Abstract: According to the doctrine of infallibility, one is permitted to believe $p$ if one knows that necessarily, one would be right if one believed that $p$. This plausible principle—made famous in Descartes’ *cogito*—is false. There are some propositions one can’t be wrong about but shouldn’t believe anyway: believing them would make one’s overall epistemic state worse. That we are sometimes required to forego acquiring guaranteed true beliefs suggests that epistemic value cannot fully be explained in terms of accuracy.

1. The Doctrine of Infallibility

A proposition is epistemically infallible for an agent just in case that it’s impossible for that agent to falsely believe it:

\[
\text{Infallibility: A proposition } p \text{ is infallible for } S \text{ iff it’s impossible that } S \text{ falsely believes that } p.
\]

This definition closely resembles the following:

\[
\text{Infallibility*: A proposition } p \text{ is infallible for } S \text{ iff it’s necessary that if } S \text{ believes that } p \text{ then } S \text{ truly believes that } p.
\]

If, necessarily, all beliefs are either just true or just false—as I believe they are—then infallibility and infallibility* are equivalent. I shall suppose that they are equivalent throughout the rest of the paper, although there is interesting territory to explore for those who believe that some propositions are neither or both true and false.

Since one can’t be mistaken about one’s infallible beliefs, it’s tempting to think we should always believe them. When there’s no risk of false belief, why not? At least, if we’re in a position to recognize that a proposition is infallible for us, surely then we may believe it. As Alston says, “one could hardly have a stronger (epistemic)

---

justification for holding a certain belief than the logical impossibility of the belief’s being mistaken” (Alston, 1971: 229). Call this the doctrine of infallibility:

**Doctrine of Infallibility:** (For all S and for all p) If S knows that <necessarily, if she herself were now to believe that p then she would truly believe that p> then it is thereby (rationally) permissible for S now to believe that p.

In slogan form: You’re always permitted to believe (known) guaranteed truths.

I argue, however, that not all propositions known to be infallible may be believed. In fact, some propositions that are known to be infallible should be disbelieved!

The doctrine of infallibility emerges as a battleground between two otherwise attractive philosophical theses: one the one hand, *veritism*, the thesis that accuracy is the fundamental epistemic good, and on the other hand what I shall call *reflectivism*, the thesis that one’s reflective attitudes about one’s first-order beliefs make at least some difference to the epistemic quality of those first-order beliefs. I’ll argue that the best way to resolve the conflict is to give up on veritism, but I shall also try to act as a guide who points out the various paths that may be trod to escape the puzzles that the doctrine of infallibility provokes.

2. Clarifying “Infallibility”

Some clarifications are in order. First, infallibility is sometimes ascribed to agents in relation to a proposition or subject matter (e.g. “The Pope is infallible about matters of faith when he speaks *ex cathedra.*”) Thus Alston writes: “one can be said to be infallible *vis-à-vis* a certain subject matter provided one cannot be mistaken in any beliefs he forms concerning that subject matter” (Alston, 1971: 229).

But as we’re using the term, this is exactly backward. Infallibility is a property that propositions have in relation to agents. Infallibility requires no special competence on the part of the agent for whom a proposition is infallible. All necessary

---

2 Alston focuses on logical impossibility, although it seems that the strength of epistemic justification would remain just as high if the agent recognizes that it is metaphysically impossible for the belief to be mistaken.

3 I say “herself” rather than “S” here to emphasize that the agent must recognize that she herself is the person for whom p would be infallible. The difference is apparent when an agent does not know who she is. Suppose that Victoria knows that it’s infallible for Victoria that Victoria exists but does not know that she herself is Victoria (e.g. because she has recently had an amnesic episode and forgotten who she is). It’s not obvious that she is permitted to believe that “Victoria exists” even though she knows that the proposition is infallible for Victoria (but not, in the relevant sense, infallible for herself).

4 Thanks to Charity Anderson for suggesting this kind of slogan.
mathematical truths are infallible for infants, for instance, though of course infants do not know that those truths are infallible for them.

Second, what I mean by “infallibility” is related to but distinct from what some authors mean by “incorrigibility.” It’s useful to briefly disentangle them.⁵

Frank Jackson discusses the view that “it is logically impossible to be mistaken about certain of one’s current mental states” under the guise of an “incorrigibility thesis” (Jackson, 1973: 51). This is an obviously related notion, but one that comes apart from my definition of infallibility. For instance, that I exist is infallible for me, but it’s at least not obvious that I couldn’t be mistaken about it. Perhaps I believe that I am a fiction, for instance, and that strictly speaking I don’t exist.⁶ I’d then be mistaken about whether I exist even though, necessarily, if I believed <I exist> I’d be right.

Shoemaker’s definition of incorrigibility is closer in that it is conditional upon a kind of affirmation: “If a person sincerely asserts such a statement it does not make sense to suppose, and nothing could be accepted as showing, that he is mistaken” (Shoemaker, 1963: 215). Shoemakerian incorrigibility is a kind of public infallibility. If a person believes an infallible proposition, their (sincerely believed) assertions of the same will be incorrigible in Shoemaker’s sense—after all, no proposition that is true can be shown to be mistaken. But not all (Shoemakerian) incorrigible propositions are infallible, for a proposition might be incapable of being disproved but false nonetheless.

There may, however, be a sense in which my infallibility is a kind of incorrigibility. Incorrigribility is (according to one way of thinking) the inability to be corrected. An agent might be uncorrectable with respect to a proposition because (a) they are unable to change their mind (or, at any rate, to have their mind persuasively changed from the outside) or (b) any change of mind would not constitute a correction, i.e. because the belief in question was already true. Any belief in an infallible proposition is incorrigible in the second sense.

3. On Behalf of the Doctrine of Infallibility

---

⁵ “Incorrigribility” and “infallibility” have sometimes been either conflated or used interchangeably in the literature. Alston, for instance, in his own detangling work, explicitly uses “infallibility” to identify a common strand through an assorted collection of (other people’s) definitions of “incorrigribility” (Alston, 1971: 225), although he is clear on his own preferred usage. I don’t claim that my use of these terms is better English than anyone else’s. These are technical terms, and what’s important is that we’re clear on what they mean (and what they don’t) in a given context.

⁶ Consider, for instance, a view in the vicinity of Lebens (2015).
But before defending the surprising conclusion that the doctrine of infallibility is false, let’s give the doctrine its due. The doctrine is deeply intuitive and for good reason. First, the doctrine has historical pedigree. It is to infallibility that Descartes appeals at the climax of the cogito:

[Suppose] there is a deceiver of supreme power and cunning who is deliberately and constantly deceiving me. ...[L]et him deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something. ...I must finally conclude that this proposition, I am, I exist, is necessarily true whenever it is put forward by me\(^7\) or conceived in my mind\(^8\) (CSM, 17: Second Meditation, emphases mine).

It’s the infallibility of “I exist” that makes the wheels of the cogito turn and—even at least one way of reading Descartes’—makes it the first item to fully survive the scrutiny of Descartes’ method of doubt.\(^10\)

The cogito illustrates the benefit of having a principle whereby guaranteed truth is sufficient all by itself to license belief. If the doctrine of infallibility were false, that would mean that Descartes had not yet done enough (at this point in the Meditations, at least) to show that we are permitted to believe that we exist! Descartes is often accused of making the standard for defeating skepticism too high: rarely has he been accused of making the standard too low.

Contemporary epistemologists continue to appeal to Cartesian infallibility at crucial junctures. Ernest Sosa, for instance, appeals to infallibility to ward off dream skepticism. Having argued that in dreams “we do not really believe; we only make-believe” (Sosa, 2007: 8), Sosa claims that if we really believe that we are dreaming we must not be dreaming: If we were dreaming, our belief wouldn’t really

---

\(^7\) I take believing a proposition to be at least one relevant way of putting it forward.

\(^8\) Believing a proposition plausibly involves conceiving of the believed proposition in one’s mind.

\(^9\) Descartes’ total anti-skeptical project is, of course, richer than just the cogito. And there’s an important exegetical question of how Descartes’ notion of infallibility employed in the cogito relates to the clarity and distinctness criterion that takes center stage in Meditation III. But proven infallibility seems central at least to the cogito even if not Descartes’ total project. If this is not what Descartes himself thought, it is at least how many have received his words. (I’m grateful to conversation with Charity Anderson, Laura Callahan, Ram Neta, and Ernie Sosa on this point.)

\(^10\) Admittedly, Descartes does not defend the doctrine of infallibility as such in the Meditations. It’s compatible with what Descartes says that although the infallibility of the proposition <I exist> is sufficient reason to believe it, the infallibility of many other propositions is not sufficient reason to believe them. (Similar observations apply to Sosa’s work on dream skepticism.) Nevertheless, I think it would be puzzling if Descartes did not implicitly endorse something like the doctrine of infallibility. Descartes simply appeals to the infallibility of the proposition <I exist.> He nowhere suggests that it is the proposition’s infallibility plus it’s being a proposition of such-and-such a kind that makes it proper to believe. If Descartes was not at least implicitly committed to the doctrine of infallibility, it’s puzzling that he says nothing about why infallibility in this case is significant. (I’m grateful to Christopher Frugé for encouraging clarity on this point.)
be a belief at all, but only a make-belief. We should affirm (and neither suspend nor deny) that <I am awake> “since only about that option [i.e. affirming] is it obvious to me now that if I take it I will be right” (Sosa, 2007: 19, emphases mine).\textsuperscript{11}

A second reason for the doctrine of infallibility emerges from a particular picture of epistemic value. “Believe truth! Shun error!” says Williams James (1907: 18). Epistemically, “these are our first and great commandments” (James, 1907: 17),\textsuperscript{12} and—a more ambitiously reductive epistemologist might have added—the only ones.\textsuperscript{13}

James’s aphorisms are suggestive of one of the main characters in our dialectic: epistemic veritism. According to veritism, the only fundamental epistemic value is accuracy. Whatever else can be said for and against this view, it is attractively simple. James’s aphorisms are telic: believing truth and shunning falsehood are goals that epistemic agents ought to promote. James notes that an agent might emphasize one of these goals more than the other. A cautious believer might be hesitant to believe when there is even a small chance of error.\textsuperscript{14} Infallible beliefs, however, are such that there is some chance they will lead to true belief and no chance that they will lead to false belief. If the only—or, at any rate, the most fundamental—epistemic goods are believing truly and avoiding false belief, then there’s seemingly always a good decision-theoretic reason to take a chance on infallible beliefs: There’s an opportunity (indeed a certainty!) to gain an epistemic good without any risk of epistemic harm.

One could accept the doctrine of infallibility without being a veritist of this stripe, but veritists should be especially attracted to the doctrine of infallibility. Adding a true belief always improves accuracy, at least if one can do so without adding any false beliefs or removing true ones. And there’s no obvious reason why believing an infallible proposition would require one to also acquire false or abandon true beliefs. Indeed, the very plausibility of the doctrine of infallibility is likely to be seen as an argument for veritism. If guaranteed truth is by itself sufficient to license belief, a good explanation is that accuracy is what matters most in epistemology.

\textsuperscript{11} Compare Sosa’s response to Wittgenstein’s in On Certainty (1969: 383): “The argument ‘I may be dreaming’ is senseless for this reason: if I am dreaming, this remark is being dreamed as well—and indeed it is also being dreamed that these words have any meaning.” Whereas for Sosa, having a “belief” while dreaming isn’t really a belief at all, for Wittgenstein the content of the belief in a dream is senseless. Wittgenstein’s remarks require parsing, but one interpretation is that, for Wittgenstein, beliefs in dreams are neither true nor false. If this is so, then Wittgenstein is a philosopher for whom there is a real distinction between infallibility and infallibility*.

\textsuperscript{12} Intriguingly, James himself fluctuates between believing the truth (1907: 18) and knowing the truth (1907: 17) as the positive epistemic commandment. In this paper, I shall represent James as endorsing the commandment to believe the truth, although this is a simplifying, historical fiction.

\textsuperscript{13} I do not claim that James himself was reductive in this sense.

\textsuperscript{14} Cf. Tom Kelly (2014)
Third, the doctrine of infallibility, or at least a principle that entails it, explains the right verdict in certain tricky cases that reverse the normal direction of fit. Note that the doctrine of infallibility is a weak version of a family of principles that permit agents to believe when they know that their belief would have some truth-oriented property or other if believed. In particular, the

**Doctrne of Infallibility**: If S knows that *necessarily, if she herself were now to believe that p then she would truly belief that p* then it is thereby (rationally) permissible for S now to believe that p.

is entailed by the

**Doctrne of Truth**: If S knows that *if she herself were now to believe that p then she would truly believe that p* then it is thereby (rationally) permissible for S now to believe that p.

For if S is permitted to believe p in virtue of knowing that she *wouldn’t* be wrong about p, then she is surely permitted to believe p in virtue of knowing that she *couldn’t* be wrong about p.

It turns out that the Doctrine of Truth is extremely useful in explaining why we are permitted to believe certain propositions when the ordinary direction of fit between mind and world is reversed. In “Infinitely Permissive” (forthcoming), Morten Dahlback considers cases that reverse the ordinary direction of fit in light of a principle like the doctrine of known truth.\(^\text{15}\) He argues that the principle shows that some sets of evidence permit both believing that p or believing that not-p (although not, of course, both at once).

Dahlback considers a case in which a demon guarantees that the result of a coin flip will match your belief about whether it is heads or tails. Dahlback reasons that if we knew we were in such a situation, we would be permitted either to believe that the

\(^{15}\) Dahlback (forthcoming) considers this principle:

**Known Correctness Entails Permissibility (KCEP)**: If you know that if you ϕ, it will be correct of you to ϕ, then you are permitted to ϕ. (Dahlback, 1)

One assumption I will make in interpreting Dahlback’s principle is that a belief’s being true is sufficient to make the belief correct in the relevant sense. And indeed, this interpretation fits well with the examples Dahlback uses. Even so, KCEP isn’t quite the same as the doctrine of truth—KCEP is formulated as a more general thesis about permissible ϕ-ings and not beliefs in particular—but the differences need not detain us here: the spirit is the same. To emphasize the relationship between the doctrine of truth and the doctrine of infallibility, I will discuss Dahlback’s examples through the lens of the doctrine of truth rather than his KCEP.
certain point in the reasoning process, it becomes impossible to believe and thereby make oneself right. So it isn’t as though evidence has no role to play. But at a 20
plausible factor that determines rationality favoring \(<\text{the coin lands heads}>\) over \(<\text{the coin lands tails}>\). Thus, I think we can be cautiously optimistic that focusing on non-favoring evidence does not favor neither \(p\) nor \(\neg p\) we are required to suspend judgment. One can’t reason one’s way on the basis of the evidence to the conclusion that the coin will land heads or tails—one simply believes and thereby makes oneself right. That the antecedent evidence favors neither thesis is important: It makes clear the role that the doctrine of truth plays in certain cases. 16

16 Although it seems intuitive to me that one is permitted to believe, e.g., that the coin will land heads even before one forms the belief, one could insist that there’s an important difference in one’s epistemic situation before and after one has the target belief. Perhaps before you have the belief that the coin will land heads it’s impermissible for you to form it (after all, you don’t have favoring evidence for it yet) but you are allowed to hold onto the belief that the coin will land heads in conjunction with your knowledge that the demon will guarantee this outcome. On this interpretation, you’re not always permitted to acquire guaranteed truths but you’re always permitted to hold onto them. (Thanks to Matt McGrath for raising the possibility of this interpretation.) There’s an interesting debate to be had here about the dynamics of belief formation. One thing that requires explanation on this new proposal is why the difference between acquiring and preserving makes an epistemic difference (e.g., is one process more likely to be truth conducive?). In any case, the ultimate argument of the paper can survive the revised principle. See fn 24.

17 I introduce the terminology “favoring evidence” because there is some pro-relation (other than favoring) between the evidence and both the proposition \(<\text{the coin will land heads}>\) and the proposition \(<\text{the coin will land tails}>\). The evidence permits or licenses either belief. But the evidence does not favor one of these beliefs over the other.

18 Not all distinct attitudes are competitors. For instance, although being confident in \(p\) and believing that \(p\) are distinct attitudes (one can be confident in \(p\) without believing that \(p\)), they are not competitive: someone who believes that \(p\) may and indeed should be confident in \(p\). Examples of competitive doxastic attitudes include: believing \(p\) and disbelieving \(p\); believing \(p\) and suspending on \(p\); having a .7 credence in \(p\) and having a .6 credence in \(p\); believing \(p\) and having a .1 credence in \(p\).

19 This is too quick. Evidentialists should think that if the evidence favors neither \(p\) nor \(\neg p\) we are normally required to suspend judgment (apart from the direction of fit cases considered here). Non-evidentialists think that something (other than or) in addition to the evidence (partially) determines what it is rational to believe. (I argue for a form of non-evidentialism in Willard-Kyle (2017).) This leaves space for the evidence to favor neither \(p\) nor \(\neg p\) but for the full set of factors that determine rationality (perhaps including but not limited to the evidence) to favor one attitude over the other. The coin-flip case is intriguing, however, in part because it’s hard to imagine any plausible factor that determines rationality favoring \(<\text{the coin lands heads}>\) over \(<\text{the coin lands tails}>\) or vice versa. Thus, I think we can be cautiously optimistic that focusing on non-favoring evidence (as opposed to non-favoring bases of rationality more generally) is a harmless simplification in this case.

20 Admittedly, in Dahlback’s case one has evidence that one is in the sort of situation that one can believe and thereby make oneself right. So it isn’t as though evidence has no role to play. But at a certain point in the reasoning process, it becomes impossible to follow the evidence to a conclusion.

16

17

18

19

20
plays (or seems to) in explaining the permissibility of believing either that the coin will land heads or that it will land tails.  

As Dahlback notes, an important feature of the case is that the ordinary direction of fit between mind and world is reversed. We are guaranteed to be right not because our mind is tracking the world but because the world is tracking our mind. That the doctrine of truth can deliver the right verdict in such cases—when the favoring evidence seems to run out—is a strong point in its favor. Considered broadly, such cases fit neatly with the reductive gloss on James sketched earlier. True and false beliefs are what matter. Who cares (our imagined Jamesian shrugs) whether how you got there was by following the evidence? What’s important is that you knew you’d be right at journey’s end.

A final virtue of the doctrine is that it permits epistemic agents to believe in virtue of knowing information that is entirely non-normative. Agents don’t need to know that a belief is good or reasonable or better than some alternative: they only need to know that, necessarily, if they believe it, it will be true. The doctrine wears on its sleeve what non-normative property (truth) is the basis for the epistemic action licensed by it. This makes it especially useful as a principle to consider in discussions about epistemic value.

4. The Problem of Easy Downgrade

So there’s much to be said for the doctrine of infallibility. It anchors prominent anti-skeptical arguments, encourages us to take smart epistemic bets, and helps explain some otherwise tricky cases when following the evidence isn’t enough.

Despite this, the doctrine of infallibility is false. Although we usually make our overall epistemic state better by believing in such a way we couldn’t be wrong, we can also make our overall epistemic state worse.

The doctrine of infallibility is false because it makes it too easy to permissibly acquire defeaters for our beliefs (or at least too easy to rationally downgrade them). A belief is defeated by another belief, in the stipulative sense of this paper, when the second belief makes the first lose some positive epistemic status. Suppose I believe it is noon but then learn that the clock I am looking at, and upon which I had based my belief that it is noon, runs an hour late. My belief that the clock is running...

---

21 Dahlback uses the cases to argue for extreme rational permissivism, the thesis that contradictory propositions can both be rationally permitted by the same evidence. Andrew Reisner (2014) uses similar cases to argue against evidentialism. Another similar case is Sherrilyn Roush’s (2005) fairy godmother who guarantees that all of an agent’s beliefs are safe (although in Roush’s version, the agents do not know that they have such a benefactor), which she employs in an argument concerning the limits of safety as a solution to certain problems in the analysis of knowledge.
an hour late defeats the justification for my belief that it is noon. Losing justification is one way of losing a positive epistemic status, so the belief has been (in our sense) defeated. (Note that our sense of “defeat“ is intentionally broader than those that require loss of some particular epistemic quality like justification or knowledge.)

It’s controversial just when defeat happens. To get a grip on the problem for the doctrine of infallibility, let’s begin by considering a very permissive defeat principle. Although many (myself included) will find this principle ultimately unconvincing, it will help us to identify a recipe for finding infallible propositions that ought not be believed. Here’s our easy defeat principle:

**Easy Defeat:** Necessarily, (for all S and for all p) it is irrational for S to believe that p if S believes that it is irrational for S to believe that p.

The principle **Easy Defeat** has some plausibility. Suppose S believes that it is irrational for her to believe some proposition, but she believes it anyway. The agent apparently displays a lack of appropriate epistemic reflection. She believes she has no good reason for believing p and yet goes ahead and believes it anyway. Something seems to have gone wrong epistemically.

**Easy Defeat** is thus one avatar (though not the only one) of our second character: reflectivism. It’s one way of expressing the intuition that one’s reflective attitudes about one’s first-order beliefs make at least some difference to the epistemic quality of those first-order beliefs.

Let’s not worry too much about whether this principle survives scrutiny. What I want to argue here is that if **Easy Defeat** is true then there are certain infallible propositions that should not be believed.

Suppose an agent knows both p and **Easy Defeat**. They consider whether they may believe, in addition, that it is irrational for them to believe that p. They ask themselves, “Suppose I were to believe that it is irrational for me to believe that p. Would that belief be true?”

Absolutely! For according to **Easy Defeat**, believing that it’s irrational to believe that p is enough to make believing that p irrational. Simply having the belief makes it so. More than that, since **Easy Defeat** is a necessary truth, it’s impossible that the belief could be false. The agent knows that the belief <it is irrational for me to believe that p> is infallible for them. The agent knows that, necessarily, if she herself were to believe that <it is irrational for me to believe that p> then she would truly believe that <it is irrational for me to believe that p>.
But obviously, it’s wrong to believe that a belief is irrational just because Easy Defeat makes that higher-order belief infallible. This would lead an agent to have a worse set of beliefs overall if \( p \) was otherwise rational to believe (and if there wasn’t independent reason to doubt that it was rational to believe that \( p \)). Something’s gone wrong.

Maybe what’s gone wrong is Easy Defeat. After all, many epistemologists think defeat is hard to come by. Imagine, for instance, that a philosopher accepts an argument that “rational” is an ideal term:²² beliefs are only rational when they manifest exactly the right level of confidence in a proposition. This philosopher would think of nearly all their beliefs that they were irrational. But this wouldn’t necessarily mean that it would actually be irrational for them to continue to hold any of their first order beliefs. After all, a belief’s irrationality in this thin sense—of simply not being perfectly ideal—does not offer any advice about whether the agent should invest more or less confidence in their first-order attitudes.

Indeed, there are many other reasons some philosophers reject Easy Defeat. Nevertheless, thinking about Easy Defeat was valuable, for it has given us a template for thinking about how certain higher-order beliefs could in principle be infallible. Consider the following infallibility template:

**Infallibility Template:** Necessarily, (for all \( S \) and for all \( p \)) if \( S \) believes that it is \( F \) for \( S \) to believe that \( p \) then it is \( F \) for \( S \) to believe that \( p \).

Easy Defeat is true just in case we can plug in “irrational” for “F.” As noted, if defeat is hard then simply believing that a belief is irrational might not be enough to make the first-order belief irrational. But it still seems that negative higher-order epistemic appraisals make their corresponding first-order beliefs worse in some way, even if it doesn’t always make them irrational. Our task is to find some \( F \) that captures whatever way it is that first-order beliefs become worse upon receiving negative higher-order appraisals.

There are four broad ways that one might argue that no negative, epistemic property satisfies the infallibility template. First, one might endorse extreme level splitting, the view that our first-order beliefs are completely unaffected by our higher-order beliefs about them.

Extreme level splitting seems unduly strong. If we think it’s a total disaster to believe to believe that \( p \) but believe \( p \) anyway, surely that lowers the quality of our first-order belief in some way. If extreme level splitting is true, then we can

---

²² I.e. imagine that they believed of “rational” something like what Peter Unger (1978) has defended regarding “knowledge” as an absolute term.
completely ignore our higher-order beliefs when evaluating their first-order counterparts. That stretches credulity. Surely there are better ways to argue that nothing satisfies the infallibility template.

In other words, endorsing extreme level-splitting violates the intuitive thesis we’ve called epistemic reflectivism: our reflective beliefs about first-order beliefs have at least some impact on the quality of our corresponding first-order beliefs. (But hold this thought: we’ll return to it in the end.)

Second, one might argue that although negative higher order beliefs affect first-order beliefs in some way, no negative property is such that believing a belief has that property makes the first-order belief bad in that same way. For instance, imagine someone believes that they are not in a position to know that p. According to this strategy, one should accept that this makes one’s corresponding belief that p worse but deny that it automatically makes the corresponding belief worse in virtue of removing them from a position to know that p. Instead, believing that you are not in a position to know p makes your belief that p bad in some other way, e.g. by making it irrational or akratic or strategically suspect. More generally, believing that one’s beliefs are bad in an F way makes them worse but only guarantees that they are worse in a G-ish way and not in an F-ish way.

There are good examples of philosophers defend specific instances of this strategy. For instance, Maria Lasonen-Aarnio (2010) argues persuasively that we shouldn’t confuse our evidence being such that it’s unlikely that we know that p with our actually not knowing that p (Lasonen-Aarnio 2010: 10). An agent might have evidence that their visual capacities are misfiring, but if there visual capacities are operating well, and if the agent bases their belief solely on their visual capacities, it might be that “being stubborn pays off” (Lasonen-Aarnio 2010: 2). If this is right, then even though an agent’s belief is epistemically criticizable—it was unreasonably stubborn of them to believe given how unlikely it was, conditional on their evidence, that they knew—the agent’s first-order knowledge is not thereby impugned (in the good case).

Note, however, that it’s extremely important to Lasonen-Aarnio’s project that we can genuinely criticize the agent who believes when it’s unlikely on their evidence that they know—they are (according to her project) being unreasonable. It’s just that they might not be criticizable in the way that we first thought. If Lasonen-Aarnio could not explain why agents who believed (and thereby came to know) in the face of significant (but ignored) counter-evidence were criticizable in some way or other, that would be a mark against her account. Whether or not we agree with Lasonen-Aarnio’s particular thesis, I take one consequence of her paper to be that
we need to be cautious about saying *how* higher-order criticisms negatively impact first-order attitudes—but that doesn’t give us reason to doubt *that* they do so.

Motivated by this insight, our second strategy for avoiding the problem of easy defeat has said, in effect, that negative higher-order epistemic appraisals make their first-order counterparts worse in some way just not automatically in the way we believed them to be worse. To counter this strategy, we can move to a principle that uses a sufficiently *general* negative, higher-order appraisal. For instance, consider this rule:

**Easy Problems:** Necessarily, (for all S and all \( p \)) if S believes that it is problematic for S to believe that \( p \) then it is problematic for S to believe that \( p \).

Our second strategy was to argue that any negative higher-order appraisal would make the corresponding first-order belief worse but not necessarily worse in the way believed. But “problematic” is such a general, negative term that *any* way of making a first-order belief worse counts as problematic. So it seems that Easy Problems is true.

Unless our third strategy for resisting the infallibility template succeeds. Our second strategy was to argue that although believing that one’s beliefs are bad in an F-ish way makes them worse, it always makes them worse in a G-ish way and not an F-ish way. Our third strategy admits that there is some F (such as “problematic”) such that believing a belief to be F makes it worse in an F-ish way. But it denies that the first-order belief must be bad enough to make it F outright. So, for instance, the defender of the third strategy insists that although believing that it is problematic to believe \( p \) entails that it is *more* problematic to believe that \( p \) than it might have been otherwise, it doesn’t make believing \( p \) problematic outright. I treats “problematic”—or any F that might otherwise satisfy the infallibility template—as a threshold term, such that being more F does not entail being F (e.g. being *taller* than something doesn’t entail being *tall*).

It’s unclear to me whether this strategy is properly motivated—it’s far from obvious that “problematic” is relevantly like “tall.” But instead of pressing this line, we’ll look for a term that avoids the objection altogether: an F such that being more F (than whatever) entails its being F outright.

We can do this by stipulating new terminology that does not operate with this sort of threshold. Let’s introduce the term “besmirched”: An agent’s belief is

---

23 I’m indebted to Ernie Sosa for suggesting a principle like this and to Laura Callahan for discussion on this issue.
besmirched when that agent has any negative, higher-order, epistemic appraisal about that belief. Here are some examples of besmirched beliefs that $p$:

1) S believes that $p$ even though they believe it’s irrational to believe that $p$.
2) S believes that $p$ even though they believe it’s bad to believe that $p$.
3) S believes that $p$ even though they believe it’s sorta silly to believe that $p$.

It’s easy to have besmirched beliefs. The slightest stain besmirches. Not all kinds of besmirchment are serious. S needn’t be terribly worried when they discover that their belief that $p$ is sorta silly—that may even be the epistemically appropriate combination of attitudes to have given their epistemic position (just as for “irrational” and “problematic”). Nevertheless, there’s something unfortunate about besmirched beliefs. The best of the best beliefs are unbesmirched: they aren’t accompanied by any negative, higher-order, epistemic appraisals.

Here’s another way for S to have a besmirched belief:

4) S believes that $p$ even though they believe it’s besmirched for them to believe that $p$.

After all, being besmirched is one way a belief can be attributed a negative, higher-order, epistemic appraisal. So the following thesis is true:

**Easy Besmirchment:** Necessarily, (for all S and all $p$) if S believes that it is besmirched for S to believe that $p$ then it is besmirched for S to believe that $p$.

After all, surely an unbesmirched belief would be unaccompanied by any negative, higher-order epistemic appraisals. Recall that we stipulated that the slightest stain besmirches: being more besmirched (than whatever) entails being besmirched outright. So Easy Besmirchment escapes our third objection.

Suppose an agent knows that $p$ and knows that Easy Besmirchment is true. They consider whether to believe, in addition, that it’s besmirched for them to believe that $p$. They ask themselves, “Suppose I were to believe that it is besmirched for me to believe that $p$. Would that belief be true?”

Absolutely! For according to Easy Besmirchment, believing that it’s besmirched to believe that $p$ is enough to make believing that $p$ besmirched. Simply having the belief makes it so. And since Easy Besmirchment is a necessary truth, it’s impossible
that the belief could be false. The agent knows that, necessarily, if she herself were
to believe that <it is besmirched for me to believe that p> then she would truly
believe that <it is besmirched for me to believe that p>.

The doctrine of infallibility faces a problem: it insists that we are permitted to
believe that our beliefs are besmirched just because believing it would make it so.
But taking this path makes our total epistemic state worse even if we acquire true
beliefs in the process. We’re not rationally permitted to needlessly downgrade our
beliefs in this way.24

Is there any other way out? Not, I think, if we’re committed to the quality of our
first-order beliefs being at least somewhat impacted by our higher-order beliefs
about them. But perhaps it’s possible to deny this while preserving the spirit of the
reflectivist insight that our higher-order beliefs matter on our epistemic record. One
could insist that what is impacted by negative, higher-order epistemic appraisals is
not (necessarily) the corresponding first-order beliefs themselves but the whole
network of beliefs. This is our fourth strategy to avoid the problem of easy
downgrade. For example, suppose I believe without any reason apart from it’s
presumed infallibility that it’s problematic (or besmirched, etc.) for me to believe
that p. What’s made problematic on this new way of thinking is not my belief that p
itself but my network of p-related beliefs. The damage done is wholly holistic. 25

This is, in some ways, an attractive position. What’s bad about believing both <p>
and <it’s problematic for me to believe that p>. Part of the answer is that the beliefs
do not mesh together as well as they might. There’s tension. When the agent has
very good reason to believe both, the best choice might be to live with that tension,
but there’s tension just the same. The tension metaphor suggests that the problem
is with how beliefs (or potential beliefs) fit together and not, in the first instance,
with the beliefs themselves. This is at least suggestive of the view above that the
thing damaged in easy downgrade cases is a belief system and not necessarily any
particular belief.

On this view, propositions like <it is problematic for me to believe that p> are not
infallible: the believer will always be (partially) right that there is a problem
somewhere, but they aren’t guaranteed to be (fully) right about where the problem

24 We’re now in a position to make good on the promissory note in fn. 15. Recall that we considered
a version of the doctrine of infallibility that said it was always permissible to hold onto known,
guaranteed truths but not always permissible to acquire them. Suppose that someone’s beliefs are
perfect except that, for no good reason, they believe that their belief that p is besmirched. Since
nothing is making their belief in p besmirched except for their higher-order attitude (that it is so
besmirched), they ought to discard this belief. They are not permitted to hold onto the belief that p
is besmirched for them, even though they are guaranteed to be right in holding onto it.
25 I’m grateful to conversation with Matt McGrath on this point, both for helping me to articulate the
problem and the solution.
is. The problem would infect the agent’s $p$-related belief system, not the $p$-belief itself.

But the objection can be overcome by slightly altering the example. For suppose, on this view, I form the belief: <it is problematic for me to have $p$-related beliefs>. I thereby make it so that there is a problem somewhere affecting my $p$-related beliefs, even if there isn’t a problem for any particular one of them. But notice that the revised proposition under consideration is silent about whether there’s a problem for any particular proposition. It only says that having a $p$-related belief would create a problem for me, somewhere. So the problem for infallible, easy downgrade propositions arises even if what is downgraded are belief systems rather than beliefs.

There’s a tension, therefore, between epistemic reflectivism and epistemic veritism that is unearthed by the dispute over the doctrine of infallibility. If reflectivism is true, then acquiring certain higher-order beliefs can make our first-order beliefs worse in some way—and this way need not involve making our first-order beliefs false. So we shouldn’t form such beliefs unless we have to. But if we should only care about accuracy, then we are permitted to form these problematic higher-order beliefs anyway, so long as we’re guaranteed to be right about them.

One could be complacent about this result. Beliefs are easily besmirched. And we’re far from ideal agents. It might turn out that the vast majority of our beliefs are already besmirched whether we believe that they are or not. And if so, why worry that the doctrine of infallibility permits us to believe that our beliefs are stained in precisely the way they already are?

But such complacency is unmerited. First, if the doctrine of infallibility were true, then even perfect knowers—oracles, supercomputers, gods—would be permitted to believe (truly, once believed) that their beliefs were besmirched. But surely such powerful agents would not have antecedently besmirched beliefs, even if we were mortals often do.

But second, when one believes, e.g., that <it is problematic to believe that $p$> solely because that proposition is infallible, the proposition is not made permissible thereby. The doctrine of infallibility is supposed to be an explanatory thesis: S’s knowing that <necessarily, if she herself were to believe that $p$ then she would truly believe that $p$> explains why S is permitted to believe $p$.\(^{26}\) Guaranteed accuracy explains permissibility. But if the doctrine of infallibility only comes out true because we are permitted to believe some infallible propositions for reasons unrelated to

\(^{26}\) The explanatory ambition is represented by the “thereby” in our formulation of the doctrine.
their guaranteed truth (e.g. because our beliefs happen to already be besmirched for other reasons), then infallibility does not play the explanatory role that we first thought. The problem is not that we shouldn’t believe our beliefs are besmirched (for all I’ve said, epistemic humility nearly always demands this of us!) but that we clearly shouldn’t do so just because we’d be guaranteed to be right. Either way, the chain linking guaranteed truth with permission is severed.

5. The Direction of Fit Solution

The problem of easy downgrade shows that the doctrine of infallibility is flawed. Given how unassailable the doctrine appeared at the start, this itself is a significant conclusion. But those sympathetic to the doctrine may hope that its flaws can be mended—or at least safely ignored in most contexts. In particular, it’s noteworthy that the counterexample involves a reversal of the ordinary direction of fit between mind and world. Our aim in believing is (at least in part) to form a representation in our mind that appropriately matches the outside world—to tailor our minds to fit the world. Ordinarily, the world cares little how our mind represents it. But not always. Sometimes, the world tailors itself to fit our representation. And indeed, if any of the defeat principles proposed are true, the quality of our epistemic states depends in some part on how we represent those states to ourselves. Perhaps, then, even though the doctrine of infallibility is false, we can treat the doctrine as true when there’s no reversal of the ordinary direction of fit between mind and world. We replace the original doctrine with this revised principle:

**Doctrinal of Infallibility, Fit Edition:** (For all S and for all p) If p bears the normal world-to-mind fit relation to S, then if S knows that <necessarily, if she herself were to believe that p then she would truly believe that p>, then it is thereby (rationally) permissible for S now to believe that p.

The ambition behind this strategy is a sensible one. It seems that there is something right about Descartes’s appeal to infallibility in the *Meditations*, and given that the doctrine of infallibility is false, epistemologists should be eager to find a more restricted principle that allows the Cartesian inference while avoiding the problem of easy downgrade.

But the problems with the doctrine cannot be excised by merely restricting its domain to cases with a normal direction of fit. For we need the doctrine of infallibility (or something that entails it) when dealing with other cases with a reversal of fit. The revised principle does not neatly divide the good cases from the bad. Recall the coin-flip case. In these cases, the agent knows that they will be right whether they believe that the coin will land heads or that the coin will land tails, and the fact that the agent knows this seems to license belief in either proposition. As
noted earlier, the most straightforward explanation of this seems to be the doctrine of truth, which entails the doctrine of infallibility. But the coin-flip cases themselves have a reversed direction of fit! The agent’s belief—through the demon’s intervention—causes the world to match the belief!

This frustrates our attempt to save the doctrine of infallibility by restricting its scope to situations with a normal direction of fit. The restriction is too severe, for it’s precisely when there’s a reversal of direction of fit that the doctrine of truth seems especially appealing. But the doctrine of truth entails the doctrine of infallibility, which we’ve already seen is false! The puzzle remains.

6. The Disbelief Solution

Perhaps we’ve been focusing too much on aiming for the good outcome of a true belief and not enough on the bad outcome of believing a falsehood. With this thought in mind, we recall that the cogito has not just one but two things going for it. First, as we’ve noted, if one believes that <I exist> then one is guaranteed to be right. But equally, if one believes its negation, <I don’t exist>, then one is guaranteed to be wrong (Sosa, 2007: 18).

Moreover, this distinguishes the cogito from the easy downgrade propositions of this paper. Notice that <it is besmirched for me to believe that > is not such that if I disbelieve it I am guaranteed to be wrong.

This suggests a new doctrine:

**Doctrine of Infallibility, Disbelief Edition:** (For all S and for all p) If S knows that <necessarily, if she herself were to believe that p then she would truly believe that p> and S knows that <necessarily, if she herself were to believe that ¬p, then she would falsely believe that ¬p, then it is thereby (rationally) permissible for S to believe that p.

This successfully saves the cogito without endorsing besmirching propositions. But much like the attempted fix by fit, it leaves unexplained why it’s permissible to believe in cases when one knows one would be right either way. In Dahlback’s case, I can believe that the coin will land heads—seemingly, just on the basis that I know I will be right—even though I will also be right if I disbelieve that the coin will land heads.

---

27 Or at least is a causal factor
So although Sosa’s insight tells us something important about the *cogito*, this revised principle cannot neatly divide all cases of good infallible propositions from the bad.²⁸

7. The Grounding Solution

Perhaps this is what is different between the *cogito* and easy downgrade propositions: only believing the latter ground their own truth.²⁹ When I believe, for instance, that <it is besmirched for me to believe that p> my very believing (at least partly) explains why it is besmirched for me to believe that p. In contrast, that I believe <I exist> does not explain (even partly) the fact that I exist. The opposite is true.³⁰ Perhaps, then, the correct version of the doctrine of infallibility reads thus:

**Doctrine of Infallibility, Grounding Edition** (For all S and for all p) If believing that p would not partially explain the truth of p, then if S knows that <necessarily, if she herself were to believe that p then she would truly believe that p>, then it is thereby (rationally) permissible for S to believe that p.

Does this get the right verdict in the coin-flip case? It perhaps depends on how we think of the grounding relation. If the demon is basing the result of the coin flip on my belief, then my belief is part of the causal explanation. So however we conceive of the grounding relation, it had better not be in a way that subsumes causal explanation.

This in and of itself is not a problem. It’s quite plausible that there’s a unified notion of grounding that includes, e.g., constitution and truthmaking but not causation.³¹ But it is unclear why, epistemically speaking, we should care about the difference. In the coin-flip case, my belief that the coin will land heads, holding fixed the friendly environment (i.e. the demon who will make the world match my beliefs), causally guarantees that my belief will be right. But what I care about, surely, is that my belief guarantees the truth not whether it does so causally.

Moreover, there seem to be cases in which a belief grounds its own truth in ways that are unproblematic: indeed, there is just such a case in the *cogito*.³² For before Descartes concludes that <I exist>, he concludes that <I am thinking>. But <I am

²⁸ I am deeply indebted to Ernie Sosa for conversation on this point.
²⁹ I’m grateful to Ezra Rubenstein for raising a version of this possibility.
³⁰ Thus the old joke:
   Bartender: Another beer?
   Descartes: I think not. *vanishes into thin air*
³¹ Perhaps the former are essentially synchronic whereas causation is (at least normally) diachronic, for instance.
³² Thanks to Ernie Sosa for pointing out the relevance of the *cogito* here.
thinking> is the sort of proposition whose truth is (at least partially) grounded in the belief itself. If I believe <I am thinking>, my very belief grounds the truth of the believed proposition. Moreover, it is permissible to believe <I am thinking> even if <I am thinking> is the only thought one is having at the moment, so the belief could even be the full grounds for the truth of the proposition.

Once again, our revised version of the doctrine of infallibility does not neatly distinguish the good cases of infallible propositions from the bad. For it excludes <I am thinking>, a paradigmatically good infallible proposition.

8. Veritism and Reflectivism Revisited

We've encountered a puzzle. The doctrine of infallibility seems overwhelmingly plausible. It is the basis for the cogito, it makes sense of epistemology when the normal direction of fit between mind and world is reversed, and it gives voice to the enticingly straightforward thought that accuracy is what matters most (epistemically) when deciding what to believe.

But the problem of easy defeat shows that the doctrine of infallibility is false. If reflectivism is true—if reflective, higher-order beliefs matter for assessing the quality of their first-order counterparts—then there are some guaranteed truths one should refrain from believing.

Does the falsity of the doctrine of infallibility lead to skepticism? Perhaps if one comes into the problem in a Cartesian mood. If one can’t automatically trust even infallible propositions (the disillusioned Cartesian asks), what can we trust? But most epistemologists have (wisely) not demanded that some of our beliefs must be infallible to count as knowledge. Indeed, fallibilists—who already believed that infallibility was not necessary for right belief—may feel justly emboldened knowing that it isn’t sufficient for permissible belief either. Infallibility isn’t always worth having even when it can be had.

No, the primary puzzle is not, “How can we really know that we exist or know that the coin will land heads if the doctrine of infallibility is false?” We were (rightly) pre-theoretically confident that these were good judgments, and we need not abandon them just because the principle that we thought explained their permissibility turned out to be false. Rather, the puzzle is how to separate the good infallible propositions from the bad. We’re left to wonder: why wasn’t the guarantee of truth good enough to license belief? What was so valuable, epistemically, that it was worth foregoing a guaranteed true belief?

What do we learn from this puzzle?
First, there’s something epistemically good about recognizing the quality of one’s first-order beliefs, something over and above the addition of a true, higher-order belief that such recognition entails. And there’s something epistemically unfortunate about believing that one’s first-order beliefs are lacking, even if that higher-order belief turns out to be true. We might call this epistemic value *reflective appreciation*.

Compare these two agents: A and B both have true and unbesmirched beliefs that \( p \). And so far neither has any higher-order beliefs about \( p \). Then A (impermissibly) adds the belief that it is besmirched for them to believe that \( p \) for no reason other than it’s infallibility. In doing so A has gained a true belief and (assuming, as is surely possible, that A hangs onto their belief that \( p \) without losing any other true beliefs or gaining any false ones. A has improved their overall accuracy.

B instead adds the belief that it is unbesmirched for them to believe that \( p \). Since we stipulated that this was so, B also adds one true belief without giving up any other true ones or acquiring any false beliefs. B has improved their overall accuracy to exactly the same degree.

We would all rather be like B than A. But this preference can’t be explained by appealing to accuracy, for B’s beliefs are no more accurate than A’s. The difference is that B is better able to reflectively appreciate their first-order success with respect to believing that \( p \). Of course, A’s reflective assessment is no less accurate than B’s. But B’s assessment is less appreciative: it does not allow A to evaluate or enjoy their first-order epistemic success in the same way. And this is so even if A and B both know full well that their beliefs that \( p \) are true: what is appreciated is not mere truth-having. That reflective appreciation is epistemically worth having even when accompanied by no comparative advantage in accuracy makes trouble for straightforward versions of epistemic veritism.\(^3\) Whatever James’s aphorism suggests, accuracy—believing truly and avoiding falsehood—are not the only aims governing our epistemic life.

Second and relatedly, we learn that there’s tension between veritism and reflectivism. Perhaps this shouldn’t surprise us so much in the end. Veritism says that what matters most at bottom is accuracy. Reflectivism says that it also matters how our first-order beliefs and higher-order beliefs fit together. Nevertheless, one might

\(^3\) Of course, not all veritists are flat-footed veritists. Consider, e.g., Sylvan (2018) whose form of veritism admits derivative but non-promotional goods such as respecting the truth. By making space for epistemic goods that are derivative of but not promoting of truth, Sylvan’s view does not as easily run into trouble with the discovery that certain ways of promoting maximal accuracy are impermissible. By being liberal with the number of epistemic goods that are derivative of truth, Sylvan’s view aligns more closely with epistemic value pluralisms.
have thought that the reason it matters how our different levels of beliefs fit together is because such relationships can help us to become more accurate. But although this might be part of the story, it can’t be the whole story. For believing easy-downgrade propositions on the basis of their infallibility guarantees accuracy while damaging reflective fit.

Veritism and reflectivism aren’t outright incompatible: the veritist could just bite the bullet. On this view, we really are permitted to believe, e.g., <it’s besmirched for me to believe that ρ> solely on the basis of the proposition’s infallibility (or, if we’re careful, its infallibility plus the fact that believing it won’t cause us to forego more true beliefs or acquire false ones). After all, doing so (usually) improves (or at least does not hurt) our overall accuracy! But this is a hard tack to take. It seems unnecessarily self-defeating to believe that one’s beliefs are besmirched or problematic just because believing it makes it so.

Those committed to veritism should, I suggest, make a different move: reject reflectivism. Adopt extreme level-splitting to save veritism from the problem of easy defeat. Our puzzle began on the assumption that negative higher-order beliefs have some impact on the overall quality of their corresponding first-order beliefs. Granted this assumption, we saw it is very hard to avoid the conclusion that there are some infallible, easily-defeating propositions, even amidst a healthy skepticism about our ability to accurately characterize the epistemic quality of our first-order beliefs.

The thesis that negative higher-order beliefs have no impact at all on the overall quality of their corresponding first-order beliefs is surprising: intuitively, our assessment of a belief makes a difference to its epistemic status for us. But then veritism is an elegant, compelling vision of epistemic value. We should at least feel the temptation to adopt extreme level-splitting to save veritism. Once we recognize that accuracy—getting truths and avoiding falsehoods—is, fundamentally, all that matters (says the committed veritist), we should just give up on the idea that our higher-order beliefs make a real impact on the quality of our first-order ones. Looking for fit between higher-order and first-order beliefs may have a heuristic function, but it confers no real, epistemic value.

That’s what it would take to hang onto the Jamesian picture. I find the price too high. Higher-order assessments of a belief really do have an impact on the quality of their corresponding first-order belief—even if sometimes in modest ways. How well you’ve done in forming a belief depends in some way on how you evaluate the belief you form. In the ideal case, one should take one’s reflective perspective into consideration when forming first-order beliefs.
If we hang onto reflectivism, we must give up the idea that truth-getting and falsehood-avoiding are the only ultimate guides in epistemology. The problem of easy defeat shows that accuracy isn't the only thing that matters epistemically: Reflective appreciation matters too. Sometimes, one shouldn't believe even when one knows one would be right.

Bibliography


James, William (1907). The Will to Believe and Other Essays in Popular Philosophy, Longmans Green and Co.


