Why No True Reliabilist Should Endorse Reliabilism

Jeff Dunn & Kristoffer Ahlstrom-Vij

DRAFT of 10/4/2017

1. Reliabilism and the Intra-personal Trade-off Problem

Consequentialists believe that what's *right* should be understood in terms of what's *good*. For example, for the classic *utilitarian*, it's right to give to charity when it maximizes happiness. Similarly, in epistemology, the *reliabilist* believes that one is justified in believing something when the belief is formed by a process that tends to lead to true belief (e.g., Goldman 1979). Recently, opponents of reliabilism have suggested that this similarity lands her in trouble. Utilitarians infamously face interpersonal trade-offs where the suffering of some must be traded for the benefit of others, for example by condoning the surgeon who harvests an innocent person for organs to save five (Thomson 1976). According to her critics, the reliabilist faces *intra*personal trade-offs, where unjustified beliefs must be formed by a person to increase her accuracy with respect to future beliefs (Berker 2013*a*, *b*, Littlejohn 2012, Jenkins 2007, Firth 1981; cf. Greaves 2013).

As it turns out, however, these critics are mistaken. For one thing, the supposed trade-off cases put forward to date do not present a problem for the reliabilist; they're all either not trade-offs the reliabilist needs to make, or they're not problematic trade-offs (Ahlstrom-Vij and Dunn 2014; Goldman 2015). For example, a simple trade-off problem would consist in the reliabilist needing to condone a belief formed in light of excellent evidence to the contrary, but that would have as a causal consequence a great many true belief in the future. But since the reliabilist evaluates the justificatory status of beliefs, not with reference to *its* consequences, but rather with reference to the direct (more on this below) consequences of the type of process that generates it, the reliabilist doesn't have to condone the formation of such a belief, since forming a belief in light of excellent evidence to the contrary arguably constitutes an unreliable process.

Further, and perhaps more importantly, while reliabilism is indeed a form of consequentialism, it's not of a kind on which we should even expect trade-off problems to arise in the first place (Dunn and Ahlstrom-Vij forthcoming). More specifically, the type of consequentialism on which we should expect trade-off problems is one that doesn't impose any *side-constraints* (e.g., Nozick 1981). In ethics, imposing side-constraints on actions is to maintain that it can be wrong to do something, even if it has very good (including the best) consequences (e.g., because it violates people's rights). In epistemology, imposing side-constraints amounts to holding that there are some things you are not to believe, even if doing so would have very good (including the best) consequences from the perspective of epistemic value. But, as we've already seen, the reliabilist does in effect impose side-constraints: she maintains that a belief can be unjustified, even if it might lead to many true beliefs, if that belief would be formed by way of an unreliable process.

Consequently, it shouldn't come as a surprise that reliabilism doesn't fall prey to the intra-personal trade-off problem, contrary to what her detractors have suggested recently. That said, what hasn't been noticed is that the reliabilist *does* face a trade-off problem, albeit of a kind that is different from the type that has been discussed so far. In Section 2, we'll show how the problem arises once the reliabilist attempts to evaluate social institutions, and becomes

forced to make a variety of trade-offs between individuals and groups. In Section 3, however, we'll also argue that the problem runs deeper than this, in that it is most fundamentally not about a trade-off between individuals and groups but rather between the motivating idea behind reliabilism and the theory proper. We'll suggest that these trade-offs force the reliabilist into a dilemma: either she can hold on to her reliabilism by imposing side-constraints in all epistemic evaluation -- including at the social level -- or she can reject side-constraints on both the individual and social level, in effect ceasing to be a reliabilist but in so doing actually staying true to the motivating idea behind reliabilism. We'll argue, in Section 4, that anyone moved by the considerations that probably attract people to reliabilism in the first place -- very roughly, the idea the true belief is good, and as such should be promoted -- should go for the second horn of that dilemma.

2. The Social Trade-off Problem

In fleshing out the trade-off problem the reliabilist faces it's helpful to consider Alvin Goldman, undoubtedly the most prominent reliabilist. Goldman (1999) suggests that social practices are to be evaluated—*exclusively*, as far as we can tell—by how well they raise 'the *aggregate* level of [true belief] of an entire community' (93). But we can imagine a variety of unreliable means to that end. To see why, keep in mind two facts we introduced at the outset: first, that the reliabilist evaluates beliefs, not with reference to *their* consequences, but with reference to the consequences of the types of processes that generate them; and, second, that such processes are evaluated with reference to their *direct* consequences. Elsewhere (in Ahlstrom-Vij and Dunn 2014), we've put this point in terms of reliabilism being both *indirect* (in its evaluation of beliefs) and *direct* (in its evaluation of processes). While the indirect nature of reliabilism might be fairly obvious, its directness might not. So, a word is in order by way of motivating the latter.

Say that I'm trying to locate some particular book in my bookshelf. Some psychological process will be at work, and when evaluating the belief 'There's the book I'm looking for', the reliabilist will look to evaluate how reliable the process involved is. Specifically, if the visual processes involved in locating the book are reliable, then that belief will be justified. But, of course, there's a sense in which the consequences of that process extend far beyond the aforementioned visual belief. Once I've located the book, I might start reading it and form a great number of beliefs as a result. Those beliefs might, in turn, lead me down a variety of different lines of inquiry, that in turn will have a multitude of doxastic consequences. But the reliabilist doesn't factor in all of *those* consequences when trying to determine whether the visual process that originally led me to the book is reliable. The only consequences relevant to the reliability of that process are the *direct* doxastic consequences of that type of process being instantiated, which, very roughly, will be evaluated in terms of the truth-ratio of the set of beliefs formed as a direct result of looking for purposes of visually locating things.

Keeping that point in mind, we can see why the reliabilist is forced to accept that there are a variety of ways to raise the raise the aggregate level of true belief in a community by way of unreliable belief-formation. In making our case for this claim, it will help to start with an overly simplified example. In particular, consider the crucial role played by idealisations and simplifications in education. The physics teacher will tell her students that $F=ma$ and the ethics teacher may tell her students that all consequentialists are utilitarians. Both of these things are false, and the teachers involved are well aware of this, of course. But they also know that, in telling their students these particular falsehoods, they are furnishing crucial epistemic stepping-stones to students, which thereby facilitates their learning of many true things in

the future when the students have reached a level of sophistication that enables them to grasp a fuller but also far more complex picture. Note, however, that even if -- as seems eminently plausible -- forming these initial, false beliefs will have a very high epistemic pay-off in the long term, they may come out as unjustified on the reliabilist picture. That is, the *direct* doxastic consequences of the psychological processes involved will tend *not* to be the formation of true belief; after all, to be successful, idealised and simplified testimony needs to give rise to false belief in the recipient.

Of course, in making that claim, we are making certain assumptions about how to determine the process type for purposes of epistemic evaluation -- a notorious problem for the reliabilist (e.g., Conee and Feldman 1998), and possibly for her opponents as well (e.g., Bishop 2010). For example, the most convenient assumption for our argument would be that the students have and use a *dedicated* psychological process for hearing and coming to believe idealised and simplified content. Since idealised and simplified content is *false*, such a process would have a reliability of 0. Of course, it is doubtful there is such a dedicated process -- that's exactly what makes the example overly simplified -- but that does not harm our argument *per se*. What we need for our argument to go through is that it's *possible* to promote the aggregate level of true belief in a community by having people form unjustified beliefs. And it seems that the structure of reliabilism plus several other facts, make it certain that such situations can arise.

To see this in the abstract (we'll consider a more realistic example shortly), notice that there are certain facts that are incredibly important for an individual to believe in order for that person to gain many further true beliefs. We could call such beliefs 'cornerstone beliefs' for their role in helping to build structures of true belief. Now, recall that according to reliabilism, a belief is unjustified if and only if it is the direct result of an unreliable process. Take a particular cornerstone belief. This belief, if widely adopted by members of a population, will lead members of that population to have many more true beliefs than they otherwise would have had. But it is surely possible that the best way to get a large section of the population to hold this particular cornerstone belief is to have the belief be the direct result of an unreliable process. In such a situation, we (vastly) increase the aggregate level of true belief in a community by having people form unjustified beliefs. And since we haven't committed ourselves to any particular way of typing processes in running this argument, questions about how to type processes are irrelevant.

Note, too, that in making this argument we are exploiting the very fact that enables the reliabilist to avoid trade-off problems at the individual level: what determines whether a belief is justified is not the forward-looking matter of whether *it* leads to true belief, but the backward-looking matter of whether *the process* that generates it tends to do so. Nevertheless, the fact that the cornerstone beliefs will tend to lead to true belief in the longer term is of course exactly what makes them a good and possibly indispensable means towards, as Goldman puts it, raising the aggregate level of true belief in the community. And this is what presents a trade-off problem for the reliabilist, who has to be prepared to trade off justified belief in the group now for a greater aggregate of true belief in the future. This is the social trade-off problem for the reliabilist.

In the next section, we'll consider natural replies the reliabilist might make. But before doing so, it may be useful to see a more realistic example of the kind of problem we have in mind, now that its basic structure is clear. The particular example we'll present draws on the phenomenon of *motivated reasoning*. As Dan Kahan (2016) writes: "'Motivated reasoning" refers the tendency of individuals to unconsciously conform their assessment of information to

some goal collateral to determining its truth' (2). *Politically* motivated reasoning occurs when one forms or updates one's beliefs in virtue of one's political affiliations rather than in virtue of a search for the truth. Kahan explains this in terms of a simple Bayesian model of belief updating. According to this model, your degree of belief in A, $c(A)$, after learning evidence E ought to be $c(A|E)$ (the degree of belief in A conditional on E). What is the value of $c(A|E)$? Bayes theorem tells us that $c(A|E) = \frac{c(A)\times c(E|A)}{c(A)\times c(E|A) + c(\neg A)\times c(E|\neg A)}$. Since $c(A)$ and $c(\neg A)$ are already determined, the value of $c(A|E)$ is wholly determined by the values of $c(E|A)$ and $c(E|\neg A)$. This can be seen by rewriting Bayes theorem in the following way:

$$\frac{c(A)}{c(\neg A)} \times \frac{c(E|A)}{c(E|\neg A)} = \frac{c(A|E)}{c(\neg A|E)}$$

The first term is the *prior odds*, which when multiplied by the *likelihood ratio* results in the *posterior odds*. Once you have your prior odds on some proposition, A, the likelihood ratio tells you how to set your beliefs with respect to A upon receiving evidence E. It's basically a summary of how you view the evidential relationship between E and A. If the ratio is greater than 1, you think E favors A over its negation; if the ratio is less than 1, you think the opposite; and if the ratio is 1, you think E is evidentially irrelevant to A. Here again, is Kahan: 'The distinctive feature of "politically motivated reasoning" is the disposition of individuals to derive the likelihood ratio for new information from their political predispositions rather than from truth-convergent criteria' (6). By way of example, suppose A is the proposition that the death penalty reduces the murder rate and suppose that E is the proposition that a Harvard study indicates no relationship between states with the death penalty and those states' murder rates. Normally, we would think that $c(E|A)$ is much less than $c(E|\neg A)$, resulting in a likelihood ratio less than 1. On Kahan's model, politically motivated reasoning happens when you nevertheless set your likelihood ratio in such a situation to 1 (or higher), perhaps discounting the study as due to liberal bias (*e.g.,* "Well, *they* would've come to that conclusion no matter what."). There is evidence that politically motivated reasoning affects people's beliefs on issues such as climate change, gun control, the safety of nuclear power, the efficacy of vaccines, and many other politically charged issues. Further, given that this is an active area of study amongst psychologists, there is some reason to believe that there is a psychologically real belief-forming process, *politically motivated reasoning*. We'll assume that is so in the case to be considered.

With this background, consider the following example. Suppose it is true that climate change is primarily caused by human activities. At least with respect to some range of propositions, this fits the description of a cornerstone belief. Politically motivated reasoning leads many in the U.S. on the political left to accept that this is true and many in the U.S. on the political right to believe that it is false. Since we are assuming that it is true that climate change is primarily due to human activity, and since this belief leads to many more true beliefs, we take it that the reliabilist social epistemologist wants to cultivate this belief in the population. Here are some options for policies the reliabilist social epistemologist might undertake:

Option 1: Take measures to package information about climate change, when delivered to those on the political right, in ways that are more appealing to those on the right. For instance, have messages about climate change given by people who look more like traditional conservatives, and in a way that appeals to conservative values.

Option 2: Take measures to educate people about fallacies in reasoning and expert testimony. Make classes in reasoning and informal logic required at the high school level.

Option 3: Status quo: keep doing what we're doing.

There is some evidence that the "packaging techniques" described in Option 1 make it more likely that information is believed by those on the right (Kahan 2010; Cohen, *et. al.* 2000). Further, there is some evidence that classes about reasoning and logic are only minimally effective in changing people's beliefs and belief-forming strategies (cite). Hence, in such a situation, it seems plausible that the Option 1 is to be preferred by the reliabilist social epistemologist. And note, too, that even if you think Option 2 will increase the aggregate level of true belief, it is possible to pair Option 2 with Option 1, which the evidence suggests will lead to even more true belief.

But consider the belief-forming process: *believing p when p is testified by someone from my culture and in a way that affirms my values*. This process is not necessarily a reliable process. And so the reliabilist must say that beliefs formed by such processes are unjustified, even while she encourages such belief formation in order to raise the aggregate level of true belief in a community.

In response, one might insist that better than Option 1 is some sort of beefed-up version of Option 2 whereby the reliabilist social epistemologist recommends some strategy to neutralize politically motivated reasoning across the board. That is, one may think what is best is to promote less politically motivated reasoning on both the right and on the left. There are two things to note about this suggestion. First, there is nothing incompatible with in general promoting less reliance on politically motivated reasoning and yet still, in certain cases, promoting the kind of packaging techniques mentioned above. It is not as if we can promote only one kind of belief-formation for all circumstances and for all people.

The second thing to note is that even if it were possible to get people to stop relying on politically motivated reasoning, it could easily turn out that the use of politically motivated reasoning by those on the left, with respect to climate change, actually gets those on the left to have more accurate beliefs about climate change than if they were to engage in some other form of reasoning. For it is certainly plausible that some on the left would make mistakes, were they to cooly examine the evidence and consult the experts. So, the reliabilist social epistemologists should encourage politically motivated reasoning on the left -- at least in the case of climate change beliefs -- even though politically motivated reasoning is not a reliable belief forming process.[1] Again, considerations of accuracy promotion seem to require the promotion of unreliable belief forming processes.

[1] And note that the type of process, *politically motivated reasoning*, is not changed by the fact that it is promoted in a situation where it yields a true belief. It is not as if in promoting such a belief forming method in a case where the method gets it right, changes the psychological process into something like *politically motivated reasoning in cases where it yields true belief*. The psychological process is what it is independent of when it is promoted.

If this is correct, we have a trade-off problem for the reliabilist that's different in kind from what has been put to her so far. Unlike those that have been figuring in the literature of late, this trade-off problem is not an intra-personal one. Rather, it involves interpersonal trade-offs between individuals and the groups the are part of. More specifically, the reliabilist seems committed to promoting the formation of unjustified beliefs by individuals so that the aggregate accuracy of the community increases.

3. Means, Ends, and a Dilemma

The reliabilist might respond that the trade-off identified in the previous section is not a particularly difficult one. After all, we pursue justification as a *means* to true belief, so when we trade off justification for true belief we're not really trading one good for another, as opposed to simply opting for one means to a good over another. Indeed, to see this more clearly, contrast the social trade-off from the previous section with the intra-personal trade-off, where we trade off true belief for true belief. *That* seems a real trade-off (whether or not it's ultimately to be considered *problematic* for the reliabilist is of course a different matter), while trading justification for true belief does not.

But this response fails to show that there's no problem for the reliabilist here. If anything, it does the opposite. Remember, Goldman suggests that, on the social level, we should promote whatever raises the aggregate level of true belief. The response above reinforces the reliabilist's commitment to this idea, by suggesting that, since justification is a mere means to the end of true belief, we should promote whatever raises the aggregate level of true belief, even if the way to do this is by having people form unjustified cornerstone beliefs. But this means the reliabilist should see to it that people form unjustified beliefs.

If we include a widely-accepted principle linking epistemic justification and epistemic obligation, we can make the problem especially acute for the reliabilist. The widely-accepted principle is this:

(*) If S's belief that *p* is/would be unjustified, then S epistemically should not believe *p*.[2]

If (*) is true, then the reliabilist is committed to saying that we should in certain circumstances see to it that people form beliefs that they should not form. Things get even worse if we endorse a plausible principle about interpersonal epistemic obligations :

(†) If S epistemically should not believe *p*, then it is not the case that R (possibly identical to S) epistemically should see to it that S believe *p*.

If both (*) and (†) are true, the reliabilist is really in trouble. For then we get a contradiction. Since the cornerstone belief is unjustified, (*) says that the believer, S, should not believe it. And, given (†), it follows that it is not the case that we should see to it that S believes it. And yet, the reliabilist social epistemologist is committed to saying that we should see to it that S believes the cornerstone belief.

---

[2] Goldman (1986, p. 59) writes: "Calling a belief justified implies that it is a *proper* doxastic attitude, oen to which the cognizer has an epistemic right or entitlement. These notions have a strong deontic flavor...They are naturally captured in the language of 'permission' and 'prohibition'..." See also Goldman (1986, p. 5) where he states that he treats 'justification' as having the kinds of deontic connections in (*). This also tells in favor of Goldman's support for (**) below. Further (**) below is entailed by Goldman's "framework principle" for justification, (P3) (1986, p. 63). (*) is entailed by (P3) supposing that there are no underminers at play, which is plausible in our cases.

Hence, the social trade-off problem, plus (*), shows that the reliabilist is committed to saying that we should under certain circumstances see to it that people form beliefs in ways that they shouldn't. If we add (†), the social trade-off problem shows the reliabilist view to be contradictory. How did we end up here? By accepting that there are epistemic side-constraints on the individual level -- that, remember, is exactly how the reliabilist avoids intra-personal trade-off cases -- while denying that there are any such constraints on the social level.

This all raises a question for the reliabilist: if we are invariably to go with whatever raises the aggregate level of true belief when evaluating social institutions, even if that means forming unjustified belief, why not do the same on the individual level? That is, if we are to follow the motivating idea behind reliabilism at the social level why not follow it at the individual level? Why is it *not* right to form beliefs that will result in lots of true belief even if they themselves are not produced by reliable processes? We can put this point in the form of a dilemma: the reliabilist needs to either (*a*) hold on to reliabilism, and with it the prohibition against unreliable belief-formation, but then also try to explain why we should allow the pursuit of justification to get in the way of that of true belief, if justification is merely valuable as an instrument to true belief; or (*b*) say that it's aggregate level of true belief that matters, and then drop the prohibition of against unreliably formed but truth-conducive belief on the individual level -- and in so doing in effect give up on reliabilism. In what follows, we will make a case for the second horn of this dilemma.

[Insert a section about the related case where a justified belief should not be formed?]

It is perhaps worth noting that there are cases related to the ones we discussed in section 2, that create some tension for the reliabilist, but not quite in the same way discussed in section 3. These are cases where there are certain beliefs that people could form that would be justified for them, but were they to form those beliefs, it would lower the aggregate level of true belief for the group.

Here's one example. Suppose we have a group of 12 jurors who are hearing evidence in a murder trial. Suppose that it is true that the victim's ex-husband had written threatening messages to her several days before the murder. Suppose that this is reliably testified to by a police officer who searched the victim's phone. If the juror's were to believe that the victim's ex-husband had written threatening messages to her several days before the murder, this belief would be true, and by the reliabilist's lights, justified. However, suppose it is also true that the ex-husband did not commit the murder and in fact was in a different state at the time of the incident. But suppose, too, that there is no one to corroborate this. It is very likely that if the jurors come to justifiably believe that the ex-husband sent threatening messages to the victim, they will go on to form many false beliefs about the case, *e.g.,* that the ex-husband did commit the murder, that the other suspects are innocent, that the ex-husband is lying about his whereabouts, etc. The reliabilist social epistemologist must say that in this kind of case, the goal of raising the aggregate level of true belief among the group of jurors is best served by them *not* forming this justified belief about the threatening messages. In general, we will get such cases whenever some proper subset of the complete body of evidence points away from the truth, but nevertheless there are justified ways of believing the propositions that constitute the proper subset of the complete body of evidence. In such cases, the reliabilist social epistemologist must claim that we should see to it that people do not adopt those beliefs even though the beliefs would be justified.

Now, add to that:

(**) If S's belief that *p* is justified, then it is epistemically permissible for S to believe *p*.

From this we get that in some cases we ought to see to it that people do not hold beliefs it is permissible for them to hold. That is odd on its own, but gets even worse if we add:

(††) If it is epistemically permissible for S to believe *p*, then it is epistemically permissible that R (possibly identical to S) sees to it that S believes *p*.

Then, given standard deontic logic, we get a contradiction. For suppose that ϕ is the action of S justifiably believing *p*, O is the operator "we are epistemically obligated to see to it that", and P is the operator "it is epistemically permissible that we see to it that". Then, from the case above we have that in certain situations $O\neg\phi$. Since ϕ is the action of S *justifiably* believing *p*, it follows from (**) that it is permissible for S to believe *p*, and from (††) that it is permissible that we see to it that S believes *p*. That is, we have $P\phi$. But, this is equivalent on standard deontic logic to $\neg O\neg\phi$, and so we have a contradiction.

4. Orthodox Reliabilism and the Basic Problem

A natural response for the reliabilist to what we have argued so far is to suggest that there is no problem for reliabil*ism* as such, but merely for certain reliabil*ists* who also want to evaluate social-epistemic phenomena along consequentialist lines. Strictly speaking, reliabilism is a view—and *only* a view—about the justification of the beliefs of individuals. As such, it carries no implications for how we should go about evaluating social institutions. Consequently, anyone embracing reliabilism will not, simply on account of doing so, face the social trade-off problem outlined above.

There's a sense in which this response is completely right: reliabilism is strictly speaking only an account of individual justification. But though thinking about reliabilist social evaluation brings this problem into focus, we will argue that it reveals a more fundamental problem that is not so easily avoided by the reliabilist. This fundamental problem is a tension between the motivating idea behind reliabilism and the theory proper. To get at this problem, think about what would drive one to embrace reliabilism in the first place. It is a theory of justification motivated by the conviction that justification is a means to the epistemic good, where the epistemic good is understood as truth or accuracy in belief. This is exemplified, for instance, by Sandy Goldberg in his recent book on reliabilism:

> Throughout our discussion [...] we have been noting that reliabilist views of doxastic justification get much of their motivation from the way they honor the link between truth and justification. Belief aims at truth, and particular beliefs are justified to the extent that [they] are formed (and sustained) in such a way that they are likely to be true (Goldberg 2010, 151).

This conviction also explains why reliabilists -- including Goldman (e.g., 1992, 167-168; 1986, 27) -- are interested not just in reliable but also in *powerful* processes, as in processes that generate a lot of true belief. But if the value of true

belief and a desire to see more of it is the underlying motivation, then it's not clear why any reliabilist would want to impose side-constraints on the individual level. Imposing such constraints commits one to saying that what's right about believing with justification is that it helps us get more of a good thing, but that we should not generally strive to realise more good things -- because only if we say the latter is there any motivation for a prohibition on unreliable belief-formation where the beliefs formed give rise to lots of true belief in the future. Of course, some philosophers have said things to the effect of it not always being right to realise more good rather than less good when accounting for side-constraints outside of epistemology. For example, Bernard Williams (1973) suggests that 'with respect to some type of action, there are some situations in which that would be the right thing to do, even though the state of affairs produced by one's doing that would be worse than some other state of affairs accessible to one' (90). Along a similar line, Foot (1985) denies 'the rather simple thought that it can never be right to prefer a worse state of affairs to a better' (198). But to side with Williams and Foot here is of course to reject exactly the type of consequentialist story that the orthodox reliabilist is relying on in motivating her particular account of justification, i.e., that to believe with justification is to believe in the right way because it puts us in a position of getting us more of a good thing. So, the orthodox reliabilist would seem to be buying into the idea of there being side-constraints at the expense of doing away with the basic motivation for her view. Put simply, what we have here is a tension between the motivating idea behind reliabilism and the theory itself. The tension is this: accuracy may sometimes be best promoted by forming a belief that is the direct product of an unreliable process but where that belief itself leads to many other true beliefs. The theory says the belief is unjustified even though it is a means to the epistemic good of accuracy.

Seeing the fundamental problem as a tension between the motivating idea behind reliabilism and the theory proper also helps us see why the social cases we have discussed so naturally display the problem. Suppose you're a reliabilist thinking about how to epistemically evaluate social practices and institutions. The motivating idea behind reliabilism is that justification is a means to the promotion of accuracy and this idea provides a recipe for reliabilists to perform this social evaluation. We evaluate social practices and institutions according to how well they promote accuracy in belief, aggregated across a group or society. But then we are going to run into the problem we identified in the section above: situations where accuracy in belief can be promoted within a group by having group members form beliefs that are produced by unreliable processes of belief formation. The social trade-off problem is thus a symptom of a deeper ailment.

At this point, the reliabilist might object that her position is not motivated by the idea that the right should be understood in terms of the good, or any desire to see more good things. She might be a reliabilist simply because she believes that the job of the epistemologist is to generate theories that fit with our intuitions, and that reliabilism best fits our intuitions about relevant hypothetical cases. But this is not a successful strategy. Reliabilism is a revisionary theory. This is so because what processes are reliable will be partly an empirical matter, which is why a reliabilist account of epistemic categories can yield surprising results, and indeed has generated surprising results about, among other things, the extent to which we're not particularly well-served by reflection (Kornblith 2012), the many cases in which we would do well to think less and instead rely on deceptively simple reasoning aids (Bishop and Trout 2004), and the circumstances under which blindly deferring to people can constitute an epistemic virtue (Ahlstrom-Vij 2015).

Revisionary theories require motivation to overcome their clash with widely held intuitions -- and that's exactly where the reliabilist needs to appeal to the value of true belief, and our desire to see more of it. Without that appeal, it's not clear with reference to what she would seek to convince those not already on board with her framework. And if that's so, then we're of course back with the problem we outlined above: either the reliabilist accepts that what's motivating her reliabilism is a commitment to the value of true belief and a desire to see more of it, and in so doing faces up to the basic problem, or she has to accept that her reliabilism lacks any motivation.

Needless to say, the reliabilist is not in a good position if she endorses a theory which is unmotivated. This seems to leave only the former route, which is to embrace the commitment to the value of true belief and its promotion. More generally, it would seem that anyone who's serious about the value of truth should reject reliabilism and say exactly what Goldman (1999) says, which is that social practices are to be evaluated with reference to the extent to which they raise the aggregate level of true belief in the relevant community. But of course if one says that about social practices, one must say the same about the individual level. This requires the reliabilist to drop the prohibition against unreliably formed but truth-conducive beliefs on the individual level, which is, in effect to give up on reliabilism. This is why no true reliabilist should endorse reliabilism.

5. Three Options

We've argued that it is an untenable position for the reliabilist to maintain orthodox reliabilism and the standard consequentialist motivation for it. That consequentialist motivation doesn't motivate orthodox reliabilism. It seems, then the reliabilist has two options. First, to stick with the consequentialist motivation and revise one's theory accordingly. Second, to come up with some other motivation for orthodox reliabilism. We've already argued that the second option won't work if the new motivation is that orthodox reliabilism fits best with our intuitions; it doesn't. But perhaps there are other ways to motivate orthodox reliabilism. In the next subsection we'll canvas three such alternative motivations. In the section after this, we'll pursue the first option -- to revise the theory -- and ask what a truly consequentialist theory in epistemology would have to look like.

5.1 Alternative Motivation for Orthodox Reliabilism: Naturalistically Acceptable Intuition Satisfying

Above we argued that since reliabilism is a revisionary theory, it does not best capture our intuitions about justification. Hence, the motivation for reliabilism cannot just be that it best captures our intuitions about justification. However, there is a more nuanced motivation for reliabilism, which might seem to fare better. This motivation comes from Alvin Goldman's seminal (1979) paper, "What is Justified Belief?"

According to this motivation, we should adopt a theory of justification that (a) is naturalistically acceptable and (b) best captures our intuitions about justification. What does it mean for a theory of justification to satisfy (a) and so be naturalistically acceptable? Goldman's basic idea is that a theory of justification is naturalistically acceptable if it makes no use of evaluative or deontic terms or concepts in its statement. As Goldman writes:

The term 'justified', I presume, is an evaluative term, a term of appraisal. Any correct definition or synonym of

> it would also feature evaluative terms. I assume that such definitions or synonyms might be given, but I am not interested in them. I want a set of *substantive* conditions that specify when a belief is justified. ...I want a theory of justified belief to specify in non-epistemic terms when a belief is justified. (p. 90)

Reliabilism seems to fit the bill here, since justification is specified in terms such as 'psychological process' and 'ratio of true beliefs to false beliefs', none of which are evaluative epistemic terms.

The idea, then, is that we have certain intuitions about justification: my belief that I have hands is justified, beliefs in the predictions of astrology are unjustified, etc. Reliabilism about justification vindicates these intuitions and yet is naturalistically acceptable. Other theories of justification, such as evidentialism, may do better in terms of capturing intuitions, but, the thought goes, they are not naturalistically acceptable. Reliabilism is the unique theory that does best with respect to (b) while satisfying (a).

One way to see the problem with this motivation for reliabilism is to point out how odd the project starts to look once we consider cases like those that occupy Goldman towards the end of section II of his classic paper. After presenting his reliabilist theory of justification, he considers what to say about processes of belief formation, like wishful thinking, that though unreliable in our world may be reliable in other possible worlds. Goldman is indecisive in the face of such counterexamples, and notes that we can opt for a version of the theory according to which what matters is a process's reliability in the world it is used in or a version of the theory according to which what matters is a process's reliability in our world. In the end he suggests that perhaps this shows that the method of conceptual analysis has shortcomings. He then writes:

> What we really want is an *explanation* of why we count, or would count, certain beliefs as justified and others as unjustified. Such an explanation must refer to our *beliefs* about reliability, not to the actual *facts*. The reason we *count* beliefs as justified is that they are formed by what we *believe* to be reliable belief-forming processes. … What matters, then, is what we *believe* about wishful thinking, not what is *true* (in the long run) about wishful thinking. (p. 101)

What is worth pointing out is that if we defend reliabilism in this way, it is no longer clear why we want a naturalistically acceptable theory in the first place. For no longer are we trying to determine *which* beliefs are reliably formed and so get us the kind of epistemic improvement we want. Instead, we are trying to determine what we believe about justification. But there is no reason why our beliefs about justification should be naturalistically acceptable, in the sense that 'justification' is defined using non-evaluative terminology.

In addition, if this is our motivation for reliabilism, theories of justification seem quite a bit less interesting. For we get a theory about which beliefs get to be called 'justified beliefs', but such a theory tells us little about which beliefs are epistemically good for us to form. Put another way, it is hard to see how this motivation for reliabilism gets us the deontic connections that Goldman seems to want between justified beliefs and epistemic obligation and permission (as captured in (*) and (**)).[3]

[3] In section 6 we consider whether justification understood in this way, while epistemically uninteresting, might be interesting in some other sense.

5.2 Alternative Motivation for Orthodox Reliabilism: Consequentialist Intuition Satisfying

In Goldman (1986), we get a different kind of motivation for reliabilism, and in fact end up with a view that is subtly different from the view proposed in Goldman (1979). In his (1986) Goldman explicitly describes his approach as one of reflective equilibrium (p. 60) according to which intuitions about justification have an important role to play. And he continues to want a theory of justification according to which justification is defined using non-evaluative terms, though this plays a less prominent role. So there are affinities with the (1979) approach.

But there is also a new ingredient motivating reliabilism, which is a commitment to some form of epistemic consequentialism. As Goldman puts it, he wants a theory of justification that is "truth-linked" (p. 69). Here, for instance, is what Goldman (1986) says about why he doesn't go for a coherence theory of justification:: "The fundamental standard concerns the formation of true belief. Coherence enters the picture only because coherence considerations are generally helpful in promoting true belief." (p. 100) He goes on to say: "True belief is the value that J-rules [rules dictating which beliefs are justified] should promote--really promote--if they are to qualify as right." (p. 103)

These comments about epistemic consequentialism seem to go *against* a reliabilist account of justification, however, and instead in favor of an account that evaluates each belief in terms of its epistemic consequences. Why, then, does Goldman (1986) reject such a view? His explicit reason is rather underwhelming: "I ignore entirely the suggestion that the justificational status of each belief is a function of that very belief's consequences." (p. 97).

Why is such a view ignored? Goldman is not explicit about this, but two ideas suggest themselves. First, right before he says he will ignore the straightforward consequentialist picture, he claims that what we should be interested in is a kind of *rule* consequentialism, which might seem to rule out such a view. We should be interested in rule consequentialism, according to Goldman, because we are interested in the rightness of (what he calls) J-rules, rules dictating which beliefs are justified. But this is just a mistake. We could certainly have a J-rule that says: believing *p* is justified iff the consequences of believing *p* maximize the number of true beliefs held. *That* is a J-rule that yields a version of epistemic consequentialism that is analogous to act utilitarianism.

I suspect that the real reason that Goldman ignores the view according to which each belief is evaluated in terms of its consequences has more to do with the fact that the theory we are left with does very well with respect to its consequentialist credentials, but fares poorly with respect to our intuitions about justification. In a telling quote, Goldman considers proposals to "regiment" the concept of justification in various ways that makes it more theoretically pleasing (it doesn't matter for our purposes what these regimented proposals look like). He writes: "Either of these [regimented] approaches might seem preferable from a systematic or theoretical point of view. Nonetheless, they do not seem to be what is implied by the ordinary conception [of justification] as it stands; and that is all I am currently trying to capture." (p. 109) So in Goldman (1986) we seem to be getting the following kind of rationale for reliabilism: it is the theory that (a) is in some sense consequentialist in that it promotes true belief, and yet (b) also satisfies our intuitions about justification.

What can be said in response to this motivation for reliabilism? First, though he relies heavily on intuition-satisfying, he nevertheless says things that cut against it. For instance, in dismissing a theory of justification

according to which a belief is justified just in case it is in conformity with the belief-forming rules accepted by one's society, Goldman writes: "Any such proposal invites an obvious objection. Why should we assume that what is accepted as justification-conferring by the members of a particular community really is justification-conferring? Can't such a community be wrong?" (p. 68). But what holds of a community's beliefs about justification surely hold of our intuitions about it.

Furthermore, in explaining why he doesn't go for a coherence-based account of justification, Goldman writes: "The fundamental standard concerns the formation of true belief. Coherence enters the picture only because coherence considerations are generally helpful in promoting true belief." (p. 100) If this is what we say about coherence, however, then why do we not say this about reliably-formed belief. Usually such beliefs promote true belief, but in some cases, unreliably-formed beliefs do. And in other cases, reliably-formed beliefs do not promote true belief. So, our criterion of what makes a belief epistemically right is not that it coheres with other beliefs nor that it is reliably produced.

What's more, Goldman's preferred criterion of rightness actually seems to permit unreliably formed beliefs as justified. Here it is:

> (ARI) A J-rule system R is right if and only if R permits certain (basic) psychological processes, and the instantiation of these processes would result in a truth ratio of beliefs that meets some specified high threshold (greater than .50).

Though this has not been widely noted[4], it is certainly possible that some set of psychological processes yields a truth ratio of beliefs above the threshold and yet one of the psychological processes itself is unreliable. So Goldman's own arguments against other views, and indeed his official statement of his theory seems to give significant weight to the truth-linkedness motivation for reliabilism. This is exactly the motivation that undermines orthodox reliabilism.

However, this is all directed at Goldman (1986). Might there be a plausible motivation for orthodox reliabilism that takes it to be the theory that (a) is in some sense consequentialist in that it promotes true belief, and yet (b) also best satisfies our intuitions about justification? The objection to this is similar to the objection to the proposal in section 5.1. If this is our motivation for reliabilism, theories of justification are quite a bit less interesting. We get a theory about which beliefs get to be called 'justified beliefs', but such a theory tells us little about which beliefs are epistemically good for us to form. To his credit, Goldman (1986) seems to notice this in his criticism of coherence theories of justification and theories of justification that make it culturally relative. What we claim is that the same criticism is effective against orthodox reliabilism.

5.3 Alternative Motivation for Orthodox Reliabilism: Knowledge as a Natural Kind

Here's a kind of motivation for orthodox reliabilism that doesn't depend on the consequentialist idea that true belief is to be promoted and that also doesn't depend on the dubious claim that reliabilism best captures our intuitions. This distinct motivation for reliabilism comes from Hilary Kornblith's (2002) proposal to investigate knowledge as a natural

[4] Though see Dunn (2012) for an in-depth examination into this issue.

kind. On Kornblith's view, epistemology is like natural science. Just as the subject matter of astronomy is the heavenly bodies themselves (and not our concept of them), so too the proper subject matter of epistemology does not consist of our epistemic concepts, but rather of *knowledge* itself. According to Kornblith, we can study knowledge itself by looking at the attributions of knowledge that biologists and cognitive ethologists are required to make to explain the survival and evolution of animals with robust cognitive systems. Here is Kornblith, summing up where he thinks this approach takes one:

> The knowledge that members of a species embody is the locus of a homeostatic cluster of properties: true beliefs that are reliably produced, that are instrumental in the production of behavior successful in meeting biological needs and thereby implicated in the Darwinian explanation of the selective retention of traits. The various information-processing capacities and information-gathering abilities that animals possess are attuned to the animals' environment by natural selection, and it is thus that category of beliefs that manifest such attunement—cases of knowledge—are rightly seen as a natural category, a natural kind. (pp. 62-3)

On this view, we get a naturalistic, non-intuition-based motivation for the claim that knowledge is reliably produced true belief. How could could one leverage this into a non-intuition-based motivation for reliabilism about justification? The natural way to go is to claim that justification just is whatever we add to true belief to get knowledge.

In evaluating this approach, let us grant Kornblith's (controversial) claim that knowledge is a natural kind, which just is reliably produced true belief. Even granting this, there are problems in getting a motivation for reliabilism about justification. First, and perhaps most obvious: since Gettier, there is now wide consensus that adding justification to true belief *doesn't* get one knowledge. So, the claim at the end of the last paragraph seems false. But more than this, the claim that justification is whatever we add to true belief to get knowledge is itself a kind of intuition-backed claim of the sort we're supposed to be disavowing here. The appeal of Kornblith's approach is that it would gives us a motivation for reliabilism *not* backed by intuition. For that to work, however, we'd need an argument that to explain the survival and evolution of animals with robust cognitive systems we need to appeal to justification as understood as reliably produced belief. But of course we don't need to appeal to justification for this. At best, we need to appeal to the capacity for various animals to *have* beliefs that are reliably produced. Whether such beliefs are to count as justified or not adds nothing to the explanation for their survival and evolution.

5.4 Alternative Motivation for Orthodox Reliabilism: Justification and Information Sharing

Let's now consider a different kind of motivation for orthodox reliabilism that doesn't depend on the consequentialist idea that true belief is to be promoted and that also doesn't depend on the dubious claim that reliabilism best captures our intuitions.

The key to this idea is to not focus on the value of truth or on intuitions about justification, but rather on why we might care about reliably-formed beliefs. Perhaps, the story goes, we care about justified beliefs -- and so reliably-formed

beliefs according to reliabilism -- because this is a way of determining which things to trust others about.[5] That is, if I tell you *P* is true and you know that I'm justified in believing that *P*, then this is good information for you that *P* is true. But if that's *why* we care about justification, then it can't just be that a justified belief is one that is conducive to increasing the overall amount of true beliefs the agent has. My belief that *P* can be conducive to overall accuracy and not be the kind of thing about which you should trust me. If justification is understood according to reliabilism, then for each belief of mine that is justified, it is true that the proposition believed is likely to be true. And so, on this way of going, to say that my belief that *P* is justified is to give others information about whether I should be trusted with respect to *P*.

This proposal attempts to drive a wedge between individual level epistemic evaluation and group level epistemic evaluation. When it comes to the group level, we are trying to evaluate various things we can do (institutions we can set up, beliefs we can inculcate, communication we can facilitate) that will lead to the overall maximization of epistemic value. When it comes to the individual level, we are trying to evaluate whether a person can be trusted with respect to certain propositions. These two different kinds of evaluation call for different structures and so there is no inconsistency in imposing side-constraints at the individual level while eschewing them at the group level. Put another way, when we talk about epistemic evaluation at two different levels (the individual and the group), what we're really considering are two questions:

1. What is it right for me to believe?, and
2. How should we (i.e., is it right for us) to set up social institutions?

Reliabilism answers the first question: it's right for me to believe what's formed by way of a reliable process. And reliabilists have been tempted by a particular answer to the latter question: we should set up institutions in a way that raises the aggregate level of true belief in society. We've been arguing that whatever motivates your answer to one of these questions should be the same as what motivates your answer to the other. The proposal now on the table suggests this is mistaken. What motivates reliabilism at the individual level is not raising the aggregate, but helping us identify reliable informants. In short, if your belief is reliably formed, it's likely to be true, so I can safely rely on it.

But there are worries with this response, too. In particular, the reliabilist who endorses such a response needs to tell us why we care about being able to identify reliable informants. Presumably the thought is that in so-doing we are able to get (more) true belief. If we didn't care about true belief, it wouldn't matter whether those we trust were accurate in what they say or not. In other words, the only reason to buy into the story about justification described above, is because we already buy into the story about the value of true belief as it spreads throughout a group. But if it is true belief that we want, then why not approve of beliefs that are unreliably formed and yet conducive to more true belief? The problem, then, with this response, is that it relies on the very notion about the value of increasing the total amount of true belief, that leads us to reject orthodox reliabilism in the first place.

[5] In his (1986) *Reliability and Cognition*, Goldman suggests this (p. 59).

6. Advice for Recovering Reliabilists

What's a recovering reliabilist to do, then? We think that she should acknowledge the consequentialist motivations that become clear in the case of social evaluation, and adopt an aggregative framework on the individual level as well. This would in effect mean that she would have to bite the bullet on the intra-personal trade-off cases that don't affect the orthodox reliabilist, on account of her denial that the justification of belief is a function of the truth-conduciveness of that belief. This will result in a counterintuitive view, to be sure, but reliabilists shouldn't be in the intuition swapping game anyway. So, the reliabilist solves the social trade-off problem by embracing the intra-personal trade-offs she could previously reject.

But, we think even if she does bite the bullet, she needs to say something about why we find the type of principles generated by reliabilism plausible. One promising route is to say something analogous to what J.S. Mill says about the virtues: the idea of forming beliefs by way of reliable processes is a good way to form beliefs on purely consequentialist grounds. So, it's a good thing that people internalize the type of principles of justification that orthodox reliabilism generates (don't engage in wishful thinking, listen to experts, proportion your beliefs to your evidence, etc.). And we can surely talk of such beliefs as justified, in that manner, but then need to keep separate the question of what beliefs are justified and what beliefs are ones we should have or form. In intra-personal trade-off cases, these two come apart: I should believe in unjustified ways, if the belief in question will lead to many true beliefs down the line.

Future Ideas

Suppose we evaluate beliefs using the aggregative framework according to which each belief is evaluated in terms of its consequences. Here is a problem for such a view. Beliefs that are rated favorably according to this standard of evaluation will not always be beliefs that we think of intuitively as justified. Above we argued that this is so much the worse for justification. But we might think there is a role that justification plays, though not solely within the epistemic realm. In particular, one might think we need some notion of justification to make sense of certain legal and moral evaluations and practices. For instance, suppose I push a large rock over a cliff and it strikes and seriously injures a hiker below. On some views, I am blameworthy only if I was justified in believing there were hikers below (or, perhaps, that there *might* be hikers below), whereas I am not blameworthy if I was justified in believing there were not hikers below. Similar examples can be constructed regarding legal culpability. The difference, for instance, between manslaughter and murder may turn on whether someone had certain justified beliefs. This suggests that some theory of justification is needed, but that it is not needed for purely epistemic purposes. This, in turn, supports our a kind of "pure" act utilitarian style consequentialism with respect to purely epistemic evaluation.

Works Cited

Ahlstrom-Vij, Kristoffer. 2015 "The Social Virtue of Blind Deference", *Philosophy and Phenomenological Research* 91: 545–582.

Berker, Selim. 2013a. "Epistemic Teleology and the Separateness of Propositions," *Philosophical Review* 122: 337-393.

Berker, Selim. 2013b. "The Rejection of Epistemic Consequentialism," *Philosophical Issues* 22: 363- 387.

Bishop, Michael. 2010. "Why the Generality Problem is Everybody's Problem", *Philosophical Studies* 151: 285-298.

Bishop, Michael and Trout, J. D. 2004. *Epistemology and the Psychology of Human Judgment*. Oxford University Press.

Cohen, Geoffrey, and Aronson, Joshua and Steele, Claude. 2000. "When Beliefs Yield to Evidence: Reducing Biased Evaluation by Affirming the Self", *Personality and Social Psychology Bulletin* 26: 1151-1164.

Conee, E. and R. Feldman. 1998. "The Generality Problem for Reliabilism", *Philosophical Studies* 89: 1-29.

Dunn, Jeff. (2012) "Reliabilism: Holistic or Simple?" *Episteme* 9: 225-233.

Dunn, Jeff and Ahlstrom-Vij, Kristoffer. *forthcoming*. "Is Reliabilism a Form of Consequentialism?" *American Philosophical Quarterly*.

Firth, Roderick. 1981. "Epistemic Merit, Intrinsic and Instrumental," *Proceedings and Addresses of the American Philosophical Association* 55: 5-23.

Foot, Philippa. 1985. "Utilitarianism and the Virtues," *Mind*, 94: 196-209.

Goldberg, Sanford. 2010. *Relying on Others*. Oxford University Press.

Goldman, Alvin. 1979. "What Is Justified Belief?" in *Justification and Knowledge*, ed., George Pappas (Springer), pp. 1-23.

Goldman, Alvin. 1986. *Epistemology and Cognition*. Harvard University Press.

Goldman, Alvin. 1992. *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. MIT Press.

Goldman, Alvin. 1999. *Knowledge in a Social World*. Clarendon Press.

Greaves, Hilary. 2013. "Epistemic Decision Theory," *Mind* 122: 915-952.

Jenkins, Carrie. 2007. "Entitlement and Rationality," *Synthese* 157: 25-45.

Kahan, Dan. 2010, "Fixing the Communications Failure", *Nature* 463: 296-297.

Kahan, Dan. 2016. "The Politically Motivated Reasoning Paradigm", *Emerging Trends in Social & Behavioral Sciences* 1-16.

Kornblith, Hilary. 2002. *Knowledge and Its Place in Nature*. Oxford University Press.

Kornblith, Hilary. 2012. *On Reflection*. Oxford University Press.

Littlejohn, Clayton. 2012. *Justification and the Truth Connection*. Cambridge University Press.

Nozick, Robert. 1981. *Philosophical Explanations*. Belknap Press.

Thomson, Judith Jarvis. 1976. "Killing, Letting Die, and the Trolley Problem," *Monist* 59: 204-217.

Williams, Bernard. 1973. "A Critique of Utilitarianism." in *Utilitarianism: For and Against*, eds., J.J.C. Smart and Bernard Williams (Cambridge University Press), 77-150.