Reliabilism and the Problem of Defeaters

Prof. Dr. Thomas Grundmann
Philosophisches Seminar
Universität zu Köln
Albertus Magnus Platz
50923 Köln
E-mail:
thomas.grundmann@uni-koeln.de

4.454 words

Reliabilism and the Problem of Defeaters

It is widely assumed that justification (and on some accounts even warrant[1]) is defeasible by counterevidence. If an epistemic agent is justified in believing that p at time t and if at time t' she acquires either evidence for the falsity of p (a rebutting defeater) or evidence for the unreliability of the source of her belief that p (an undercutting defeater), then the belief's justification is removed at t'. In short, defeaters are evidence which removes justification. In what follows, I will first explain why the existence of defeaters seems to be incompatible with simple reliabilism. Then I will try out different reliabilist strategies to deal with this problem, among them Goldman's (1979) solution.

Let us first consider a typical example of a *rebutting defeater*: David sees at some distance what he takes to be a sheep and thus forms the belief that there is a sheep in the field. He knows that Frank is the owner of the field. On the next day, Frank tells David that there has never been a sheep in that field, although Frank owns a dog that looks like a sheep from the distance and often strolls around in the field. David thereby acquires a rebutting defeater for his belief that there was a sheep in the field. In this case, what David is told by Frank is incompatible with the truth of what he believes. If David holds on to his belief it becomes unjustified. Thus, it would be epistemically appropriate for David to believe that there was a dog, not a sheep in the field.

Consider now the following example of an *undercutting defeater***:** Betty knows that she has taken a drug which has a 50 % chance of causing hallucinations. Suddenly, she happens to have a completely unexpected experience. While she is on a lonesome hiking-trip, it suddenly seems to her as if the ground is shaking. On the basis of this impression she believes that she just experienced an earthquake. She is not justified in her belief. Her knowledge of the drug's side-effects undermines her experiential reason for believing that an earthquake just occurred. It would be epistemically appropriate for her to withhold her belief on the matter.

Why do reliabilism and defeaters seem to be incompatible? According to simple reliabilism, S is justified in believing that p at t if and only if S's belief at t is based on a reliable belief-

[1] See Plantinga 2000, p. 359.

producing mechanism. In effect, simple reliabilism claims that reliable processes are necessary *and sufficient* for a belief to be justified. Now recall the examples of defeaters given above. In both cases – the sheep case as well as the drug case – it seems possible that the agent's belief was produced by a reliable process. Consider the following true story about the sheep case: Unknown to Frank, there was one of his sheep in the field, but not his dog. And if his dog had been in the field, David would have been able to distinguish it from a sheep. Hence, David's belief that a sheep was in the field was reliably produced. According to simple reliabilism, his belief would be classified as justified. But this can't be true, since intuitively its justification was defeated by what David was told about the situation. Being reliably produced is thus not sufficient for being justified. And this contradicts simple reliabilism. Consider next the drug case. Let us assume that Betty is resistant to the hallucinatory side-effects of the drug, though she does not know about it. Moreover, she really experienced an earthquake on her hiking-tour. Hence, her belief that she was facing an earthquake was reliably produced. Again, given simple reliabilism, her belief would count as justified. But intuitively it isn't justified, because its justification is removed by Betty's knowledge about the general side-effects of the drug. In this case, the defeater may be misleading, but it successfully neutralizes the justificatory quality of Betty's belief nonetheless. So, contrary to what simple reliabilism claims, being reliably produced is again not sufficient for being justified. To put the problem in a nutshell: Simple reliabilism is incompatible with the widely acknowledged defeasibility of justification. Whereas simple reliabilism implies that reliably produced belief is *sufficient* for justification, the defeasibility of justification implies that reliable production is *not sufficient* for justification.

How should the reliabilist respond to this problem of compatibility? BonJour (1980, 1985) recommended giving up on reliabilism. But there are certainly more promising strategies available to the reliabilist. Firstly, she could insist that simple reliabilism and defeasibility of justification are in fact *compatible* and only appear to be incompatible on first glance. This would be the strategy of *conservative reliabilism*. Secondly, the reliabilist could *bite the bullet* and simply deny the existence of defeaters in general. Thirdly, the reliabilist could try to *revise* her position trying to integrate defeaters into reliabilism. Call this *revisionary reliabilism.*

What are the prospects of *conservative reliabilism?* The reliabilist might argue that, according to simple reliabilism, the justificatory status of a belief depends on the reliability of the whole

process responsible for entertaining it. Very often, the relevant process is the originating cause of the belief. But this need not always be the case. It also might happen that a belief was acquired in a certain way and is now sustained by a completely different process. In such cases, the current justificatory status of the belief depends completely on the reliability of the *sustaining* process. In the sheep case, David's belief was originally acquired by employing perception. When David is later informed that there was no sheep in the field, but continues to believe the contrary, then the relevant belief-sustaining process has changed. Ignoring the available counterevidence is at least part of the new belief-sustaining process. Now, the proponent of simple reliabilism might claim the following about David's cognitive situation: whereas the original perceptual process is reliable, it is plausible to assume that believing p in the face of ignored counterevidence is unreliable on the whole. Therefore, it might seem as if simple reliabilism was able to explain why a reliably acquired belief becomes unjustified when counterevidence is being ignored.

I am not that optimistic about maintaining conservative reliabilism. In fact, I believe that such a conservatism is a dead end. Here is the most severe objection to this view: it relies on the general assumption that *ignoring the available counterevidence* is an essential part of the cognitive process sustaining the belief. If such ignorance were an essential part of the sustaining process, then it would be quite reasonable to expect that this type of process often leads to error and thereby is unreliable. But ignoring counterevidence can also just be an accidental by-product of an otherwise very reliable process. Consider the following two cases. *Case One*: John is a highly ambitious philosopher. He wants to come up with some new, original theory about epistemic defeaters. Finally, he happens to have a very good idea on this topic. Now, there are some obvious objections to his new theory around and he knows of these objections. Nevertheless he is so strongly biased towards his own theory that he just ignores the available counterevidence and holds on to his theory. He ignores the counterevidence because of a strong confirmation bias. In this case it seems obvious that the sustaining process of John's belief is highly unreliable. But now consider *case two*: Frank has worked hard on the details of a scientific theory which he holds to be true. He is exhausted and needs all his concentration to go on. Although in principle he is very sensitive to counterevidence, he happens to overlook some relevant counterevidence just by accident. In this case it seems likely that the sustaining process of Frank's belief is reliable because this type of process does not regularly make people ignoring the available counterevidence. If this is true, then the following conclusion seems reasonable. It is possible that someone (namely

Frank in case two) is ignoring available counterevidence, even though his belief is sustained by a reliable process. If we assume that ignoring available counterevidence generally removes justification, then being sustained by a reliable process is not sufficient for a belief being justified. I therefore don't think that conservative reliabilism is a tenable position.

What about denying the existence of defeaters, a strategy Fred Dretske (2000) refers to as "Mad Dog Reliabilism"? A proponent of this strategy would insist that reliably produced beliefs remain justified, *even if counterevidence is available to the believer*. This is an odd view, since the defeasibility of justification is strongly suggested by our intuitions about justification. Of course, one need not take all these intuitions at face value. But then one should better have a story at hand that can explain away these intuitions. Mylan Engel (1992) offered such a story. Distinguishing between *personal* and *doxastic* justification he maintains that our intuitions about defeasibility concern personal, rather than doxastic justification. Hence, a *person* is not justified (rational or responsible) in holding on to her belief in the face of counterevidence she is aware of. But according to Engel, this does not imply that the *belief* she holds on would itself become unjustified. Engel's position differs from "Mad Dog Reliabilism" in so far as it tries to accommodate our intuitions about defeaters. But I think that Engel still does not take these intuitions seriously enough, since we obviously have the intuition that the epistemic quality of the belief as such is affected by available counterevidence.

So far we have seen that neither conservative reliabilism nor a position that denies defeaters are promising strategies for the reliabilist. Therefore, it seems unavoidable that the reliabilist changes her position to a certain extent in order to *integrate* defeaters into her account. So, let us look more closely at revisionary accounts of reliabilism. Goldman (1979, p. 20) suggests the following modification of simple reliabilism:

(G) S is justified in believing that p at time t, if and only if

1. S's belief is based on a reliable process, and
2. there is no conditionally reliable process which S *could* have used and which, if it had been used, would have resulted in S's not believing p at t.

Clause (2) is an extension of simple reliablism which pays tribute to the defeasibility of justification.

In his comment on this proposal Goldman makes it sufficiently clear that he does not mean clause (2) to imply that justification is defeated by the fact that newly gathered evidence would yield a different doxastic attitude. According to Goldman, justification is only defeated by *already acquired counterevidence* that would make belief-revision internally rational. Goldman's suggestion seems to be extensionally adequate. As far as I can see, it licenses the right cases as justified. Consider David's case again: since Frank told him that there has never been a sheep in the field there is a conditional reliable process at David's disposition which, if it had been used by David, would have resulted in David's not believing that there was a sheep in the field. David just could have inferred that there was no sheep in the field from what Frank told him. So, condition (2) of Goldman's account is not satisfied. Hence, we get the desired result: according to (G) David's belief is no longer justified. Yet, (G) is still not fully satisfying. First, it seems to be *fairly ad hoc*. For, the suggestion amounts to the claim, in reliabilist terms, that a belief is justified if and only if it is reliably produced and there are no defeaters (which would lead to belief-revision in internally rational agents). It fails to explain *why internally rational counterevidence removes justification*.[2] Second, it is *not clear why Goldman's proposal (G) is still a version of reliabilism.* If we put aside the technical details, reliabilism explains all justificationally relevant features as being objectively conducive to the goal of truth. But one does not see, how condition (2) fits into this general picture. (2) excludes cases in which someone does not adapt his beliefs to her internally available evidence. But (2) does not tell us why this internal adaptation is instrumentally good with respect to the goal of truth. The example used by Goldman seems to point in the opposite direction.[3] In that example Jones has reliable memory beliefs about his past. But his parents try to deceive him by telling him a false story according to which his memory is completely corrupted. Jones does not believe his parents, but persists in believing his memory. From a reliabilist point of view Jones' reaction seems perfectly in order. He does not care about misleading counterevidence and persists in believing the truth. But Goldman admits that after having heard what his parents told him Jones is no longer justified in holding his memory beliefs. Now, it might be possible to go this way even as a reliabilist. But the reliabilist then owes us an answer to the question why sensitivity to counterevidence (no matter whether it is true or false) is objectively truth-conducive. Goldman does not give this answer. Thirdly, it seems natural to say that Jones did not do what he epistemically *should* do, when he persists

[2] Notice that on Goldman's view a defeater may be a false or even highly unreliable evidence. Condition (2) only requires that the available process is *conditionally* reliable. So, the inference from the evidence to not believing p must be valid. But the input can be false and unreliable.
[3] Goldman 1979, p. 18.

in believing his memories. The defeaters he has got are *normative defeaters*. Interestingly, Goldman himself describes the case in normative terms:

> "So what we can say about Jones is that he fails to use a certain (…) process that he (…) *should* have used. (…) So, he failed to do something which, epistemically, he *should* have done. (…) The justificational status of a belief is not only a function of the cognitive processes actually employed in producing it; it is also a function of processes that (…) *should* be employed."[4]

Goldman here implicitly accepts that Jones is committed to certain epistemic obligations. But he doesn't tell us where these obligations come from. Furthermore, (G) does not imply any normative statements. So, within Goldman's account the normativity of normative defeaters remains unexplained.

Here is another suggestion of how to integrate defeaters into the general reliabilist framework which comes close to proposals by Alvin Plantinga (2000, pp. 359-366) and Michael Bergman (2006, Ch. 6):

(IR) S is justified in believing that p at t, if and only if

- (1) S's belief is based on a reliable process, and
- (2) there is no mental state at t in S's representational system which makes believing that p internally irrational. (no-defeater condition)

This conception seems to be closely related to the solution suggested by Goldman's (G). I even think that both are approximately equivalent. The interesting thing about Plantinga and Bergman is that they offer explanations for condition (2) which can answer the question why counterevidence removes justification and also, at least in Plantinga's case, explain the normativity of certain defeaters. For Plantinga a justified belief must not depend on a malfunction of the cognitive system, and properly functioning cognitive systems would remove internally irrational beliefs. Since a defeater for believing that p makes that belief internally irrational, the system can tolerate that belief only if it is not properly functioning, i.e. if it is not working as it *should*. Hence, believing that p in the face of internally rational counterevidence is unjustified.

Yet although Plantinga does explain why defeaters remove justification, his explanation remains problematic. It depends on a supra-naturalistic account of proper functions.[5] Moreover, the normative notion of proper functioning has nothing to do with reliability.

[4] Goldman 1979, p. 20 (my emphasis).
[5] See Plantinga 1993b.

Especially the proper functioning of internal rationality has nothing to do with getting at true beliefs. Plantinga understands *internal* rationality as a matter of proper function "downstream from experience."[6] Since internal representations may be radically false, there is no truth-connection inherent to internal rationality.

In contrast to Plantinga, Bergman does not need the normative concept of proper functioning. According to him, a belief is justified if it is reliably produced and, in addition, rational "from the inside." There have to be accessible evidential states, like beliefs or experiences, whose contents support the truth of the beliefs that are based on them from a first-person perspective. My perceptual belief "There is something red in front of me" is justified if I have the experience of something red in front of me and my visual faculties are reliable on that occasion. We may call this position "Evidential Reliabilism."[7] If a belief is held without sufficient evidential support, it is internally irrational. From this perspective, defeaters can be understood as pieces of evidence that destroy or neutralize the evidential support of a belief and thereby remove its prior justification. Consider again the sheep case: When David acquires the belief that a sheep is in the field by using visual perception, both necessary conditions of justification are satisfied: (i) David possesses the supporting visual evidence that something in front of him looks like a sheep and (ii) his visual faculties are working reliably in those circumstances. However, when David is told that there was no sheep in the field, his evidential basis has changed. If he considers both that something in the field looked like a sheep and that there was no sheep in the field, then his belief that a sheep was in the field is no longer evidentially supported. His belief is still reliably produced, but lacks the necessary evidential support.

Although Evidential Reliabilism gives a cogent answer to the question why defeaters remove justification, there are a number of strong objections to this position. *First*, it is simply not true that every justification requires supporting evidential states, as Evidential Reliabilism claims. Consider, for example, introspective beliefs. It is a widely held view that they are not based on any evidence. If I acquire the belief that I experience something red right now and if I acquire this belief via introspection, then I do not base my belief on something like an inner experience of the experiential state in question.[8] Rather, I have an immediate belief about my

[6] Plantinga 2000, p. 365.
[7] I owe this term to XXX.
[8] See Shoemaker 1996, p. 207. Some philosophers, e.g. Sosa 2007, p. 45, claim that in introspection the mental state itself is the evidence for the introspective belief about it. But there are severe objections to this view. First, introspective beliefs do not only represent the content of the first order state, but also that it is a mental state and

current perceptual state. Even if this belief were false, since it might get something wrong about the content of my state, it would nevertheless have some positive epistemic quality. If it cannot be knowledge (since the belief is false), it must be justification. Hence, it must be possible for an introspective belief to be justified without being based upon evidence. Or consider testimonial justification. In that case, we often do base our beliefs on evidential states, namely the utterances we hear. But even if we directly recognize their meaning, this does not evidentially support the truth of what we are told. Assume that you hear that someone utters that p. This evidence alone does not support your belief that p. There is no evidential connection between uttering that p and p being the case. Therefore, even in the case of testimony we lack *supporting* evidence. Or consider finally the case of self-evident propositions, as e.g. "1+1=2". They seem to be justified. But their justification does not rely on any evidence (contrary to what the term suggests). We are just attracted to assent to the propositional content of self-evident propositions by entertaining them in thought. But entertaining a propositional content surely is not the evidence which justifies the belief.[9] For, we are entertaining all kinds of propositions which may be unjustified. In short: Evidential reliabilism cannot explain why non-evidential justification, which obviously exists, is defeasible. *Second*, Evidential Reliabilism is a mixture of reliabilism and internalism. On this view, defeaters have an explanation that is completely internalist in nature.[10] Therefore, it does not give us a thoroughly reliabilist account of defeaters. *Third*, Evidential Reliabilism does not explain the normativity of normative defeaters.

So far I have argued (1) that a promising reliabilist account should leave room for defeaters and (2) that all existing reliabilist accounts that satisfy (1) either do not give an adequate explanation of why defeaters remove justification or give an explanation which is not reliabilist in spirit. Finally, I want to present my own account of defeaters which is supposed to overcome these shortcomings and remain completely reliabilist in spirit. Here is my own suggestion:

---

which kind of mental state it is (desire, experience, belief etc.). Now, since two mental states are related evidentially solely in virtue of their content, my first-order experience of something red cannot be evidence for my introspective belief that I have an experience of something red. The evidential basis for the claim that it is an experience is simply missing in the first-order state. Second, it seems reasonable to assume that there are justified introspective beliefs which are false, maybe due to mistakes of inattention or bias through strongly misleading expectations. Let us, for example, assume that I have the experience of a certain shade of red_4, but I represent it as an experience of red_{12}. In this case, the content of my first-order experience is not an evidential basis for my mismatching introspective belief. Still my introspective belief may be justified, though false.

[9] For this view compare Sosa 2007, p. 55.

[10] Compare Alston 1989.

(TG) S is justified in believing that p at time t, if and only if

(1) S's belief is based on a reliable process, and

(2) there is no conditionally reliable process available to S which (i) *a properly functioning cognitive system* of the kind to which S belongs *would have used* and (ii) which would have resulted in S's not believing p at t, and

(3) the proper function mentioned in (2) can be explained with respect to getting at true beliefs.

In order to demonstrate that (TG) is completely reliabilist in spirit, one first has to show how belief-inhibitory processes can be classified as reliable. Of course, these processes do not lead to true beliefs more often than to false beliefs. But we can call them "reliable", if and only if they *eliminate* false beliefs more often than true beliefs, when their input is true. Secondly, clause (2) mentions in, contrast to Goldman's account, belief-inhibitory processes which a *properly functioning* cognitive system would have used. (TG) thereby pays tribute to the normative dimension of defeaters. Thirdly and most importantly, clause (3) requires that being sensitive to internal counterevidence is itself truth-conducive or at least somehow valuable in getting at the truth (and avoiding errors). This clause distinguishes (TG) from, for example, Plantinga's view, according to which avoiding defeaters is a purely internal affair. How can internal rationality of belief-formation, as required by (2), be understood as truth-conducive? Let me roughly sketch how such an explanation might look like. My intention is to give a completely naturalist explanation of proper functions, as it has been suggested by Ruth Millikan (Millikan 1993). According to Millikan, A has the proper function F if and only if A originated as a reproduction of some prior item that has performed F in the past, and A exists because of this prior performance. Let us apply this definition to belief-revising cognitive systems. By correcting errors or sources of error, the cognitive system usually improves the overall truth-ratio of its beliefs. The overall reliability of a cognitive system will be massively improved if its beliefs are rationally sensitive to errors or sources of error. This capacity can be implemented by the cognitive system, only in so far as the system is sensitive to what looks to be counterevidence from the inside. Under normal conditions (i.e. if the cognitive perspective on the world is reliable on the whole), avoiding internal irrationality will massively eliminate error. Now, a cognitive system that eliminates error is better adapted to its environment than a cognitive systems that does not and thereby the former gains a reproductive advantage. This explains in a naturalist manner how subsequent cognitive

systems acquire the proper function *of being rationally sensitive to counterevidence.* That the system functions properly only if it avoids internal irrationality even holds in cases (like Jones' case) where the available counterevidence would be misleading. If counterevidence is ignored in any particular case, the cognitive system is malfunctioning in sustaining that belief. In this case, the belief is no longer justified, since condition (2) of (TG) is not satisfied.

Let me conclude by pointing out what, on my view, are the advantages of (TG) over (G) and (IR). Whereas Goldman's (G) cannot explain the normative dimension of defeaters, (TG) says that a cognitive system is not sensitive to counterevidence as it *should* since it does not fulfil its proper function. Neither (G) nor (IR) really explain why the no-defeater condition (2) is reliabilist in spirit. (TG) promises with clause (3) an answer to that question. We have seen why and how a cognitive system may adopt the proper function of being rationally sensitive to internal counterevidence in order to get at the truth and reproduce itself.[11]

References

Alston, Williams 1989: "An Internalist Externalist," in: Alston: Epistemic Justification, Ithaca/London, pp. 227-245.

Alston, William 2002: "Plantinga, Naturalism, and Defeat", in: James Beilby: Naturalism Defeated?, Ithaca/New York, pp. 176-203.

Bergman, Michael 2006: "Defeaters", in: Bergman: Justification without Awareness, OUP, pp. 153-177.

BonJour, Laurence 1985: The Structure of Empirical Knowledge, Cambridge (MA).

Dretske, Fred 2000: "Epistemic Rights without Epistemic Duties," in: Philosophy and Phenomenological Research 60, pp. 591-606.

Engel, Mylan 1992: "Personal and Doxastic Justification in Epistemology," in: Philosophical Studies 67, pp. 133-150.

Goldman, Alvin 1979: "What is justified belief?", in: G. Pappas (ed.): Justification and Knowledge, Dordrecht, pp. 1-23.

Millikan, Ruth 1993: "In Defense of Proper Functions," in: Millikan: White Queen Psychology and Other Essays for Alice, Cambridge (MA), pp. 13-29.

Plantinga, Alvin 1993b: Warrant and Proper Function, Oxford.

Plantinga, Alvin 2000: "The Nature of Defeaters", in: Plantinga: Warranted Christian Belief, OUP, pp. 359-366.

Shoemaker, Sydney 1996: The First Person Perspective and Other Essays, Cambridge.

Sosa, Ernest 2007: A Virtue Epistemology. Vol. I, Oxford.

[11] Acknowledgements.