confirm H .²⁹ So UNDERCUTTING3's condition is easily satisfied; if UNDERCUTTING3 is true, this entails that the voicemail is an undercutting defeater for the evidence for rain provided by the email, which is counterintuitive.

What should we say about this sort of case? One option, of course, is to "bite the bullet" and accept that redundant evidence does undercut, since it reduces the amount of confirmation that the original evidence confers on the hypothesis. But I think we can do better.

Here's one possible fix. UNDERCUTTING3 entails that we have undercutting defeat whenever $dc(E, H, K) > dc(E, H, K \wedge D)$. In cases like the Rex case, the problem was that the only reason that D lowered the extent to which E confirmed H was that D was (entirely) redundant of E , and hence confirmed H to the same extent that E itself did. So one natural idea is that undercutting defeaters are such that the amount that E confirms H is less than the *sum* of the amount that E confirms H when D is assumed *and* the amount that D *itself* confirms H . Thus:

UNDERCUTTING4: D is an undercutting defeater for the evidence that E provides for H (relative to background information K) just in case $dc(E, H, K) > dc(E, H, K \wedge D) + dc(D, H, K)$.

In cases like the original rain case where my learning that John is an unreliable testifier about the weather undercuts his testimony that it's raining, the third term $dc(D, H, K)$ equals 0, since the information that John is unreliable about the weather neither confirms nor disconfirms the proposition that it's raining. But in cases like the Rex case, where D is completely redundant of E , the second term $dc(E, H, K \wedge D)$ equals 0, and the first term $dc(E, H, K)$ equals the third term $dc(D, H, K)$ (since D alone has the same confirmatory effect on H as E alone does), so the inequality in UNDERCUTTING4 fails. More concretely, $dc(E, H, K \wedge D) = 0$, since Rex's email does nothing to confirm rain if I've already taken the voicemail into account. And $dc(E, H, K) = dc(D, H, K)$, since the email and the voicemail have the same confirmatory effect on the hypothesis of rain when they're considered alone. So $dc(E, H, K) = dc(E, H, K \wedge D) + dc(D, H, K)$, so UNDERCUTTING4's condition fails.

The problem with UNDERCUTTING4, however, is that it stumbles with cases which involve hybrids of undercutting defeat and evidential support, such as the

²⁹If dc is a relevance measure (which I've been assuming it is), then $dc(E, H, K \wedge D) = 0$ when $p(H|D) = p(H|E \wedge D)$. But it's clear here that $p(H|D) = p(H|E \wedge D) = .9$.

second ghost/elf case from above. In that case, you have a visual experience as of a ghost, which is some evidence that there are ghosts. Then, when you see a chalk mark on your back, that both undercuts your visual experience, and also serves as some independent evidence that there are ghosts. As before, let E be your visual experience as of a ghost, let H be the proposition that there are ghosts, and let D be the chalk mark on your back. Suppose, for simplicity, that visual experiences as of ghosts are just as good evidence for ghosts as chalk marks on your back are, so that $dc(E, H, K) = dc(D, H, K)$. As long as the visual experience as of a ghost is still *some* evidence that there are ghosts even after the chalk mark has been taken into account, then $dc(E, H, K \wedge D) > 0$, so UNDERCUTTING4's condition fails. But this isn't what we want. In the case above, the chalk mark *does* undercut the visual evidence for ghosts; it's just that, since the chalk mark is a hybrid of undercutting defeat and evidential support, it is able to violate the inequality in UNDERCUTTING4.

What we need is a way to distinguish between, on the one hand, cases like the Rex case, where some D lowers the amount of confirmation that E confers on H only because D is redundant of E , from, on the other hand, cases like the ghost case, where D does genuinely undercut the evidential force of E (i.e., *not* for reasons of redundancy), but this effect is "hidden" by the fact that D also provides direct evidential support for H .

What this seems to require is that, instead of adding $dc(D, H, K)$ to the right-hand side of the inequality in UNDERCUTTING3 (yielding UNDERCUTTING4), we instead add some function f of $dc(E, H, K)$ (and perhaps other parameters as well) to the right-hand side, which quantifies the extent to which D confirms H *in virtue of the sort of redundancy effects that made problems for* UNDERCUTTING3. Thus:

UNDERCUTTING5: D is an undercutting defeater for the evidence that E provides for H (relative to background information K) just in case $dc(E, H, K) > dc(E, H, K \wedge D) + f(dc(E, H, K))$.

What are the constraints on this function $f(dc(E, H, K))$? First, f should monotonically increase with the extent of redundancy between E and D ; the more redundancy there is between E and D , the larger that the value of $f(dc(E, H, K))$ should be. In the limiting case where they are completely redundant, $f(dc(E, H, K)) = dc(E, H, K)$. In the limiting case where E and D are probabilistically independent (i.e., where $p(E|D) = p(E)$ or, equivalently, where $p(D|E) = p(D)$), $f(dc(E, H, K)) = 0$, so that we get UNDERCUTTING3 when there are no redundancy effects.

One natural idea is to use some function of $p(E|D)$ to characterize the extent of the redundancy between E and D —more specifically, the extent to which D is redundant of E . The simplest linear function of $p(E|D)$ and $dc(E, H, K)$ that has the above properties is $\frac{p(E|D)-p(E)}{1-p(E)} \times dc(E, H, K)$. (The left multiplicand of this function can be understood to quantify the fraction of the distance between $p(E)$ and 1 that D increases E 's probability.)

Plugging this function into UNDERCUTTING5, we get:

UNDERCUTTING5*: D is an undercutting defeater for the evidence that E provides for H (relative to background information K) just in case $dc(E, H, K) > dc(E, H, K \wedge D) + \frac{p(E|D) - p(E)}{1 - p(E)} \times dc(E, H, K)$.

In the original rain case where John’s testimony that it’s raining and my learning that John is an unreliable testifier about the weather are probabilistically independent, $p(E|D) = p(E)$, so the third term in UNDERCUTTING5* goes to 0, and UNDERCUTTING5* is equivalent to UNDERCUTTING3. In the Rex case, the voicemail saying that it’s going to rain tomorrow makes the email saying that it’s going to rain tomorrow certain, so $p(E|D) = 1$, so the third term is equal to $dc(E, H, K)$. The second term, $dc(E, H, K \wedge D)$, equals 0, since the email doesn’t confirm rain on the assumption of the voicemail as background information (that’s the whole reason that we were forced to abandon UNDERCUTTING3). So UNDERCUTTING5* entails that there is undercutting defeat just in case $dc(E, H, K) > dc(E, H, K)$, which is always false, so UNDERCUTTING5* entails that there is no undercutting in the Rex case, as desired.

Unfortunately, UNDERCUTTING5* stumbles with certain cases where D has both a redundancy effect on E and also an undercutting effect. For example, if we let $D = E \wedge U$, where U is some genuine undercutting defeater, then UNDERCUTTING5* wrongly entails that D doesn’t undercut. Since $D = E \wedge U$, $p(E|D) = 1$, so the third term in UNDERCUTTING5* becomes $dc(E, H, K)$. And because D entails E , $dc(E, H, K \wedge D) = 0$. So UNDERCUTTING5*’s condition again becomes $dc(E, H, K) > dc(E, H, K)$, which is never satisfied, so UNDERCUTTING5* entails that there is no undercutting defeat. But, intuitively, $D = E \wedge U$ is an undercutting defeater here, since it includes some information (i.e., U), which is by hypothesis undercutting.

Intuitively, the reason that $E \wedge U$ is an undercutting defeater is that it entails U , which is by hypothesis an undercutting defeater. So if I were to come to know that $E \wedge U$ obtains, then (by closure) I’d thereby come to know that U obtains. So maybe we can say this: UNDERCUTTING5* doesn’t characterize *all* undercutting defeaters, but rather characterizes only “simple” undercutting defeaters. Then, we can say that “complex” undercutting defeaters are propositions that entail simple undercutting defeaters.³⁰ Undercutting defeaters, then, are just those propositions that are either simple or complex undercutting defeaters.

The trouble here is that the account is too inclusive; lots of things that don’t count intuitively as undercutting defeaters will count as (complex) undercutting defeaters on this account. For instance, let’s return to our rain case, and consider the proposition (call this O) that the number of grains of sand on Daytona Beach is odd. O is completely irrelevant to John’s testimony about the weather, and hence shouldn’t be counted as an undercutting defeater of that testimony. But O entails $O \vee U$, where U is the proposition that John is unreliable about the weather. And it’s very plausible that $O \vee U$ is an undercutting defeater; if I were to learn $O \vee U$ (without inferring it from O), that would at least partially

³⁰We don’t need to allow for complex undercutting defeaters that entail other complex undercutting defeaters but not any simple undercutting defeaters, since entailment is transitive.

undercut John’s testimony. So it’s not the case that any proposition that entails an undercutting defeater is itself an undercutting defeater.

One plausible explanation for why O doesn’t strike us as an undercutting defeater in the case above even though O entails the undercutting defeater $O \vee U$ is that O also entails $O \vee \neg U$. And since $O \vee \neg U$ has the exact opposite of $O \vee U$ ’s undercutting effect (let’s call this effect “bolstering”), O itself has no net undercutting effect. By contrast, $E \wedge U$ entails the undercutting defeater U , but fails to entail $\neg U$.

This suggests that complex undercutting defeaters are those propositions that entail simple undercutting defeaters *but don’t entail any bolsterers*. But there are two problems here. First, $E \wedge U$, a complex undercutting defeater, entails $(E \wedge U) \vee B$, where B is some genuine bolsterer. And if B is a strong bolsterer whereas U is a comparatively weak undercutter, then it’s possible that $(E \wedge U) \vee B$ will be a net bolsterer. Second, if U is a strong undercutter and B is a comparatively weak bolsterer, we still want $U \wedge B$ to count as an undercutter (since it entails U , and U is a stronger undercutter than B is a bolsterer), even though it entails B , which is a simple bolsterer.

The natural thing to say here is that complex undercutting defeaters are those propositions that entail simple undercutting defeaters but don’t entail any bolsterers that are at least as strong. Or, more precisely, the combined strength of the simple defeaters that a complex defeater entails is higher than the combined strength of the simple bolsterers that that complex defeater entails.

And fortunately, UNDERCUTTING5* gives us a natural way to quantify the *strength* of a defeater, since we can just let the strength of an undercutting defeater be the *difference* between the two sides of the inequality, when that inequality is satisfied. Thus:

UNDERCUTTING5* STRENGTH: When D is an undercutting defeater for the evidence that E provides for H (relative to background information K) according to UNDERCUTTING5*, D ’s strength as an undercutting defeater is $dc(E, H, K) - dc(E, H, K \wedge D) - \frac{p(E|D) - p(E)}{1 - p(E)} \times dc(E, H, K)$.

Similarly, we can characterize bolsterers as follows:

BOLSTERING1: B is a bolsterer for the evidence that E provides for H (relative to background information K) just in case $dc(E, H, K) < dc(E, H, K \wedge B) + \frac{p(E|B) - p(E)}{1 - p(E)} \times dc(E, H, K)$.

And, just as with UNDERCUTTING5*, we can let the strength of the bolsterer be the difference between the two sides of the inequality in BOLSTERING1, when that inequality is satisfied:

BOLSTERING1 STRENGTH: When B is a bolsterer for the evidence that E provides for H (relative to background information K) according to BOLSTERING1, B ’s strength as a bolsterer is $\frac{p(E|B) - p(E)}{1 - p(E)} \times dc(E, H, K) + dc(E, H, K \wedge B) - dc(E, H, K)$.

We can then characterize undercutting defeaters as follows:

UNDERCUTTING6: D is a simple undercutting defeater for the evidence that E provides for H (relative to background information K) just in case $dc(E, H, K) > dc(E, H, K \wedge D) + \frac{p(E|D) - p(E)}{1 - p(E)} \times dc(E, H, K)$.

D is a complex undercutting defeater for the evidence that E provides for H (relative to background information K) just in case the combined strength of the simple defeaters that D entails is higher than the combined strength of the simple bolsters that D entails, as measured by UNDERCUTTING5* STRENGTH and BOLSTERING1 STRENGTH (and D is not a simple undercutting defeater of the evidence that E provides for H).

We can then, of course, make the appropriate changes to DEFEATER2.

7 Conclusion

In this paper, I have tried to give a framework for thinking through how notions of evidential defeat that have been deployed mostly in the context of binary belief should be generalized to partial belief contexts. Part of my goal has been taxonomic, but I don't think that the issue here is just the terminological one of whether this or that piece of information should deserve the title of "undercutting" or "opposing" defeater. Rather, I think that these are real notions that play important roles in epistemology, and I hope that I've made some progress on tracing the contours of those notions. My approach has been unabashedly Bayesian, and though I share the skepticism of many philosophers about how much of our epistemic lives can be modeled in standard Bayesian terms, I do think that Bayesianism is the best thing going with respect to nondeductive inference and that it is useful to see just how far we can push the Bayesian program. To the extent that the account above falls short, I hope that its shortcomings will give us some useful clues about which Bayesian or non-Bayesian approach we should take toward evidential defeat in the future.

Matthew Kotzen
New York University
Department of Philosophy
5 Washington Place, Office 408
New York, NY 10003
matthew.kotzen@nyu.edu

References

- [Achinstein 1994] Achinstein, P. (1994). "Explanation v. Prediction: Which Carries More Weight?," in Hull, Forbes, and Burian (eds) 1994, *Proceedings of the Philosophy of Science Association* Vol. 2. East Lansing, Mich.: Philosophy of Science Association, pp. 156–64.
- [Audi 1993] Audi, R. (1993). *The Structure of Justification*. Cambridge: Cambridge University Press.
- [Bonini, Crupi, Osherson, and Tentori 2005] Bonini, N., Crupi, V., Osherson, D., and Tentori, K. (2005). "Comparison of Confirmation Measures," *Cognition* 103, pp. 107–119.
- [Bergmann 2005] Bergmann, M. (2005). "Defeaters and Higher-Level Requirements," *The Philosophical Quarterly* 55, pp. 19–436.
- [Chisholm 1989] Chisholm, R. (1989). *Theory of Knowledge, 3rd edition*. Englewood Cliffs: Prentice-Hall.
- [Collins 1994] Collins, R. (1994). "Against the Epistemic Value of Prediction over Accommodation," *Noûs* 28, pp. 210–24.
- [Earman 1992] Earman, J. (1992). *Bayes or Bust?* Cambridge: MIT Press.
- [Fitelson 1999] Fitelson, B. (1999). "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity," *Philosophy of Science* 66, S362–S378.
- [Eells and Fitelson 2000] Eells, E. and Fitelson, B. (2000). "Comments and Criticism: Measuring Confirmation and Evidence," *The Journal of Philosophy* 97, pp. 663–672.
- [Harker 2006] Harker, D. (2006). "Accommodation and Prediction: The Case of the Persistent Head," *British Journal for the Philosophy of Science* 57(2), pp. 309–321.
- [Harman 1988] Harman, G. (1988). *Change in View: Principles of Reasoning*. Cambridge: MIT Press.
- [Harman 1999] Harman, G. (1999). "Rationality," in his *Reasoning, Meaning, and Mind*, pp. 9–45. Cambridge: MIT Press.
- [Horwich 1982] Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge University Press.
- [Howson and Urbach 1993] Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach, 2nd Edition*. Chicago: Open Court.
- [Kotzen ms] Kotzen, M. (manuscript). "Dragging and Confirming." Chapter of Ph.D. Thesis.

References

- [Kvanvig 2007] Kvanvig, J. (2007). "Two Approaches to Epistemic Defeat," *Philosophy in Focus*, ed. Deane-Peter Baker. Cambridge: Cambridge University Press.
- [Maher 1988] Maher, P. (1988). "Prediction, Accommodation, and the Logic of Discovery" in Fine and Leplin (eds) 1988, *Proceedings of the Philosophy of Science Association Vol. 1*. East Lansing Mich.: Philosophy of Science Association.
- [Milne 1996] Milne, P. (1996). " $\log[P(h/eb)/P(h/b)]$ Is the One True Measure of Confirmation," *Philosophy of Science* 63, pp. 21–26.
- [Plantinga 1993 a] Plantinga, A. (1993). *Warrant and Proper Function*. Oxford: Oxford University Press.
- [Plantinga 1993 b] Plantinga, A. (1993). *Warrant: The Current Debate*. Oxford: Oxford University Press.
- [Pollock 1987] Pollock, J. (1987). "Defeasible Reasoning," *Cognitive Science* 11, pp. 481–518.
- [Pollock 1995] Pollock, J. (1995). *Cognitive Carpentry: A Blueprint For How To Build A Person*. Cambridge: MIT Press.
- [Pollock and Cruz 1999] Pollock, J. and Cruz, J. (1999). *Contemporary Theories of Knowledge, 2nd edition*. Lanham: Rowman and Littlefield.
- [Pollard 1999] Pollard, S. (1999). "Milne's Measure of Confirmation," *Analysis* 59, pp. 335–338.
- [Pryor 2000] Pryor, J. (2000). "The Skeptic and the Dogmatist," *Noûs* 2000, pp. 517–49.
- [Pryor 2004] Pryor, J. (2004). "What's Wrong with Moore's Argument?" *Philosophical Issues* 14(1), pp. 349–378.
- [Pryor ms a] Pryor, J. (ms). "Is Moore's Argument an Example of Transmission Failure?," available at <http://www.jimpryor.net/research/papers/Moore2001.pdf>.
- [Pryor ms b] Pryor, J. (ms). "Uncertainty and Undermining," available at <http://www.jimpryor.net/research/papers/Uncertainty.pdf>.
- [Schlesinger 1987] Schlesinger, G. (1987). "Accommodation and Prediction," *Australasian Journal of Philosophy* 65, pp. 28–42.
- [Silins 2005] Silins, N. (2005). "Deception and Evidence," *Philosophical Perspectives* 2005, pp. 375–404.
- [van Fraassen 1980] van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

References

- [Weatherson ms] Weatherson, B. (ms). "The Bayesian and the Dogmatist."
Available at SSRN: <http://ssrn.com/abstract=997944>.
- [White 2000] White, R. (2000). "Fine-Tuning and Multiple Universes," *Noûs* 34,
pp. 260–76.
- [White 2003] White, R. (2003). "The Epistemic Advantage of Prediction Over
Accommodation," *Mind* 112(448), pp. 653–683.
- [White 2006] White, R. (2006). "Problems for Dogmatism," *Philosophical Stud-*
ies 131, pp. 525–557.