Disagreement, Reliability, and Resilience

John Pittard
November, 2015

Note: This is a pre-refereed version of an article forthcoming in Synthese. Please cite the published version, which contains non-trivial changes.

Abstract: Alex Worsnip has recently argued against conciliatory views that say that the degree of doxastic revision required in light of disagreement is a function of one’s antecedent reliability estimates for oneself and one’s disputant. According to Worsnip, the degree of doxastic revision is also sensitive to the resilience of these estimates; in particular, when one has positive “net resilience,” meaning that one is more confident in one’s estimate of one’s own reliability than in one’s estimate of the disputant’s reliability, less doxastic revision is required. I show that Worsnip’s Resilience Account, however intuitive it may be, sometimes issues prescriptions that are clearly irrational. I then argue that Worsnip’s criticisms of “extreme conciliationism” are mistaken. The discussion brings out several important lessons for the epistemology of disagreement: first, while positive net resilience does not affect the degree of conciliation required in one-shot disagreements, over multiple disagreements it may diminish or magnify the required degree of conciliation; second, a common way of framing the disagreement debate is misguided; and third, the focus of the disagreement debate should not be on whether reliability estimates should determine the degree of conciliation (they should), but on what reasons may legitimately ground reliability estimates.

1 INTRODUCTION

In a recent paper, Alex Worsnip (2014) contends that philosophers on various sides in the debate over the epistemic significance of disagreement differ very little in their judgments concerning the appropriate responses to concrete cases of disagreement. Where these philosophers do differ, according to Worsnip, is on the question of what theoretical account of disagreement best predicts when significant doxastic revision is required in light of disagreement and when it is not. Worsnip hopes to help resolve this theoretical dispute by offering his own novel account of the evidential significance of disagreement, an account that he thinks is the logical point of convergence once suitable qualifications have been

1 For what it is worth, I do not share the view that epistemologists essentially agree on how we should respond to concrete cases of disagreement. It is true that there is broad agreement on the appropriate response to most of the cases that are discussed in the disagreement literature. But this is a product of selection bias. Most of the cases discussed in the literature are cases that are offered as evidence for one or another position; and of course a case can serve this dialectical role only if all sides agree on the appropriate response. Thus, cases discussed in the literature are typically selected partly because the appropriate response is rather clear. This does not give us good reason for thinking that the appropriate response is clear in most cases of disagreement. And it may be the case that in the many less clear-cut cases, those on opposing sides in the disagreement would be inclined towards significantly different verdicts.
added to “extreme conciliationism” (which holds that one should always “split the
difference” between one’s initial credence and the initial credence of an “epistemic peer”) and to “extreme steadfastness” (which holds that peer disagreement does not provide any reason to adjust one’s credences). The crucial claim of Worsnip’s account is that in a disagreement with an epistemic peer, the degree to which you should conciliate (that is, revise your credence in the direction of your peer’s) depends on the respective resilience of your reliability estimates for your disputant and for yourself. In cases where you are much more confident in your estimate of your own reliability than you are in your estimate of your disputant’s reliability, little conciliation will be required. This is true even if you initially had the same reliability estimate for yourself and for your disputant.

Worsnip’s proposal has a significant degree of intuitive plausibility. But in this paper I aim to show (through what I take to be a conclusive counterexample) that the proposal is misguided. Differential resilience in reliability estimates has no bearing on how much one should conciliate in the face of a one-shot disagreement.2

My discussion will proceed as follows. In section 2, I describe the cases that Worsnip cites as motivation for his proposal and describe this proposal in detail. In section 3, I develop a counterexample that shows that comparative resilience (in Worsnip’s sense) does not modulate the evidential import of a one-shot disagreement as Worsnip supposes. In section 4, I explain why the case that Worsnip offers as a conclusive counterexample to extreme conciliationism is no such thing. One conclusion from this section is that disagreements over matters that one is extremely confident in will almost never qualify as peer disagreements in the relevant sense. In section 5, I address an incoherence charge that Worsnip brings against extreme conciliationism. Finally, in section 6, I draw out some lessons from this discussion for the larger disagreement debate: first, while positive net resilience does not affect the degree of conciliation required in one-shot disagreements, over multiple disagreements it may diminish or magnify the required degree of conciliation; second, a common way of framing the disagreement debate is misguided; and third, the focus of the disagreement debate should not be on whether reliability estimates should determine the degree of conciliation (they should), but on what sorts of reasons may legitimately ground reliability estimates.

My aim here is not to defend extreme conciliationism. (I do not in fact endorse it.) Rather, it is to show that Worsnip has not identified any problem with extreme conciliationism, and that it would be a mistake to follow his “moderate theory of how one should respond to disagreement” (2).3

2 WORSNIP’S PROPOSAL AND ITS MOTIVATION

To characterize various positions on how to respond to disagreement, Worsnip introduces the useful notions of “net reliability” and “net resilience,” notions that in turn depend on the notions of reliability and resilience. In this context, an agent’s reliability is defined as her propensity to get close to the credence that is ideal given her evidence. Worsnip sets aside the question of how exactly this propensity is to be measured. Worsnip then defines net reliability as the difference between your reliability and the reliability of your interlocutor. If

---

2 Though, as I will explain later and as White (2009) has shown, disparities in the resilience of these reliability estimates can mitigate or magnify the degree of conciliation required over multiple disagreements with a single individual.

3 Here and throughout the rest of the paper, page numbers in parentheses refer to the relevant page in (Worsnip 2014).
you are more reliable than your interlocutor, net reliability is positive; if your interlocutor is more reliable, net reliability is negative. In the special case where your interlocutor is your *epistemic peer*, your net reliability is 0. Often, one will not know the reliability of oneself or one’s interlocutor, but one may still have an *estimated* reliability for each participant in a disagreement. Suppose, for example, that a typical person’s reliability in some domain is 0.8. If I have no reason that would justify me in thinking that I am more reliable than average and also do not have any reason for thinking that I am less reliable than the average, I should presumably estimate my reliability at 0.8. Still, I might think it highly unlikely that my *actual* reliability is 0.8. Reliability estimates can also give us a net reliability *estimate*. For example, if I estimate my reliability at 0.8 and my disputant’s reliability at 0.8, then my net reliability estimate will be 0 even if I think that it is highly unlikely that our *actual* net reliability is 0.

With the notion of estimated reliability in hand, we can now introduce the notion of resilience. **Resilience** is a measure of the stability of your estimate of someone’s reliability in the face of future evidence. More confident reliability estimates, especially those based on a significant track record, will be more resilient. Reliability estimates based on mere hunches or very thin evidence will be more susceptible to change, and thus less resilient. Finally, Worsnip defines **net resilience** as the difference between the resilience of your estimate for your own reliability and the resilience of your estimate for your interlocutor’s reliability. Positive net resilience indicates that your confidence in your estimate of your own reliability outstrips your confidence in your estimate of your interlocutor’s reliability. Importantly, net resilience need not correlate with net reliability. If I know on the basis of a substantial track record that my reliability in some domain is 0.7, and I also know that this happens to be the population average (though there is significant variability among individuals), then my net reliability in some disagreement with an arbitrarily selected person (about whom I know nothing) may approximate 0 (since 0.7 is my best reliability estimate for an arbitrarily selected individual), but my net resilience will be quite high. After all, I could easily learn something that would drastically shift my estimate of this person’s reliability.

Having defined these terms, we are in a position to understand Worsnip’s discussion of his own view and some alternatives. Worsnip characterizes two extreme positions, put forward as foils to his own moderate alternative, in the following way:

**Extreme conciliationism:** Whenever you encounter a disagreement with someone who possesses the same evidence as you, you are required to adjust your credence in the proposition at issue in proportion with your prior net reliability estimate and your relative credences. In particular, if your net reliability estimate is 0, and your credence in p is equal to your interlocutor’s credence in not-p, then you are required to adjust your credence to 0.5.

**Extreme steadfastness:** Encountering disagreement, at least with anyone who you do not estimate as more epistemically reliable than you, is never in itself a reason to adjust your credences. (6)

Worsnip contends that both of these extreme views are obviously incorrect (a claim that he thinks would be affirmed by nearly all contributors to the disagreement literature), and

---

4 As Worsnip notes, peerhood is characterized differently by different thinkers writing on disagreement. I agree with Worsnip that peerhood is most usefully defined in terms of the disputants’ reliability (rather than in terms of intellectual virtues, say).
supports this point with counterexamples to each view that in each case he takes to be decisive. First, here is the counterexample he offers to extreme conciliationism:

**LIZARDS:** At a philosophy conference in Washington DC, I meet David. Considering myself to be a roughly average epistemic agent amongst philosophers, and David to be a randomly sampled philosopher, I am initially inclined to make a net reliability estimate of 0, though of course my net resilience is high, since I have very little to go on about David so far. After a bit of idle chatter about the weather and the conference program, I mention that on my off-day I am planning to go and see the White House. “Don’t go there!” exclaims David. “The US government is run by a sinister race of disguised lizards. All the major governments of the world are.”

(6)

Since it is obvious that one is not rationally required in this situation to adopt a credence of 0.5 for the lizard hypothesis (even if David is exceedingly confident in the hypothesis, and even there is no difference in the evidence possessed by David and oneself), Worsnop takes LIZARDS to be a successful counterexample to extreme conciliationism. Next, as a counterexample to extreme steadfastness, Worsnop offers a version of a prominent case from (Christensen 2007, 193). Here is Worsnop’s case:

**DINNER CHECK:** I go out for dinner with seventeen friends. At the end of the meal, my friend Cat and I both calculate what our third friend, Alfonso, owes. It is a slightly tricky calculation, but not one beyond either of our powers. I take Cat to be equally reliable as me at mental arithmetic, on the basis of numerous similar past occasions. I come up with $21.74, and form credence 0.9 in the proposition that Alfonso owes $21.74. Next I discover that Cat has credence 0.9 that the amount Alfonso owes is $22.74. (6)

Since it is obvious that a reduction of confidence is appropriate in DINNER CHECK, Worsnop suggests that we have good reason to reject extreme steadfastness.

Taken by itself, someone might take DINNER CHECK to support the conclusion that the degree of doxastic revision required in a disagreement is largely or wholly determined by the comparative estimated reliability of the disputants, in other words, by estimated net reliability. It seems, after all, that it is Cat’s reliability that generates the conciliatory pressure in this case. But Worsnop thinks that LIZARDS shows this to be false. Despite initially equivalent reliability estimates, no significant conciliation is required in LIZARDS.

Worsnop proposes a theory of disagreement that he thinks gives us a unified account of these two cases. Worsnop’s central claim is that significant doxastic revision is required in a disagreement to the extent that there is a low estimated net reliability and low net resilience. What generates the conciliatory pressure in DINNER CHECK is not just the fact that Cat was estimated to be highly reliable (resulting in low estimated net reliability), but also the fact that this high reliability estimate was based on significant track record data (resulting in low net resilience). The role of net resilience helps to explain the absence of any significant conciliatory pressure in LIZARDS. While the speaker’s reliability estimate for David was initially no different than the speaker’s estimate of his own reliability (resulting in an estimated net reliability of zero), his estimate for David was based on an extremely thin evidential base, resulting in high net resilience. This high net resilience explains why we are inclined to judge that the disagreement has little evidential import.
Let’s call Worsnip’s view the “Resilience Account.” The Resilience Account not only gives us a unified account of these two cases, but there is also an intuitively plausible explanation for why net resilience should modulate the degree of doxastic revision in the face of disagreement. When you find yourself in a case like LIZARDS, there is a tension between your high degree of initial confidence in the proposition under dispute and your initially high estimate of the reliability of your disputant. It seems that in light of this tension at least one of these commitments must weaken, and there is no obvious reason why it must be your confidence in the disputed proposition that must “give way” to your high estimate of our disputant’s reliability rather than the other way around. Intuitively, the “weaker” of these commitments should be the one that is more significantly revised, and the commitment that is less resistant to counterevidence—in other words, the commitment that is less resilient—would seem to be weaker in the relevant sense. Thus, we should expect that peer disagreement gives us less powerful evidence against our view in situations where our initial estimation of our disputant’s reliability is not especially resilient.

3 A COUNTEREXAMPLE TO THE PROPOSAL

However intuitively appealing the Resilience Account might be, it delivers clearly irrational prescriptions in many cases and should for this reason be rejected. Here is one case that the Resilience Account mishandles:

**ESTIMATION:** “Estimation Game” is a computer game that works as follows. For a few seconds, the player views a photograph of a glass jar full of objects (marbles, sardine cans, whatever). Then the image disappears and the player has a chance to answer a true/false question about the number of objects in the jar, such as: “This jar contains more than 123 thimbles—true or false?” The correct answer is “true” just half of the time. The questions are calibrated so that they are all the same difficulty. As it turns out, 30% of adults have an 80% accuracy rate when playing Estimation Game; 40% of adults have a 90% accuracy rate, and 30% of adults have a 100% accuracy rate. These accuracy rates are stable and do not improve with practice. Additionally, there is no way to tell which group one is part of by how things seem when answering questions in the game. This can only be discovered by confirming one’s track record. Finally, whether someone misses one question is probabilistically independent from whether someone else misses that question (this means that the fact that I missed a question has no evidential bearing on whether someone else will miss that question). I know all this. I also know that I am one of the 90% accurate players. Now, in a philosophy experiment I am asked to play Estimation Game. In another room there is a randomly selected adult who is also playing the game, and who is seeing and answering the same questions. This subject knows the same facts I know about the game, population accuracy rates, and so on. But unlike me, this other subject has never played the game before, and does not yet know his or her accuracy rate. After I arrive at my answer, I will be told what credences the other subject assigned to the two possible answers. I look at the first image, read the question, and judge that the correct answer is probably *True*. I assign *True* a credence of 0.9. I am then told that the other subject disagrees; this subject answered *False*, and assigns *True* a credence of only 0.1. I am also told that there is nothing unusual going on in the other room—the subject is not distracted, is not feeling unusually confident, etc.
Let’s assume that the correct answer is True, but that there is not anything unusual about this particular occasion that leads me to think that I am more or less reliable than is typically the case. (The question does not look unusually obvious, I do not feel more confident than I normally do in answering a question in Estimation Game, and so on.) Given this assumption, most everyone would affirm that it is completely reasonable for me to have a credence for True of 0.9 prior to learning what the other subject thinks. After all, in situations relevantly like this, I am right 90% of the time. But what should my credence be after learning that the randomly selected player assigns True a credence of 0.1?

Let’s first consider the verdict supported by extreme conciliationism. Because the other subject is randomly selected from the adult population, there is a 30% chance that s/he (I’ll use “she” from here on out) is 80% reliable, a 40% chance that she is 90% reliable, and a 30% chance that she is 100% reliable. Taking a weighted average of these leads to an expected reliability of 90%. Since I know that my own reliability is 90%, this means that the net reliability estimate in this case is 0. And extreme conciliationism says that when my net reliability estimate is 0 and my disputant is equally confident in the opposing view, my credence after learning of the disagreement should be 0.5.

Extreme conciliationism gives the right answer in this case. This judgment can be supported by appealing to Bayesian conditionalization, but a more intuitive justification may be more persuasive. Suppose I repeated this experiment one thousand different times, with one thousand different subjects, none of whom have played Estimation Game. The following table shows the expected values for how many subjects of each type I would face off against, how many disagreements I would encounter, and how many of these disagreements would be ones where I am right and how many would be ones where I am wrong:

---

5 Because Worsnip contests the standard way of applying conditionalization in this sort of case (11-12), it is dialectically more effective to give an intuitive justification of the verdict of Extreme Conciliationism. But here is how one would reach this result via conditionalization. To simplify, let’s imagine that I am told first that the other subject has a credence of 0.9 for either True or False. This shouldn’t change my estimation of her reliability. We stipulated that she does not know her own accuracy rate, and that she knows that the best guess of her accuracy is 0.9 (and, further, that how things seem to her while answering the question would provide no evidence of her reliability). Given this, I should expect that if she is reasonable she will assign a credence of 0.9 to her answer, however reliable she may in fact be. (We can assume that her being reasonable in this way does not correlate with any particular accuracy rate in Estimation Game.) So updating on this information changes neither my estimation of her reliability nor my credence for True. Next, I learn D (for ‘disagreement’), which stands for the proposition that she thinks that False is the correct answer. According to conditionalization, my new credence for True, should be equal to the conditional credence for True given D that I had prior to learning D. In other words, it should be the case that \( C_{view}(True) = C(True | D) \), where C is my credence function just before learning D. By Bayes’s Theorem, we know that \( C(True | D) = C(D | True) \cdot C(True) / C(D) \). To calculate the value of \( C(D | True) \), let’s use Low to designate the possibility that the other subject is in the 80% accurate group, Mid to designate the possibility that she is in the 90% accurate group, and High to designate the possibility that she is in the 100% accurate group. Probabilistic coherence requires that C should satisfy the following: \( C(D | True) = C(Low | True) \cdot C(D | True & Low) + C(Mid | True) \cdot C(D | Mid & Low) + C(Handy | True) \cdot C(D | True & Handy) \). Given the information stipulated in the case, we know each of these values. So \( C(D | True) = 0.3 \cdot 0.2 + 0.4 \cdot 0.1 + 0.3 \cdot 0 = 0.1 \). Calculating \( C(D | False) \) in the same way, we arrive at \( C(D | False) = 0.9 \). By the law of total probability, \( C(D) \) should be equal to \( C(D | True) \cdot C(True) + C(D | True) \cdot C(True) \cdot C(True) / C(D) \cdot C(False) = 0.9 \cdot 0.1 + 0.1 \cdot 0.9 = 0.18 \). Finally, substituting the values into the original equation (from Bayes’s Theorem), we get \( C(True | D) = 0.1 \cdot 0.9 / 0.18 = 0.5 \). So \( C_{view}(True) = 0.5 \).
As the table shows, if I took part in this experiment 1,000 times with 1,000 different subjects, the expected number of disagreements would be 180, and the expected number of disagreements where I have answered correctly would be 90. So the proportion of disagreements where I am right is expected to be 50%. Clearly, then, the reasonable response when I encounter a disagreement in the experiment is to adopt a credence for my answer of 0.5.

The table brings out another important result. I expect that 78 out of 180 disagreements will be with someone from the Low reliability group, 72 of these will be with someone from the Mid group, and only 30 of these will be with someone from the High group. So of those I disagree with, I expect that 43.33% will be 80% accurate, 40% will be 90% accurate, and 16.67% will be 100% accurate. This means that when I learn that a subject disagrees with me, the expected reliability of this subject should drop from 90% (my original estimate) to 87.33% (= 0.4333 · 0.8 + 0.4 · 0.9 + 0.1667 · 1). So after learning that another subject disagrees with me, I estimate her reliability to be lower than mine but nonetheless give her answer equal weight to my own. Worsnip finds this result to be “bizarre at best and incoherent at worst” (11), and I will address the apparent strangeness of this result in the last section. But for now, it suffices to note that however counterintuitive the result might seem, it is clearly right: on average, I will be right in only half of my disagreements, so my credence should move to 0.5 in response to disagreements; and my reliability is higher than the expected reliability of someone selected from the pool of those who disagree with me, so I also should decrease my reliability estimate of the other subject to below 0.9 when I encounter a disagreement. The table decisively vindicates both of these responses, even if we might struggle to intuitively appreciate how both of these responses could simultaneously be correct.

Let’s now consider what the Resilience Account says. According to the Resilience Account, a disagreement where the net reliability estimate is 0 should push me towards an equal weight verdict only if the net resilience is also 0—in other words, only if I am just as confident in my reliability estimate for my disputant as I am in my reliability estimate for myself. If the net resilience is positive, so that I am more confident in my reliability estimate than I am in my disputant’s reliability estimate, then I should give more weight to my own view. In ESTIMATION, my net resilience is positive. I know that my reliability is 90%, but I
don’t know this about the other subject. There is a very significant chance that the other subject could be more reliable than me, but an equally significant chance that the other subject could be less reliable than me. So according to the Resilience Account, my credence for my answer should decrease, but not all the way to 0.5. This is the wrong result. Using the straightforward math that was used to fill out the table above, I can know that in disagreements during the experiment I am expected to be right half of the time. But the Resilience Account says that in this sort of situation my credence should be higher than 0.5, a result that clearly is not rational.

4 Diagnosing the Putative Counterexample to Extreme Conciliationism

The Resilience Account delivers irrational verdicts in cases like ESTIMATION and should for this reason be rejected. Extreme conciliationism delivers the right result in ESTIMATION, and also in DINNER CHECK. But what about LIZARDS? Isn’t Worsnip right that this case gives us a definitive counterexample to extreme conciliationism? And don’t we have to appeal to net resilience in order to explain the intuitively correct verdict in this case? In this section, I defend negative answers to both of these questions: LIZARDS does not give us a counterexample to extreme conciliationism, and we do not need to appeal to net resilience in order to justify the commonsense response.

In describing the LIZARDS case, Worsnip takes it for granted that it is reasonable for the speaker (let’s call him Alex) to have a net reliability estimate equal to zero. The justification for Alex having the same reliability estimate for himself and David is that Alex takes himself to be a “roughly average epistemic agent amongst philosophers” and that David is a “randomly sampled philosopher.” If Worsnip is right in supposing that a net reliability estimate of zero is rational in this case, then he is right to maintain that LIZARDS gives us a counterexample to extreme conciliationism. But if Alex’s having a net reliability estimate of zero would in fact be irrational in this situation, then all we would be entitled to conclude is that when we start with irrational reliability estimates, applying extreme conciliationism can lead to irrational results. And that would hardly give us a reason for rejecting extreme conciliationism. As Worsnip himself rightly says, “if you feed crazy inputs into a rational process, you may get crazy outputs: that is the fault of the inputs, not the process” (17). This means that a successful counterexample to extreme conciliationism must be a case where an agent starts with rational reliability estimates and then reaches an unacceptable result by means of applying extreme conciliationism. I will argue that LIZARDS is not such a case. If Alex starts with a net reliability estimate equal to zero, he starts with an irrational position. On the other hand, if he starts with a rational net reliability estimate and applies extreme conciliationism, then he will end in a reasonable position.

Why is an initial net reliability estimate of zero irrational in the situation described in LIZARDS, contrary to what Worsnip supposes? The answer is that because Alex has an extremely high estimation of his own reliability with respect to propositions like the lizard government hypothesis, factors about David that are unknown could not cause David to be significantly more reliable than Alex takes himself to be (there simply isn’t much room between Alex’s estimated reliability value and perfect reliability), but such unknown factors could cause David to be significantly less reliable than Alex takes himself to be. Unknown factors thus have more potential to hurt David’s reliability relative to Alex than potential to help David’s reliability relative to Alex. For this reason, the fact that there are unknown
factors about David should lead Alex to have an estimate of David’s reliability that is at least slightly lower than Alex’s estimate of his own reliability, meaning that the net reliability estimate should be at least slightly positive rather than zero. And in cases where one starts with an extreme level of confidence, a slightly positive net reliability estimate can almost entirely undercut the evidential significance of a disagreement.

That is the abstract explanation of why Worsnip is wrong to suppose that an initial net reliability estimate of zero would be reasonable in the LIZARDS case. But the point can more easily be appreciated if we consider a concrete example of the sort of unknown factor that should lead Alex to a slightly positive initial net reliability estimate. Let $LG$ (for ‘Lizard Government’) stand for the proposition that all the major governments of the world are run by a sinister race of disguised lizards. Since one might reasonably think that, as a general matter, we are much more likely to be right when we affirm some highly involved and specific hypothesis than when we negate some highly involved and specific hypothesis, let’s even the playing field by supposing that Alex and David have as background knowledge the fact that either $LG$ is true or the standard view (that governments are run by human beings) is true. Given this background knowledge, denying $LG$ is equivalent to accepting an alternative view that is equally specific, thus eliminating any structural reason to suppose that the person who denies $LG$ is more likely to be correct than the person who affirms it. Even given this stipulated background knowledge, Alex’s initial credence for $\sim LG$ is no doubt very, very high. Suppose he is inclined to assign $\sim LG$ a credence of 0.9999999 (one part in ten million shy of certainty), and that he estimates that in situations where he is inclined to assign a credence of this value, he is right 0.9999999 of the time. (More generally, Alex estimates that when he is inclined to assign a very high credence value $x$ to some proposition $p$ in this domain, his judgment that $p$ is true is correct $x$ portion of the time.) So 0.9999999 is also Alex’s estimate of his reliability in this situation, and (let’s suppose) is the value of his credence for $\sim LG$.

Now, keeping in mind that Alex thinks he is completely unremarkable in his reliability with respect to $LG$, how should he estimate David’s reliability? Well, suppose Alex thinks that if David is a regular user of the drug LSD, then David is unlikely to be highly reliable on the question of whether $LG$ is true. More specifically, Alex thinks that conditional on David being a regular user of LSD and his having a view on $LG$ that is just as confident as Alex’s view on the matter, David’s reliability is approximately 0.6. Should this belief about David’s reliability in the event that he is an LSD user affect Alex’s net reliability estimate in this case? Yes, at least if Alex knows that he himself is not a user of LSD but does not know this about David. Given this assumption, Alex’s estimate of David’s reliability will be weighed down as a result of taking into account a possibility (LSD use) that Alex can dismiss when estimating his own reliability. Suppose that Alex estimates the chance that a randomly selected philosopher takes LSD as being 1 out of 10,000. In this case, a rational estimate for David’s reliability with respect to $LG$ can at most be $0.999996 (= 0.6 \cdot 1/10,000 + 1 \cdot 9,999/10,000)$. While this of course is a high reliability estimate absolutely speaking, it is four hundred times farther away from perfect reliability than Alex’s estimate of his own reliability. And it is this proportional distance from perfect reliability that is most significant in determining the

---

6 We could imagine that there is an oracle that David and Alex both trust absolutely, who informs them that one of the two is right about who runs the government.

7 Just to be clear, my aim in stipulating this background knowledge is to make it initially plausible that the initial net reliability estimate should be zero, thus giving LIZARDS the best shot at being a counterexample to extreme conciliationism. I will argue that even with this stipulation, LIZARDS is not a counterexample.
evidential significance of a disagreement. If Alex starts with a credence of 0.9999999 for \( \sim LG \) and conditionalizes on the fact that David reports a confident belief in \( LG \) (a belief report that Alex takes to be 0.99996 reliable), then Alex’s new credence for \( \sim LG \) will be 0.99975.\(^8\) Mildly positive net reliability can drastically limit the doxastic revision that is required in a disagreement where one starts out with an extreme level of confidence.

One way of thinking about what is going on here is that certain circumstances that would result in David’s belief reports being highly unreliable (drugs, mental disorders, pranks) place a ceiling on how high Alex can estimate the reliability of David’s belief reports. Since Alex knows more about his own history and internal states, he can rule many of these possibilities out in his own case, with the result that the ceiling on Alex’s own reliability estimate is higher than the ceiling on his estimate for David. Of course the ceiling on David’s reliability estimate may still be very close to 1. But in contexts where the disputed proposition is something that Alex is extremely confident in, Alex’s estimate of his own reliability may exceed the ceiling on his reliability estimate for David.

Jennifer Lackey (2010, 277) has emphasized the significance of “personal information” (information that we know about ourselves in virtue of our first person perspective) to the epistemology of disagreement, and has noted how asymmetries in personal information can mitigate the pressure to conciliate in certain disagreements. Worsnip agrees with Lackey that asymmetries in personal information can help explain the commonsense verdict in \textit{LIZARDS}, but Worsnip thinks that the asymmetry in personal information limits the conciliatory pressure this case by causing the net resilience to be positive (14-15). Since Worsnip takes it for granted that a net reliability estimate of zero is reasonable, he thinks that the significance of asymmetric personal information must be explained by its influence on net resilience. But as the above discussion makes clear, the asymmetry in personal information rationally demands that Alex have a positive net reliability estimate, and not merely positive net resilience. Thus, \textit{LIZARDS} does not put pressure on us to affirm the Resilience Account. This is a welcome result, since the Resilience Account supports an irrational verdict in estimation.

Some might have this lingering worry: even if unknown factors about David mean that Alex cannot discount certain possibilities where David is abnormally unreliable, isn’t it also the case that, because David is unknown, Alex cannot discount certain possibilities where David is abnormally reliable? If so, then wouldn’t the negative unknowns be balanced out by the positive unknowns, so that net reliability remains at zero? To answer, it is true that there are many circumstances that Alex cannot rule out that would result in David’s being abnormally reliable. But since Alex takes himself to be extremely reliable in this context (with a reliability of 0.9999999), situations where David is abnormally reliable will be situations where he is at most very slightly more reliable than Alex (specifically, at most 0.0000001 more reliable). But situations where David is abnormally unreliable may be situations where he is much less reliable than Alex. So in weighing possibilities of abnormal unreliability against possibilities of abnormal reliability, Alex will be weighing possibilities that involve David

\(^8\) Suppose that Alex learns that David has a view on \( LG \) that is just as confident as Alex’s, but Alex hasn’t yet learned what that view is. Let \( D \) (for “disagreement”) stand for the proposition that David’s view is that \( LG \) is true. Before learning \( D \), Alex’s conditional credence \( C(\sim LG \mid D) \) should, according to Bayes’s Theorem, equal \( C(D \mid \sim LG) \cdot C(\sim LG) \cdot C(D) / C(D \mid LG) \). Based on David’s estimated reliability of 0.99996, Alex should assign 0.99996 to \( C(D \mid LG) \) and 0.000004 to \( C(D \mid \sim LG) \). The law of total probability requires that \( C(D) = C(LG) \cdot C(D \mid LG) + C(\sim LG) \cdot C(D \mid \sim LG) \). Since we also know that Alex’s credence for \( LG \) is 0.9999999, we can make the relevant substitutions and calculate \( C(D) \); it is approximately 0.00004001. We can now substitute to determine that \( C(\sim LG \mid D) = 0.00004 \cdot 0.9999999 / 0.00004001 \approx 0.99975 \).
being substantially less reliable than Alex against possibilities where David is only slightly more reliable. The negative possibilities will therefore do much more to depress the reliability estimate than the positive possibilities will do to push it up.

5 **WORSNIP’S INCOHERENCE CHARGE**

Recall that in estimation, upon learning that the other subject disagrees with my answer, I should reduce my credence for True to 0.5 and also decrease my estimate of my disputant’s reliability to approximately 0.87, so that I now estimate my disputant’s reliability to be lower than my reliability. As already mentioned, Worsnip finds this sort of result to be bizarre and possibly incoherent (11). And it is easy to appreciate his worry here. All else being equal, I obviously should think that more reliable thinkers are more likely to be right than less reliable thinkers. How, then, could it possibly be reasonable for me to give equal weight to my disputant’s view while also estimating my reliability more highly? It is not as though I know of some other consideration that favors my disputant’s view other than the fact that she believes it. And this consideration surely is outweighed by the fact that the contrary view is endorsed by someone whom I estimate to be more reliable (namely, myself).

This objection is intuitively quite forceful. But the intuitive force of the objection merely underscores our susceptibility to error when we engage in “intuitive” probabilistic reasoning. As we’ve seen, a simple calculation of the expected proportion of disagreements where I am right (50%) and a simple calculation of the expected reliability of my disputants in these disagreements (0.87) shows that the objection must rely on faulty reasoning. But beyond showing that the objection is incorrect, can we diagnose where the reasoning goes wrong? I believe that we can. While the disagreement in estimation gives me reason to think that my disputant is more likely to be in the Low group than in the High group (i.e., in the 80% reliable group rather than the 100% reliable group), my reliability advantage over someone in the Low group would provide only a modest reason to favor my position while the reliability advantage over me enjoyed by someone in the High group would provide a conclusive reason to favor their position. In disagreements with someone in the Low group, I am right approximately 69% of the time, while in disagreements with someone in the High group I am right 0% of the time. Learning that I am disagreeing with someone in the Low group should therefore lead me to assign a credence for my answer of 0.69, which only modestly departs from the “equal weight” credence of 0.5, while learning that I am disagreeing with someone in the High group should lead me to assign a credence of 0 to my answer, a much larger departure from the equal weight credence. And the degree to which the Low group and High group possibilities should affect my current credence is a function of the probability of each possibility and of the degree of doxastic revision that would be called for if I learned that the possibility in question was actual. While High group possibility is less probable, this lower probability is offset by the magnitude of the doxastic revision that would be called for should I learn that my disputant is from the High group.

---

9 White (2009, 248–9) shows that Bayesian conditionalization supports the result that Worsnip finds so objectionable. But he does not offer an intuitive explanation of this result.

10 We can show mathematically that, if I start from an updated credence for True of 0.5, the “credal influence” exerted by the less probable High group possibility is equal to the influence exerted by the more probable Low group possibility. Suppose I have a 0.5 credence for True after the disagreement, as I’ve argued is reasonable. If I were to learn that my disputant is from the Low group, my new credence should be 0.692 (since I’m right in disagreements with the Low group 69.2% of the time), which is a credence movement of +0.192 away from my current position. If I were to learn that she is in the High group, then my new credence for True should be 0, a
Even if it is coherent for me to respond to a peer disagreement by giving equal weight to the view of my disputant while henceforth estimating her reliability to be less than my own, it does not follow that it is coherent for me to give equal weight to the view of a disputant who I know to be less reliable than myself. If I were to learn in estimation that my disputant is in the 80% reliable group, I would of course respond by giving a higher credence to my own answer. But as the previous paragraph shows, it is a mistake to think that having a positive estimated net reliability has the same implications for my credences as learning that the net reliability is in fact positive.

6 Lessons for the Larger Disagreement Debate

I will close with some brief reflections on the significance of the foregoing discussion for the larger debate concerning the evidential import of disagreement. I will make three points.

First, the argument of the previous sections should not be understood as showing that net resilience is irrelevant to an account of disagreement’s epistemic significance. While antecedent estimated net resilience has no immediate bearing on a one-shot disagreement, positive or negative net resilience may amplify or mitigate the degree of conciliation required across multiple disagreements (White 2009). Net resilience exercises this influence indirectly through its influence on estimated net reliability. Estimated net reliability is susceptible to change to the extent that net resilience deviates from zero. While changes in estimated net reliability that result from a one shot disagreement do not affect the degree conciliation that is required in that very disagreement, changes in estimated net reliability resulting from other (independent) instances of disagreement or agreement can affect the degree of conciliation required in the present disagreement. The degree of conciliation required in one disagreement will be mitigated to the extent that there are many other disagreements (whether they are temporally prior or later) and will be magnified to the extent that there are few other disagreements. To see why this is the case, imagine that I am to face off against the other contestant in Estimation Game not once but many times in a row. As we’ve already established, if we disagree on the first question, my credence for my answer should be 0.5. But now suppose that we agree on the second question. This agreement is new evidence that favors the hypothesis that the other contestant is from the High group, and in light of this evidence my credence for my first answer should drop below its value of 0.5. In other words, I should conciliate even further. Alternatively, if we disagree on the second question, this is evidence that favors the hypothesis that the other contestant is from the Low group, and this should lead me to increase my credence for my answer to the first question. In other words, I should now be less conciliatory. As I play against this same contestant multiple times, over time the rate of disagreement will converge to 26% (=0.9 · 0.2 + 0.1 · 0.8) if the other contestant is in the Low group; to 18% (=0.9 · 0.1 + 0.1 · 0.9) if the other contestant is in the High group. The rate of disagreement will converge to 26% (=0.9 · 0.2 + 0.1 · 0.8) if the other contestant is in the Low group; to 18% (=0.9 · 0.1 + 0.1 · 0.9) if the other contestant is in the High group. The High group possibility thus exerts equal credal influence even though it is significantly less probable than the Low group possibility, with the result that the credence of 0.5 is a stable equilibrium.
the Middle group; or to 10% if the other contestant is in the High group. In the case where the disagreement rate converges to 26%, my credence for each of my answers in cases of disagreement should converge to approximately 0.69 (the portion of time I expect to be right in disagreements with someone from the Low group); in the case where the disagreement rate converges to 18%, my credence for my answers in cases of disagreement should converge to 0.5 (the portion of time I expect to be right in disagreements with someone from the Middle group); finally, in the case where the disagreement rate converges to 10%, my credence for my answers in cases of disagreement should converge to 0 (since I am never right in disagreements with someone from the High group). What this shows is that in cases where net resilience is not zero, it is possible to learn multiple opinions of someone who was initially an epistemic peer (i.e., someone for whom estimated net reliability was initially zero) and come away giving more weight to your opinions than to the opinions of your interlocutor. But it is also possible to come away giving less weight to your opinions than to your interlocutor’s opinions. When we are learning multiple opinions, Worsnip is right that positive net resilience can make it the case that the appropriate level of conciliation should deviate from the “equal weighting” that might be required by a (crude) “equal weight view” on disagreement. But Worsnip is wrong to think that positive net resilience inevitably means that less conciliation will be required. This might turn out to be the case. But it could equally turn out that a greater degree of conciliation will ultimately be required.

Second, the foregoing discussion (and White’s (2009), which I am essentially defending against Worsnip) shows that many philosophers have mischaracterized the central question in the debate concerning the epistemic significance of disagreement. David Enoch (2010, 976) exemplifies this mischaracterization of the debate when he writes that “it is thus…common ground that the fate of the peer-disagreement issue (and in particular, that of the Equal Weight View) is pretty much determined by the answer to the question we are now considering, namely, whether the disagreement itself can count as evidence that your interlocutor is less than fully your peer.” Enoch thinks that this is the key issue because he thinks that if we can demote someone from the status of peer in response to a disagreement, then even if you take someone to be a peer prior to learning of a disagreement over p, “once you find out about the disagreement, you can justifiably demote him from the status of peerhood, and stick to your own judgement about p (after all, that someone who is your epistemic inferior disagrees with you is not a strong reason to change your mind)” (2010, 974). In other words, Enoch (like Worsnip) takes it for granted that if a disagreement justifies me in estimating my own reliability to be higher than my disputant’s reliability, then I am not rationally required to think that my disputant’s view is just as likely to be correct as my own view. If this view was right, then to show that extreme conciliationism is false, it would suffice to show that a disagreement with someone who I took to be my epistemic peer can itself be a reason for demoting that person from peer status. But we’ve seen that Enoch and others are wrong to suppose that demoting someone from peer status means that we are not required to give equal weight to their view. In a peer disagreement like the one in ESTIMATION where antecedent estimated net reliability is zero and antecedent net resilience is positive, the rational response is to both demote the disputant from peer status (i.e., estimate his reliability to be lower than yours) and give the disputant’s view equal credence to your own. Of course this demotion may mean that in future (independent) disagreements the disputant will not be an antecedent peer, and thus that you will not be required to give equal weight to your disputant’s view. But the fact that the present dispute should lead you to demote your disputant does not entail that you can assign more weight to your view on the presently disputed question.
Third and finally, the debate over the epistemic significance of disagreement should not be focused on the question of whether the evidential significance of disagreement depends on antecedent reliability estimates (it clearly does), but should instead be focused on questions about the sorts of reasons that may legitimately inform one’s antecedent reliability estimates. Let me explain. Because the credences I adopt in response to a disagreement with subject S must cohere with the proportion of time I expect to be right in disagreements with S, and because the proportion where I expect to be right straightforwardly depends on the estimated reliability of S, it should be uncontroversial that the degree of doxastic revision that is appropriate in light of a disagreement will partially depend on the estimated reliability of the disputant. Bayesian conditionalization and tables of the sort developed in section 3 both show that the key factors that determine the required level of conciliation are my own initial credence and the estimated reliability of my disputant. If we suppose that my initial credence should be equal to my estimated reliability (an assumption that was highly plausible in ESTIMATION, at least), then it would seem that the required level of conciliation will be a function of my and my disputant’s reliability, as extreme conciliationism maintains. Suppose this is right. Have we settled the disagreement debate by showing (by means of Bayesian formulae or simple tables) that estimated net reliability (rather than resilience) should determine the level of doxastic revision in response to disagreement?

No, we have not. Even if all sides were to agree that the evidential significance of a disagreement is determined by antecedent estimated net reliability, there is a substantive philosophical question that continues to divide the (broadly) conciliatory camp from the (broadly) steadfast camp. The question is whether one’s antecedent estimates of the reliability of oneself and of one’s potential disputant must be arrived at in a dispute-independent manner, so that one does not in some way presuppose the correctness of one’s reasoning on the potentially disputed matter. Like David Christensen (2009; 2011) and many others, I think that such an impartiality requirement must be affirmed in order to arrive at a version of conciliationism that is worthy of the name. If I can base a high estimate of my own reliability with respect to p on the fact that my reasoning about p is evidently sound and compelling, then my net reliability estimate in a disagreement over p is likely to be positive even in disagreements with someone who, considered from a dispute-neutral perspective, appears to be my peer. If on the other hand I am required to estimate my own reliability (and the reliability of others) in a manner that is not partial to my doxastic perspective, then my net reliability estimate will be close to zero (or less) in a much wider range of disagreements, giving disagreement much greater evidential significance.

As it happens, I think that the impartiality requirements posited by conciliationists are implausible. I therefore think that demanding conciliatory views are mistaken. But they are not mistaken because they say that the level of conciliation is determined by antecedent estimated net reliability. I’ve argued (against Worsnip) that antecedent estimated net reliability, and not antecedent net resilience, accounts for the degree of doxastic revision that is required in the wake of a one-shot disagreement. If demanding conciliatory views are mistaken, it is not in their emphasis on estimated net reliability, but in the constraints they impose on the reasons that may inform one’s reliability estimates.

---

11 Describing my reasons for thinking this is of course a task for another paper.
REFERENCES


