Forthcoming in *Episteme*

ESOTERIC RELIABILISM*

Kristoffer Ahlstrom-Vij

Department of Philosophy

Birkbeck, University of London

k.ahlstrom-vij@bbk.ac.uk

Abstract: Survey data suggest that many philosophers are *reliabilists*, in believing that beliefs are justified iff produced by a reliable process. This is bad news if reliabilism is true. Empirical results suggest that a commitment to reliable belief-formation leads to overconfident second-guessing of reliable heuristics. Hence, a widespread belief in reliabilism is likely to be epistemically detrimental by the reliabilist's own standard. The solution is a form of two-level epistemic consequentialism, where an esoteric commitment to reliabilism will be appropriate for an enlightened few, while a form of epistemic fetishism—on which some heuristics are treated as fundamental epistemic norms—is appropriate for the rest of us.

1. Some Bad News

Survey data (Bourget and Chalmers 2014) suggest that almost half (42.7%) of professional philosophers in the Western world embrace externalism about epistemic justification, the view that the factors determining whether a belief is justified need not be cognitively accessible to the believer.[1] Since the most

* I'm grateful to Jeff Dunn, Klemens Kappel, Hilary Kornblith, Peter Singer, and J. D. Trout for helpful comments on earlier versions of this draft, as well as to two anonymous reviewers for this journal.

[1] The response breakdown in full is as follows: Externalism about epistemic justification at 42.7%; internalism about epistemic justification at 26.4%; other view on epistemic justification 30.8% (Bourget and Chalmers 2014: 476). As

prominent form of externalism, many of those externalists most likely embrace *reliabilism*, thereby believing some version of the claim that a belief is justified iff it is produced by a reliable belief-forming process.[2] For committed reliabilists—and that includes the present author—this survey data might at first be taken to be cause for celebration. However, I'll argue that it's actually bad news, if reliabilism is true.

As I'll explain in Section 2, the reason is that agents deliberating in reliabilist terms, and thereby qualifying as what I, borrowing from Peter Railton (1984), will be calling *subjective reliabilists*, face a *problem of defection*. More specifically, a subjective reliabilist can be expected to selectively defect from some heuristics in favor of others, when she thinks this will increase her reliability under the relevant circumstances. The problem is that empirical results suggest that, owing to our tendencies for overconfidence, such defection will often involve people second-guessing what are in fact *reliable* heuristics, resulting in poorer epistemic performance by the reliabilist's own standard than someone treating the heuristics thereby defected from as fundamental norms.

Of course, aforementioned survey data does not speak to whether those identifying themselves as externalists, many of whom we're assuming to be reliabilists, are *subjective* reliabilists in particular. Perhaps many of them are what we—again, borrowing from Railton—may refer to as *sophisticated* reliabilists, committed to reliabilism as the correct theory of justification, while remaining agnostic on the question of whether reliabilism offers a good decision-procedure. But, as we shall see below (Section 5.1), even the sophisticated reliabilist can in the great majority of cases be expected to face a problem of defection. That's why the popularity of reliabilism—be it in its subjective or sophisticated form—is bad news if reli-

---

these numbers make clear, externalism not only represents almost half of respondents, but also constitutes the largest group of the three identified in the survey.

[2] See Goldman (1979). I'm here ignoring Goldman's further qualifications in terms of belief-independent and belief-dependent processes, as well as whether justification also is a function of what processes the agent *could* and perhaps *should* have used, since they make no difference to my argument. The same goes for other, more specific versions of reliabilism include 'normal world' reliabilism (Goldman 1986), forms of reliabilism that distinguish between 'strong' and 'weak' justification (Goldman 1988), and versions embracing some form of 'mental list' virtue reliabilism (Goldman 1992).

abilism is true, particularly in light of the risk that its popularity will penetrate beyond academia and into the population at large.

Having laid out the case in Section 3 for why defection is a problem if reliabilism is true, I'll argue that the solution is a form of *two-level epistemic consequentialism*. The components of such consequentialism are outlined in Section 4, while Section 5 presents the particular type of two-level consequentialism called for in epistemology. I'll argue that it should involve an esoteric commitment to reliabilism on the part of an enlightened few, in a manner reminiscent of Sidgwick's esoteric morality, while a form of epistemic fetishism—on which some heuristics are treated as fundamental epistemic norms—is appropriate for the rest of us.

2. The Problem of Defection

Reliabilism is a theory of justification, and justification is a *normative* phenomenon, at least in the following sense: someone who believes with justification believes in the way they *should*. Consider, then, someone who both is a committed reliabilist and wants to believe in the way that they should. What advice can we give such a person? Saying 'form beliefs in a reliable fashion' offers no more by way of substantive epistemic advice than does a recommendation like 'believe what's true.' For this reason, any reasonable advice will be framed in terms of *heuristics*, or relatively practical norms of belief-formation. Think, for example, of the variety of common-sense norms ingrained in epistemic practice, such as that we ought to be open-minded in the face of disagreement, should take heed of the beliefs of our epistemic peers, and should defer to people widely recognized as experts in their fields.

Of course, a reliabilist will see these norms for what they are: heuristics, not fundamental epistemic norms. They might in many circumstances be *reliable* heuristics, and as such good ways of satisfying the fundamental, reliabilist norm of forming beliefs in a reliable fashion. But they're heuristics nonetheless, and as such vindicated (if they are) by more fundamental facts about truth-conduciveness. For that reason, the reliabilist would recommend that we always be prepared to defect from some heuristics in favor of others—even ones instantiating patterns of belief-formation that common intuitions, and the established practices they reflect, might designate epistemically *vicious*—when the latter make for more reliable belief-formation. For example, if everyone around us is an idiot, perhaps we ought *not* to be open-minded in the

face of disagreement, or take heed of the beliefs of (what you take to be) your peers. And if *we* are the idiots, then it's all the more important that we defer to people widely recognized as experts.

In that respect, the type of reliabilist we're considering here will be what we, borrowing from Railton (1984), might refer to as a *subjective reliabilist*, in consciously aiming at satisfying the reliabilist norms by following a distinctly reliabilist form of decision-making. In so doing, how likely is she to be *successful* in switching from less to more reliable heuristics, as opposed to the other way around? Not very—or so I will argue. To start with, consider *statistical prediction rules*, one of the most well-studied form of heuristic. Statistical prediction rules are simple rules, typically operating on a small number of cues, for generating judgments on a wide variety of matters. For example, by analyzing large sets of clinical data and picking out predictive cues, you can develop a prediction rule for medical diagnosis. Instead of making a diagnosis on the basis of clinical intuition—a heuristic honed by the clinician throughout their training and career—the physician can diagnose a patient by simply reading off the relevant cues and feeding the relevant values to the prediction rule, which in turn will output a diagnosis.

As it turns out, clinicians relying on such rules tend to be more successful in arriving at accurate diagnoses than those who don't, and those who only do so selectively. As noted by Robyn Dawes and colleagues (2002), there are 'nearly 100 comparative studies in the social sciences' such that, '[i]n virtually every one of these studies, the actuarial [that is, statistical] method has equalled or surpassed the clinical method, sometimes slightly and sometimes substantially' (719). And, crucially, the superiority of prediction rules is not restricted to the clinical domain. Such rules have also been shown to outperform expert criminologists in predicting criminal recidivism (Carroll *et al.* 1982), bank officers in predicting loan and credit risks (Stillwell *et al.* 1983), admissions officers in predicting academic performance (DeVaul *et al.* 1957), and forensic psychologists in predicting violent behavior (Faust and Ziskin 1988), to name but a few examples.[3]

[3] For a helpful overview of the literature on statistical predictions rules, see Bishop and Trout (2005*a*) and, more recently, Bishop and Trout (2013). For a discussion of the implications for epistemology and philosophy of science of such rules, see Bishop and Trout (2005*b*) and (2002), respectively.

It's all the more unfortunate, then, that we have a systematic tendency *not* to rely consistently on the relevant type of prediction rules. This fact has received ample attention in the literature on predictive modeling in connection with the so-called 'broken leg problem' (Meehl 1954). The problem is illustrated with reference to an imagined prediction rule that is highly reliable in predicting a person's weekly attendance at the movies, but that should be disregarded upon finding out that the person in question has a fractured femur. There is certainly something to be said for being sensitive to information not taken into account by whatever rule one happens to be relying on, especially in light of the plausible observation that no domain-specific method will be reliable in all domains. The problem is just that we tend to see far more broken legs than there really are, and thereby end up defecting from reliable heuristics far more often than we should, from an epistemic point of view (Dawes *et al.* 2002).

So why do we defect to this extent? As it turns out, it has nothing to do with prediction rules *per se*. Instead, Winston Sieck and Hal Arkes (2005) found that the reason is that we're *overconfident* about our abilities to outperform the relevant rules by increasing our reliability through selective switching between heuristics. And if what's causing defection is as *general* a phenomenon as overconfidence, there's nothing unique about statistical prediction rules, when it comes to inappropriate levels of defection; on the contrary, such levels of defection can be expected in relation to decisions about what heuristics to rely on generally. After all, it's a well-known psychological fact that, depressed people aside (Taylor and Brown 1988), we tend to rate ourselves as above average on desirable traits (e.g., Alicke 1985; Brown 1986), and this overconfidence importantly extends to our evaluations of our own epistemic capabilities, including of the extent to which we are more objective (Armor 1999) and less susceptible to bias than others (Pronin *et al.* 2002). As Emily Pronin (2007) notes in relation to the latter, 'people tend to recognize (and even overestimate) the operation of bias in human judgment—except when that bias is their own' (37).

We are now in a position to formulate *the problem of defection*: Available evidence suggests that we'll tend to *overestimate* our ability to tell when selectively switching heuristics will actually be epistemically beneficial, and on that account often switch to what are in fact *less* reliable heuristics. We'll see more clearly why this is a problem when we consider situations in which the type of defection involved isn't considered an option, on account of certain heuristics being treated as *fundamental*. After all, for any degree of reliability *r* of the relevant heuristics, anyone treating the heuristics as fundamental norms will always be

reliable to degree *r* (assuming competency in applying the heuristics). This is because treating a heuristic as fundamental involves not being willing to defect from it, since there is no more fundamental norm with reference to which we can *motivate* such defection, in the manner we can when treating some norm as a mere heuristic.[4] So, for example, if I take the application of the relevant prediction rules to *simply be the thing to do*, as opposed to *the thing to do because it will increase my reliability*, defection won't be on my map of options, in much the same way as lying isn't on the map of options for a strict Kantian, taking any injunction against lying to constitute a fundamental norm rather than one motivated on consequentialist grounds. That's also why someone defecting from one heuristic to another in her attempt to *exceed r*, will in virtue of her subjective reliabilism end up defecting when she should not—and thereby end up reliable to a degree *less* than *r*.

To illustrate, consider, again, a clinical prediction rule. A clinician might acknowledge that the rule is reliable, yet believe (overconfidently) that she will be *even more* reliable by relying on a different heuristic, in the form of her clinical intuition. The former will get many cases right, she might think, but her clinical intuition, being (in her view) more discriminating, will get even more cases right. But if she's wrong about that—as aforementioned literature on prediction models suggests that we often are—she'll be worse off epistemically for defecting from the prediction rule. Or consider a common-sense norm of the kind mentioned at the beginning of this section: defer to people widely recognized to be experts in their field. If most people recognized as experts in their field are in fact experts, someone who consistently defers accordingly on account of treating the need to do so as an epistemically fundamental norm—as (again) *simply the thing to do*—will tend to do well epistemically. We might imagine, for example, someone having taken on board a form of epistemic Confucianism, where filial piety is a virtue that imposes fairly strict hierarchies, and in particular requires that those who know less defer to those who know more. If people widely considered to be experts in fact are, such a person will likely be better off epistemically than someone who, having noted (correctly) that at least *some* experts might not deserve that honorific, tries to sort the

[4] There's an exception: if two norms, both of which the agent takes to be fundamental, conflict with one another, she has to make a choice about which norm to violate (i.e., defect from). For more on this, see footnote 15.

genuine experts from the illegitimate ones by evaluating potentially esoteric subject matter claims they have little competency in parsing, and does worse as a result.

Indeed, Linda Zagzebski highlights how this type of phenomenon crops up in a fairly striking manner in more generic, experimental settings:

> Animals like rats and pigeons maximize. If the animal discerns that one choice is better the majority of the time, it chooses that option all the time. In contrast, humans attempt to match probabilities. For instance, if humans are trying to predict whether a red rather than a green light will flash, they proportion their choices to match the probability of the mechanism. So if the light has flashed green 75 percent of the time, humans will typically predict green 75 percent of the time, whereas in similar situations, rats will *always* choose the option that appears 75 percent of the time. The rats are right 75 percent of the time. The humans do worse! (Zagzebski 2012: 115)

So, instead of sticking with a heuristic that would yield the correct verdict in the clear majority of cases—that is: 'Always go with green'—we overestimate our ability to do even better by getting the right verdict in *every* case through some more sophisticated heuristic (e.g., one attempting to map more closely onto some probability distribution), and thereby end up doing worse than someone faithfully applying the simpler heuristic. And experimental results like these are of course just further evidence that the relevant tendencies to defect from more to less reliable heuristics in our attempt to outperform the former can be expected in general, and not simply in connection with statistical prediction rules.

3. Why the Problem of Defection is a Problem

The problem of defection, as formulated above, will immediately face four challenges, as follows:

3.1. Why think the problem of defection is at all a problem?

On the first challenge, we ask: why think that the problem of defection is at all a problem? After all, defecting from a more to a less reliable heuristic will *not* to be a reliable way to form belief. So, arguably, the reliabilist gets the right verdict here: the person forming her belief as a result of the relevant form of defection will, in cases where the heuristic defected to is unreliable, end up with unjustified beliefs. But the

present investigation is not looking to argue that reliabilism is *false* by generating incorrect verdicts about individual cases. It's looking to show that, if reliabilism is *true*—and the present author believes that it is—and there also is a widespread belief to that effect, then that's bad news by the reliabilist's own standard. Consequently, for reasons that will be developed further in what follows, the problem for the reliabilist is that the truth of her account of justification cannot be widely publicized.

However, there's a better way to formulate the present objection. After all, it might be suggested that the problem of defection has an easy solution if reliabilism is true: simply take the empirical data in question as evidence that defection under the relevant circumstances is likely not to make us epistemically better off, and that we thereby ought not to defect. If we are reliabilists trying to use reliable processes, and the empirical evidence suggests that faithfully applying some particular heuristics is the way to do that, then that's what the reliabilist recommends. However, the problem with this solution is not that it would be bad if we did, but that we in many cases predictably don't. And the reason we don't is that, even when we believe—perhaps on account of being familiar with the relevant evidence—that some heuristic is in fact reliable, we often seemingly can't help ourselves also overconfidently believing that we can do even better by opting for a *different* heuristic, even if we in so doing in fact end up reducing our reliability.

Of course, it might be thought that that there's a reliabilist solution to that as well: perhaps the reliabilist who has been made aware of the prevalence and consequences of overconfidence can put in place strategies for reducing her own overconfidence, and thereby increasing the likelihood that she will not defect inappropriately. This, however, is not a particularly promising strategy in this case. To see why, consider what exactly it takes to reduce overconfidence. While several studies have suggested that overconfidence is a very recalcitrant phenomenon, typically mitigated neither by accuracy incentives nor by simple motivational declarations (e.g., Arkes *et al.* 1986; Lord *et al.* 1984), what *have* been shown to reduce overconfidence to some degree, are rigorous regiments of feedback (Arkes *et al.* 1987). The problem when it comes to extracting a remedy to the problem of defection from this fact, however, is that, in non-experimental settings, we rarely have available any data on the previous track record of the persons involved, and will therefore not be able to provide any feedback on previous successes or failures of judgments.

More to the point, even if such data *were* available, it needs to be kept in mind that not just any kind of feedback reduces overconfidence. The kind of feedback regiment that has been shown to reduce overconfidence is what Sieck and Arkes (2005) refer to as 'enhanced calibration feedback'. Such feedback involves having people (*a*) answer several questions about their degree of calibration directly after having performed the relevant judgment tasks, (*b*) consult graphical representations of how well their answers correspond to their actual degree of calibration, and then (*c*) answer several questions about what the relevant graphs suggest about their degree of overconfidence, to ensure that they understand the feedback information. In other words, while such immersive and thorough feedback can put a dent in something as recalcitrant as our tendency for overconfidence, the rigorousness of the feedback schedule required renders the practical prospects of reducing overconfidence by way of such feedback dim.

At this point, it might be objected that, for everything that has been argued so far, the reliabilist might still be better off than non-reliabilists, in at least having a working desire to be more reliable, even if not successfully overcoming their overconfidence. And if that's so, then why think that there's a *problem* of defection? But it is not clear that the reliabilist will in fact be better off than the non-reliabilist, or at least not systematically so. To see why, we need to look more closely at the relationship between being *motivated* to succeed and *in fact* succeeding. You might think that the relationship is fairly straightforward: if you try harder, you'll do better. But matters are far less straightforward than this, as can be seen from research on the effect on accuracy of inducing motivation.

Consider, for example, attempts to increase cognitive effort through by making people feel socially accountable for their judgments. As Jennifer Lerner and Philip Tetlock (1999) note in an overview of two decades of research on accountability and cognitive effort, the problem is that 'only highly specialized subtypes of accountability lead to increased cognitive effort', and that 'more cognitive effort is not inherently beneficial; it sometimes makes matters even worse' (270). We see a similar conclusion from attempts to improve epistemic performance through financial incentives. There is some evidence that incentives improve performance on simple clerical and memorization tasks. The problem is, of course, that many judgment tasks are *not* simple clerical tasks. In light of that fact, Colin Camerer and Robin Hogarth (1999) sum up the evidence on the relation between incentives and performance through the accurate but underwhelming observation that 'incentives sometimes improve performance, but often don't' (34).

The general upshot here is, of course, that motivation alone far from always translates into success. And why should we expect it to? Being motivated to do well doesn't mean having the ability to do so. For that reason, we also can't assume that becoming a reliabilist, and as such developing a working desire to become more reliable, will in fact make you systematically more successful on that score.

But there's a more fundamental point to be made here. What if becoming a reliabilist *did* make you (systematically) more successful in achieving true belief, compared to being a non-reliabilist? Would it not then follow that a widespread endorsement of reliabilism wouldn't be a problem after all, contrary to the contention of this paper? No, but in order to see that, we need to dig a bit more deeply into the nature of and motivation for reliabilism, which is best done by considering a further objection.

3.2. Are reliabilists committed to promoting justification?

In order for defection to generate any type of problem, we need to have agents move from overconfidently believing that 'I can satisfy norm *N* better by defecting from heuristic *H* to *H**' to believing that 'I should defect from *H* to *H**', and finally to defecting accordingly. In the case of the reliabilist, we would be dealing with agents overconfidently believing 'I will be more reliable by defecting from *H* to *H**' to believing 'I should defect from *H* to *H**', and then acting accordingly. But, it might be objected, reliabilism doesn't carry with it any commitment to act or believe one way or the other. Reliabilism simply tells you that, when (and only when) you form beliefs on the basis of reliable processes, those beliefs are justified. More generally, to offer necessary and sufficient conditions for justification doesn't commit you to a thesis on which justification ought to be in any way *promoted*.

There is a sense in which this observation is completely on point. Remember that we set up the problem of defection by considering someone who both is a committed reliabilist *and* wants to believe in the way that they should. What the current challenge is pointing out is that the former commitment doesn't entail the latter. But notice the following: you can accept that justification should be promoted *in addition* to some set of necessary and sufficient conditions for justification—which is exactly what reliabil-

ists do (for good reason, as I shall argue in a moment).[5] More specifically, reliabilists embrace what I'll refer to as *the promotion thesis*, which is the thesis that we should take steps to increase our reliability.[6] The promotion thesis explains why we find within the reliabilist camp not only a definition of justification in terms of reliability, but also a variety of attempts to say constructive things about how to go about *increasing* our reliability on an individual as well as social level, be it by offering suggestions for what experts to trust (Goldman 2001), what reasoning-strategies to use (Bishop and Trout 2005), how to re-design epistemic environments in ways that protect people from bias (Ahlstrom-Vij 2013*b*), how to raise the level of reliability in science, law, education, and democratic institutions (Goldman 1999), or how implementing certain incentive structures might help increase the reliability of the scientific community (Kitcher 1990), to name but five examples.[7]

Importantly, this is not to state a mere sociological fact. If you're a reliabilist you *should* embrace the promotion thesis. After all, consider what would drive you to embrace reliabilism in the first place. It would arguably be the conviction that true belief is a good—an *epistemic* good, specifically—and that we

---

[5] While framing the point in a slightly different manner, this is what Conee and Feldman (2004) are picking up on when suggesting that they are simply in the business of '[stating] a necessary and sufficient condition for epistemic justification' (86) that 'has no implication about the actions one must take in a rational pursuit of the truth' (90), and that this makes for a contrast with Goldman, in that 'the principles he is discussing are guides to action' (86).

[6] Note that the promotions thesis involved need not involve *maximization*, be it of reliability or whatever normative phenomenon happens to be concerned. For present purposes, all that's required is a *minimal* promotions thesis, on which we *sometimes* should bring about *more* of the target property. After all nothing more than this minimal thesis is required to generate defection, involving people moving from 'I can satisfy norm *N* better by defecting from heuristic *H* to *H**' to believing that 'I should defect from *H* to *H**', and acting accordingly.

[7] Of course, if aforementioned discussion of the prevalence of overconfidence is on point, then we should be skeptical of the particular prescriptions that rely on individual agents attempting to epistemically improve themselves (Ahlstrom-Vij 2013*c*)—which of course doesn't take away from the fact that philosophers defending such strategies offers evidence of their acceptance of the promotion thesis, which is all that matters here.

on that account should bring about more of it.[8] Call this *the fundamental motivation for reliabilism*. This motivation accounts *both* for the idea of taking justification to be a matter of reliable processes (i.e., reliabilism) and for the desire to increase our reliability (i.e., the promotion thesis). But if that's so, it's not clear that anyone could plausibly embrace reliabilism while rejecting the promotion thesis. To embrace reliabilism is to be moved by its fundamental motivation, which motivates *both* reliabilism and the promotion thesis. Rejecting the latter would involve also rejecting the fundamental motivation—which in turn would mean leaving one's commitment to reliabilism without a motivation.

It might be objected that someone can accept reliabilism on grounds that have nothing to do with aforementioned motivation. At least some philosophical views are accepted simply on account of doing a good job of systematizing our pre-theoretical intuitions about the phenomenon defined.[9] But note that reliabilism is particularly badly suited for such a justification. It was noted in Section 2 that the reliabilist treats any common-sense norm about how to go about believing things as a heuristic that will stand or fall with reference to how well it actually does in terms of reliability (or truth-conduciveness more generally). It follows that her views—much like her consequentialist cousins in ethics—will in many cases be *revisionary*. For example, if reliabilism is true, then we should under certain circumstances reflect on our beliefs far less than what common sense deems appropriate (Kornblith 2012); defer to others blindly, and without any concern for their epistemic credentials (Ahlstrom-Vij 2015); and form our beliefs in ways traditionally considered epistemically irresponsible (Bishop 2000). In all of these cases, the reliabilist is pointing out that norms that have come to be considered fundamental are ultimately mere heuristics, and should be substantially revised once their epistemic merits have been properly evaluated. However, were the reliabilist not able to appeal to the fundamental motivation for her view—and, in particular, to a shared desire for true belief—her revisionary remarks could rightly be dismissed as unmotivated, exactly on account of clashing with widespread intuitions.

---

[8] This in no way entails a commitment to promoting truth *always and everywhere*, as opposed to when, say, we have non-epistemic (for example, moral) reason to do so. After all, taking true belief to be the fundamental *epistemic* value is compatible with taking the domain of epistemic evaluation to be parasitic on more fundamental values or domains, such as moral ones. (Ahlstrom-Vij 2013*b*)

[9] Many thanks to an anonymous reviewer for this journal for raising this point.

This goes to show that the problem of defection is indeed a problem if reliabilism is true. On account of the promotion thesis, and contrary to the objection considered here, reliabilism is accompanied by a commitment to act or believe in particular ways. For that reason, we are in connection with reliabilism indeed dealing with agents overconfidently believing 'I will be more reliable by defecting from *H* to *H**' to believing 'I should defect from *H* to *H**', and then acting accordingly—and, hence, the problem of defection.

3.3. Isn't the reliabilist still better off for being a reliabilist?

Return now to the objection introduced at the end of Section 3.1: so long as the reliabilist is still likely to outperform the non-reliabilist, widespread acceptance of reliabilism is not appropriately described as a problem. Section 3.1 offered some reason to be skeptical that the reliabilist is likely to outperform the non-reliabilist (or at least consistently so). But even if that were in fact the case, the objection would still be misguided. After all, given the promotion thesis introduced in Section 3.2, such a state of affairs would simply shift the reliabilist's concern to the fact that the subjective reliabilist is doing significantly *less* well than she could have done, had she not fallen prey to the problem of defection. As noted already in Section 1, the concern is about poorer epistemic performance by the reliabilist's own standard than someone treating the heuristics thereby defected from as fundamental norms. And the wider the pool of people embracing subjective reliabilism, the wider the pool of people that are falling short of the level of epistemic performance they could've been at, had they not defected at inappropriate rates—*that's* the problem we face, if reliabilism is true, and the (potentially questionable) premise of the current objection is, too.

Indeed, at this point it's helpful to look ahead at the solution to the problem of defection that will eventually be defended (in Section 5): a form of *esoteric reliabilism*, where a commitment to reliabilism will be appropriate for an enlightened few, while a form of epistemic fetishism on which some heuristics are treated as fundamental epistemic norms is appropriate for the rest of us. This solution is, of course, inspired by Henry Sidgwick's (1981/1874) analogous defense of esoteric *morality*. This parallel is relevant here because, in his reflections on the relationship between utilitarianism and common-sense morality, Sidgwick in effect takes an analogous position to the one I'm suggesting the reliabilist should take in relation to the objection under consideration.

According to Sidgwick, common sense morality is a form of 'latent utilitarianism' (438). For that reason, while someone sticking to common sense will be prone to making certain mistakes in attempting 'to correct the estimate of common opinion by the results of his own experience in order to obtain from it trustworthy guidance for his own conduct' (152), it would be wrong to suggest that common sense—representing 'the net result of combined experience' (151)—doesn't offer *some* valuable practical guidance. In that respect, we can think of widespread subjective reliabilism of the kind we've suggested would be a problem here, on the model of Sidgwick's latent utilitarianism. And when we do, we can also reframe the objection under consideration as follows: if common sense were to develop into a form of latent reliabilism, why be bothered by the fact that people are likely to go astray epistemically on account of overconfidence, so long as they (we are assuming for the sake for argument) are still better off than they would have been, had they not had a fundamental commitment to reliabilism?

Sidgwick's answer in relation to latent utilitarianism is clear: the problem with common sense morality is that it doesn't offer 'the *best* guidance we can get to the attainment of maximum general happiness' (463; my emphasis). Moreover, in so far as a morality is 'imperfectly felicitous,' as he believes common-sense morality to be, despite being latently utilitarian, it 'require[s] considerable improvement from a Utilitarian point of view' (465). Indeed, in so far as the 'actual moral order is admittedly imperfect, it will be the Utilitarian's *duty* to aid in improving it' (476; my emphasis). The point of this comparison with Sidgwick's ethics is, of course, that the promotion thesis we find with the reliabilist would have her give a similar response—indeed, the very response provided at the beginning of this section. Again, it should be stressed that that response would only be called for, in the event that we have reason to believe that the reliabilist will in fact be systematically better off from an epistemic point of view compared to non-reliabilists. And as noted in Section 3.1., it's not clear that this premise is plausible.

3.4. Is the problem of defection unique to those committed to reliabilism?

This brings us to the final challenge, on which it's maintained that what's doing the work in generating the problem of defection has nothing to do with any commitments unique to the reliabilist and everything to do with *overconfidence*. For any epistemic norm—be it reliabilist or not—overconfidence will lead us to overestimate our ability to satisfy that norm without the aid of heuristics and thereby also lead to too fre-

quent defection. And if that's so, even if selective defection raises a problem for the reliabilist, we do not have any reason to worry about the extent to which people believe reliabilism in particular. To the extent that there's a problem of defection, it's not a problem that's unique to those committed to reliabilism.[10]

Two things should be noted in response. First, even if this challenge were to stand, it doesn't affect the overall conclusion of this paper: if reliabilism is true, there's a problem of defection that—as we shall argue later on—is to be solved with reference to a form of two-level consequentialism. Second, the challenge doesn't stand. While we might certainly see defection in relation to any type of norm, it doesn't follow that there is in all such cases a *problem* of defection. The reason is that such a problem only arises for views that come with a commitment to *truth-conduciveness* in particular, where a process is truth-conducive in so far as it is reliable (i.e., generating a high ratio of true beliefs) and powerful (i.e., generating a lot of true beliefs).[11]

Why think that? Say your view comes with a commitment to promoting some particular thing, independently of its contribution to truth-conduciveness. To make matters more concrete, say that it comes with a commitment to promoting *doxastic coherence*, which doesn't imply a higher likelihood of truth, even *ceteris paribus* (Olsson 2005). In your attempt to promote coherence norms, you then overconfidently defect from heuristics that would actually help you increase coherence in favor of norms that have you do the opposite. As a result, you do a worse job of attaining coherence than you would have done had you just stuck to the original heuristic. What we're dealing with here is certainly a case of defection. But we are not dealing with a *problem* of defection.

This is because, owing to considerations having nothing to do with reliabilism or its alternatives, we have reason to believe that true belief is the fundamental, epistemic goal (Ahlstrom-Vij 2013). If that's so, someone is doing well epistemically to the extent that they're forming beliefs on the basis of processes

---

[10] Thanks to Peter Singer and J. D. Trout for both raising versions of this objection.

[11] Why define truth-conduciveness in terms of reliability *and* power? Because 'truth-conduciveness' seems a good term for what we want in so far as we are moved by the fundamental motivation for reliabilism, while someone forming belief on the basis of processes that are generating no belief (reliability but no power) or very large number of beliefs with a high proportion of false belief (power but no reliability) doesn't seem to be faring particularly well in relation to what we thereby want.

that are reliable and powerful. As a consequence, anytime someone defects from one heuristic in favor of another, we're only dealing with a *problem* of defection in so far as the relevant conduct impacts negatively on their truth-conduciveness. By contrast, when someone is doing worse in attaining some goal that's *unrelated* to truth-conduciveness, such as coherence in the case above, a person defecting from norms promoting that goal will not necessarily be less reliable or powerful, and as such epistemically worse off, for so doing.[12] In fact, depending on the exact relationship between the non-truth-linked goal in question and truth-conduciveness, she might even be *better* off. This will be the case, for example, when the former goal and truth-conduciveness are negatively correlated.

This goes to show that, if your view *only* postulates norms relating to the achievement of some goal that's unrelated to truth-conduciveness, defection from those norms won't make for a *problem* of defection. But, of course, some views involve taking truth-conduciveness to be necessary but not sufficient for justification. Indeed, don't some reliabilist do exactly this? No, what some reliabilists do is say that *the reliability of the belief-forming process involved* is necessary but insufficient for justification. For example, Goldman (1986: 62-63 and 111-112; see also 1988: 54) includes a non-undermining condition in his account of justification, requiring that the person must not believe that her belief was formed in an unreliable manner. The resulting account departs from the simple type of reliabilism we started out with, by not taking it that a belief is justified if reliably formed. However, since the condition is arguably motivated by the thought (whether true or not) that a person that forms beliefs by reliable processes and *also* meets the non-undermining condition will be *more* reliable than one that merely forms beliefs by reliable processes, we don't here have a view that takes truth-conduciveness to be insufficient for justification. Rather, truth-conduciveness remains the underlying motivation for the inclusion or exclusion of candidate criteria for justification. Indeed, Goldman (1986) is explicit about the fact that he wants a theory of justification that

[12] Such a person can of course still be said to be worse off *by their own lights*. But in cases where their own lights fail to shine on whatever factors actually determine whether someone is better or worse off epistemically—as suggested earlier, those factors relate to truth-conduciveness—it does not follow from this that they are in fact worse off. Consider an analogy: if the Stoics are right, a person is doing well to the extent that she's virtuous, but that's compatible with someone who cares deeply about social status, wealth, and the like, being worse off by her own lights in not attaining those things. The Stoics would simply deny that she is *in fact* worse off.

is 'truth-linked' (69; see also 1992: 164), and that '[t]rue belief is the value that J-rules [rules dictating which beliefs are justified] should promote—really promote—if they are to qualify as right' (103). This, of course, is all in accordance with what we earlier called the fundamental motivation for reliabilism.

This is not to deny that *non-reliabilists* who deny (*a*) that true belief is the (sole) ultimate epistemic good and, as such, also (*b*) that the reliabilist's fundamental motivation captures the full range of goods to be promoted, might hold that truth-conduciveness is necessary but insufficient for justification. In so far as a commitment to those views also includes some form of promotion thesis, it might be suggested that a problem of defection will likely arise, too. And it will—with, and only with, respect to norms that can be cashed out in terms of truth-conduciveness. But the fact remains that, while anyone will run the risk of facing the problem of defection with respect to any norms of truth-conduciveness that they might embrace, only those committed to reliabilism will face that problem *for every single norm* that she endorses.[13]

To sum up, the problem of defection is only a problem on views that come with a commitment to promoting truth-conduciveness in particular. Only for such a view will it be the case that, when we overconfidently defect from heuristics helping us satisfy some particular norms, and as a result do a worse job in actually satisfying those norms, we are also doing *epistemically* worse. The reliabilist, as we have seen, embraces a view of this kind. Given the fundamental motivation for reliabilism, the reliabilist embraces the promotion thesis. What is to be promoted on that thesis is reliability, and someone who fails to believe in a way that's reliable thereby also fails to believe in a way that's truth-conducive.[14] As a result, when we overconfidently defect from reliable norms in favor of less reliable norms, we are epistemically worse off, which warrants talking about the defections involved as making for a *problem* of defection. Other views will face that problem as well to, and only to, the extent that they embrace norms of truth-conduciveness, but only the reliabilist will face it with respect to every single norm that she endorses. So, while the problem is not unique to the committed reliabilist, the reliabilist faces it in its most serious form.

---

[13] Strictly speaking, the same would go for a view on which you are (doxastically) justified in believing that *p* iff *p* and you believe that *p*. However, since that view is highly implausible on independent grounds, I am ignoring it here.

[14] Of course, if we insist that, on any plausible reading of what it is to be a reliabilist, the relevant promotion thesis should invoke not only reliability but also power, then that merely reinforces my point that the problem of defection is a problem—and a particularly pressing one at that—for the reliabilist.

4. Two-level Consequentialism

We're now in a position to see why it's bad news that reliabilism is a popular view in epistemology, if reliabilism is true. As we have seen, subjective reliabilism combined with overconfidence can be expected to lead to the second-guessing of reliable heuristics, and, as a result, poorer performance than that of someone treating the heuristics thereby defected from as fundamental. In that respect, it makes reliabilist sense for people *not* to deliberate in reliabilist terms, and instead treat the relevant reliable heuristics as fundamental norms. In the remainder of this paper, I want to suggest that this line of reasoning moreover motivates a kind of *two-level consequentialism* in epistemology, preventing any reliabilist commitment on the part of epistemologists from spreading to the general population, lest subjective reliabilism becomes a prevalent phenomenon. But first I need to say a few words about what two-level consequentialism is.

Two-level consequentialism can take a number of forms, but on the notion relevant here, all forms embrace the following:

> 1. LEVEL INDEPENDENCE: There's a distinction between two levels of deliberation, a philosophical one attempting to pin-point *what's right*, and a practical one attempting to decide *what to do*—either in relation to *moral* conduct, or in relation to decisions about how to go about conducting *inquiry*—and the two levels might operate independently of one another.

Take the classic utilitarian, by way of illustration. What's right is a function of maximizing utility. But it doesn't follow that people ought to walk around attempting to maximize utility. As Sidgwick (1981/1874) puts the point, '[i]t is not necessary that the end which gives the criterion of rightness should always be the end at which we consciously aim' (413). We find this idea also in John Stuart Mill (1987/1861), who suggests that 'it is a misapprehension of the utilitarian mode of thought to conceive it as implying that people should fix their minds upon so wide a generality as the world' (290). What's morally right is indeed a matter of such a wide generality, according to Mill, but it doesn't follow that people ought to deliberate in such terms when deciding what to do.

As far as I can tell, Mill leaves it open whether the utilitarian has reason to positively *encourage* people to deliberate in non-consequentialist terms, or whether people doing so is merely *compatible* with utilitarianism. As such, it's not clear that Mill embraces the following:

> 2. ENCOURAGED LEVEL SEPARATION: It often makes consequentialist sense to keep the two levels separate and to *encourage* people not to deliberate in consequentialist terms when deciding what to do.

We find a version of this idea in R. M. Hare (1981). On the *critical* level, as Hare calls it, we evaluate what's right in act-utilitarian terms, while, on the *intuitive* level, we rely on common intuitions. Moreover, these common intuitions, according to Hare, 'are sound ones, if they are, just because they yield acceptable precepts in common cases. For this reason, it is highly desirable that we should all have these intuitions and that our conscience should give us a bad time if we go against them' (49). Now, Hare, as I read him, thinks of the relevant level separation in both *intra*-personal terms, as involving at least in some cases an attempt to encourage a separation of levels within *ourselves*, and in *inter*-personal terms, as encouraging such a separation in *others*. Others have been concerned with one but not the other. For example, Railton (1984) is concerned with an intra-personal separation when suggesting that 'a *sophisticated consequentialist* is someone who has a standing commitment to leading an objectively consequentialist life, but who need not set special stock in any particular form of decision making and therefore does not necessarily seek to lead a subjectively consequentialist life' (153) in the sense of 'follow[ing] a distinctively consequentialist mode of decision making' (152). Railton makes clear that, particularly in cases where consequentialist deliberation might be *self-defeating*—preventing the person from doing well on consequentialist terms—a person may wish to develop habits that *encourage* an intra-personal separation between levels of deliberation within him- or herself. (We'll have more to say about the psychological plausibility of this proposal in Section 5.1.)

Sidgwick, by contrast, famously embraces an *inter*-personal version of ENCOURAGED LEVEL SEPARATION:

> […] it may be desirable that Common Sense should repudiate the doctrines which it is expedient to confine to an enlightened few. And thus a Utilitarian may reasonably desire, on Utilitarian principles, that some of his conclusions should be rejected by mankind generally; or even that the vulgar should keep aloof from his system as a whole, in so far as the inevitable indefiniteness and complexity of its calculations render it likely to lead to bad results in their hands (Sidgwick 1981/1874: 490).

Note that Sidgwick isn't simply saying that non-utilitarian principles should be encouraged among the many—that is, he's not merely embracing an inter-personal version of ENCOURAGED LEVEL SEPARATION—but also that this practice should be kept a *secret*. Hence:

> 3. SECRET LEVEL SEPARATION: The fact that it often makes consequentialist sense to keep the two levels separate and to encourage people not to deliberate in consequentialist terms when deciding what to do should be kept a *secret* outside of an enlightened few.

This should give the reader a good sense of what two-level consequentialism is. It's a view that at the very least commits its defenders to LEVEL INDEPENDENCE, while some go further in also embracing an inter- or intra-personal reading (or both) of ENCOURAGED LEVEL SEPARATION, and in some cases also SECRET LEVEL SEPARATION. In the next section, we'll consider how this applies in the epistemic domain in light of the problem of defection.

5. Two-level Epistemic Consequentialism and Epistemic Fetishism

What type of two-level consequentialism does the problem of defection motivate in epistemology? Remember, the upshot of the problem was that subjective reliabilism combined with overconfidence will lead to the second-guessing of, and selective defection from, reliable heuristics in favor of less reliable ones, and as a result poorer performance compared to that of people treating the former heuristics as fundamental. In light of that, it was suggested that it makes reliabilist sense for people *not* to deliberate in

reliabilist terms when deciding how to go about conducting inquiry. Let's consider which of the components of two-level consequentialism outlined in the previous sections are thereby called for.

5.1. Level Independence and Encouraged Level Separation

We may start by noting that the solution to the problem of defection requires LEVEL INDEPENDENCE, and in particular that there's a level of deliberation about *what's right*, and one about *what to do*. And while the former might be a function of the consequences of choosing one heuristic over another on one's reliability, it doesn't follow that this fact should play any substantial role in the minds of people deliberating about how to go about conducting inquiry. In that sense, the two levels are *independent*.

More than that, the formulation arguably also calls for some form of ENCOURAGED LEVEL SEPARATION. In particular, if reliabilism is true, it will often make sense to *encourage* people to deliberate about how to go about conducting inquiry in non-reliabilist terms, in order to discourage people from overconfidently misjudging when they will in fact be better off for switching heuristics. But should we read that component *intra*-personally or *inter*-personally? Given that we borrowed the notion of a *subjective* reliabilist from the type of intra-personal level separation we find in Railton (1984), *sophisticated* reliabilism might be taken to be the way to go. A sophisticated reliabilist has a standing commitment to doing well by reliabilist standards, without being in any way committed to taking decision procedures formulated in reliabilist terms to be the best means towards meeting those standards. However, the relevant, intra-personal reading would be problematic, at least if understood to apply generally (we'll have more to say about a restricted version in a moment). After all, having a standing commitment to reliabilism, in a manner analogous to the case of sophisticated moral consequentialism, too easily invites the temptations to selectively defect from (reliable) heuristics that gave rise to the problem of defection in the first place. To remove the risk of such a commitment bleeding through to the practical level, the intra-personal separation would need to be so complete, and the relevant commitment be relegated so far back in the mental life of the relevant person, that it no longer makes sense to talk about her as having a standing commitment to (objective) reliabilism. We can formulate this point in terms of a dilemma: either such a person faces the problem of defection, or she is no longer a sophisticated reliabilist, since lacking the relevant commitment.

An inter-personal reading of ENCOURAGED LEVEL SEPARATION, by contrast, does *not* invite any such temptations, since no intra-personal separation between levels is necessary on that reading. Still, that reading might be taken to invite other worries. One worry can be framed in terms of yet another dilemma: Either (*a*) there still needs to be an intra-personal separation among *some*—the enlightened few, as Sidgwick might say—or (*b*) *no one* can believe that reliabilism is true, even if it is. Borrowing a term from discussions of utilitarianism, the latter horn here can be described as requiring that reliabilism be (completely) *self-effacing*. Both horns come with their own challenges, but there doesn't seem to be a way to go between them. So, which horn should we go for?

Let's consider (*b*) first. Isn't self-effacement a problematic feature of a theory? As Derek Parfit (1984) points out:

> […] to be self-effacing is not to be self-defeating. It is not the aim of a theory to be believed. If we personify theories, and pretend that they have aims, the aim of a theory is not to be believed, but to be true, or to be the best theory. That a theory is self-effacing does not show that it is not the best theory (Parfit 1984: 24).

Similarly, for reliabilism (and epistemic consequentialism generally): thinking back to the type of promotion thesis that we suggested accompanies reliabilism, we might say that the reliabilist cares, not about people being committed reliabilists, but about people being reliable (or, perhaps, truth-conducive more generally). And if it turns out that the best way to promote reliability is by having people embrace some other theory than reliabilism—such as one that treats some reliable heuristics as fundamental—then that's what the reliabilist is going to want to see happen.

At the same time, while not leading to self-defeat, such self-effacement is arguably problematic on purely consequentialist grounds. After all, consider the sheer number of principles that someone might go for, if not for straightforwardly reliabilist ones. Some of these are surely going to be reliable heuristics, and as such ones sanctioned by a reliabilist evaluation. But many of them—whether part of our commonsense epistemic practice, or academic epistemology—will not be. Moreover, since less than perfectly reliable, there will be many situations in which the heuristics will yield conflicting advice. If there's self-

effacement across the board, there will be no one around to ensure that the heuristics embraced are actually reliable, and that any conflicts between heuristics are generally resolved in a manner that's in fact conducive to reliability.[15]

As far as I can see, this renders (*b*) an unworkable option. The solution to the problem it faces is, clearly, something short of complete, across-the-board self-effacement. This brings us to (*a*), on which there's an intra-personal separation among some, perhaps in what Sidgwick refers to as the enlightened few. This, however, might be taken to bring us right back to the dilemma posed above for the sophisticated reliabilist, to the effect that such a person either faces the problem of defection, or she is no longer a sophisticated reliabilist. But this fails to be a genuine dilemma in this case. As it turns out, overconfidence is not evenly distributed—the more you know, the more aware you are of your limitations. As Justin Kruger and David Dunning (1999) put the point, 'competence begets calibration' (1127), in the sense of one's perceptions of one's ability mirroring one's actual ability. That's because becoming more informed robs you of the ignorance that not only results in bad judgments and choices but also prevents you from realizing how badly you're actually doing. This enables us to deny that the enlightened few will fall prey to the temptations generating the problem of defection. On account of being relatively free of overconfidence, sophisticated reliabilism *is* an option for them.

5.2. Secret Level Separation

I don't have anything substantive to say on the issue of *who* the enlightened few are in this case, beyond noting that, while all enlightened epistemologist arguably will be reliabilists, it's highly unlikely that all reliabilists are enlightened in the sense required for proper calibration. Of course, even that point aside, go-

---

[15] The latter job—of seeing to it that conflicts between heuristics are resolved in a way that makes for reliable belief-formation—gets to the exception called attention to in footnote 4. From the point of view of the individual agent treating those heuristics as fundamental, the situation will be akin to the one we would be in were intuitionism true, involving, on Rawls' (1999) characterization, 'an irreducible family of first principles which have to be weighed against each other' (30). This will likely require either developing priority rules, or, if that makes for a too complex system of norms (in turn making compliance less likely), the development of a small set of heuristics for which conflicts are highly unlikely—all with an eye towards safe-guarding reliability overall.

ing for (*a*) might be taken to bring us straight from the frying pan into the fire—because doesn't it commit us to SECRET LEVEL SEPARATION?

It does, and some might consider this a fatal objection to the line of reasoning pursued here. Consider what's probably the most well-known objection to such secrecy: Bernard Williams's concerns about 'Government House' consequentialism (e.g., in Sen and Williams 1982). Williams's target was Sidgwick's call for secrecy over the truth of consequentialism beyond the enlightened few, and as such for an esoteric morality. Of course, we might think that there's a relevant difference here between secrecy in morality and secrecy in epistemology, with the latter being less problematic than the former. A more forceful response to Williams's worry, however, would also call into question the very idea that there's anything in principle problematic about secrecy in morality. As pointed out by William Langenfus (1989), the colonial metaphor of a Government House is doing significant work in convincing us that there's a problem. But rejecting the colonial metaphor isn't sufficient for a satisfactory reply. As Katarzyna Lazari-Radek and Peter Singer (2014) point out: 'Some people think that the fact that something *would* be wrong if it were done openly shows that it *is* wrong, even if done in secret' (300). This is so on account of a principle of *publicity*, on which norms are only legitimate if they can be disseminated publicly—and esoteric morality can't, of course, if it's to remain esoteric. Two observations will help us determine whether this principle constitutes an obstacle to the type of *esoteric reliabilism* considered here.

First, John Rawls (1999), perhaps the most prominent defender of a publicity principle, claimed that it applies 'for the choice of all ethical principles' (112).[16] As such, it is not clear that the relevant publicity principle applies to the choice of *epistemic* principles—at the very least, we cannot simply assume that all principles applying in the domain of ethics apply also in the domain of epistemology. Second, even if we were to assume as much, then consider what reasons we have (if any) for accepting a publicity condi-

[16] Rawls defends a particularly strong version of the publicity principle, involving a need for me to know (*a*) everything I would know about the relevant rules if they were the result of an agreement; (*b*) what the rules demand of me and of others; and (*c*) that others know the same, that they know that I know these things, and so on (see Rawls 1999: 48). Needless to say, any reason to reject the weaker principle that rules or norms need to be such that they can be publicly disseminated (a necessary condition on coming to know the things Rawls requires) would count against this stronger publicity principle.

tion in the first place. Rawls himself found it 'reasonable' (112). As noted by Samuel Scheffler (1982), however, 'the consequentialist can simply deny that the condition "seems reasonable" to him, and force the discussion back to an overall comparison of the consequentialist and non-consequentialist conceptions' (47-48). And there's no reason to construe such a move on the part of the consequentialist as a sign of simple intransigence. After all, as noted by Railton (1984), 'any such condition [of publicity] would be question-begging against consequentialist theories, since it would require that one class of actions—acts of adopting or promulgating an ethical theory—not be assessed in terms of their consequences' (155). In that respect, insisting on a publicity criterion in discussions about consequentialism is possibly downright *un*reasonable.

Of course, SECRET LEVEL SEPARATION might be objectionable for reasons having nothing to do with publicity. We might, for example, feel that we're somehow violating the *autonomy* of those not let in on the relevant secrets. Something along these lines might be taken to be driving Brad Hooker's (2002) complaint that esoteric morality would constitute 'paternalistic duplicity' (85). Let's break down this complaint into two parts, one about *paternalism*, and one about *duplicity*. As for the former, it is not clear that the actions of the enlightened reliabilists would involve interfering with inquirers *for their own good*, as opposed to for the sake of others who might be at risk in so far as people make ill-informed or ignorant decisions. If that's so, then the actions in question are not paternalistic. Of course, if the enlightened reliabilists were to interfere with people for their own epistemic good—and moreover be doing so without seeking their consent—then that would indeed amount to a form of *epistemic* paternalism. However, as I have argued elsewhere (Ahlstrom-Vij 2013*b*), such interference is fully compatible with a proper appreciation both of people's *personal* autonomy and of their *epistemic* autonomy. So that part of Hooker's complain seems without warrant in the case of esoteric reliabilism.

Would the relevant form of SECRET LEVEL SEPARATION still be objectionable on account of being *duplicitous*? As noted by Langenfus (1989: 487), there are likely to be at least some cases in which the enlightened few will be able to influence the rest of us by way of argumentation and reason, rather than through interference. (How *many* cases? That would depend on how susceptible we generally are to rational argumentation as a matter of empirical psychology.) Since such argumentation cannot be conducted in reliabilist terms, there would still be some degree of duplicity involved on the part of the enlight-

ened few. Would that be *objectionable*? Not necessarily, and we can see this more easily if we consider a concrete example.

Say that we are devising a curriculum for budding scientists and that we, as enlightened reliabilists (let's assume), have identified some particular set of norms *N* as being highly reliable. At the same time, in light of the problem of defection, we realize that we cannot do both of the following: (*a*) teach *N*; and (*b*) frame our teaching of *N* in terms of their reliability. Instead, we might do either of the following. We teach *N* but frame our teaching of the component norms in ways that don't motivate them with reference to their reliability. For example, we might motivate them by suggesting that they are epistemically fundamental, and as such simply capture what's constitutive of epistemically responsible conduct. Or, if we decide this is unlikely to work—for example, people are unlikely to consider the norms in *N* as plausible candidates for fundamental norms—we might opt for a different set of norms *N** that is potentially less reliable than *N*, but that stands a better chance of being such that people can be persuaded that they are indeed fundamental.

Whichever scenario we end up in—either one where we're teaching the set we consider to have the most epistemic merit, but don't teach it with reference to that fact; or one where we're teaching some different set than the one we consider to have the most epistemic merit—there is some degree of duplicity involved in what we are doing. Specifically, there is some truth of the matter, either about the motivation for the norms in question, or about their true (relative) merits, that we are withholding. Is this objectionable? It's not clear that it would be. Whether on the scale imagined here, 'duplicity' of this kind is a frequent aspect of most educational settings. For example, we often teach a less complex and in many respects strictly speaking false account of things because doing so will bring a greater number of students towards understanding than does teaching something in all its complexity. This, too, will involve some degree of duplicity since being explicit about what we're doing—that we're teaching less complex version because they're unlikely to grasp the complex one—will likely not be conducive to having them (at least) embrace what we know to be the less complex version.

More generally, whether such duplicitous behavior is objectionable would seem to depend on some combination of the consequences of and motivations behind the relevant arrangements. We have covered both of these aspects in the imaginary case just considered: if we are right about the merits of the relevant

curriculum, there will be good consequences, in that the scientists involved will walk away significantly better off from an epistemic point of view than they arrived. And the motivation behind the curriculum design would, arguably, be unobjectionable as well: what's being done is intended to be for the benefit of those at the receiving end, as well as for society at large, since we all have an interest in a society characterized by reliable belief-formation.

5.3. Epistemic Fetishism

Moreover, if we're successful in such educational settings more generally, the result will be a strong commitment on the part of others to treating as fundamental certain heuristics that will enable reliable belief-formation, and minimize the risk of inappropriate defection. Exactly at what level of abstraction these heuristics are to be formulated is an empirical matter, contingent upon what leads to best performance. Returning again to the type of common-sense norms we started out with, we can imagine, for example, making the case for the (epistemically) fundamental status of norms about being open-minded in the face of disagreement, or taking heed of our peers, by inculcating the idea that we have a moral duty to hear people out, on account of their being fellow agents to whom we owe a rich set of obligations. Similarly, as noted already in Section 2, we might do the same for norms about deferring to people who are widely recognized as being experts within their fields, perhaps by making a plausible case for a form of epistemic Confucianism, where filial piety requires that those who know less defer to those who know more.

Again, whether norms like these—whether in granularity or in content—are the ones that it would make sense to treat as epistemically fundamental, and with reference to what type of (non-reliabilist) story, is an empirical matter, and not one that I intend to settle here. The point is simply that *some* heuristics are to be treated in this manner. To signify that the relevant commitments to those heuristics are moreover to be both *strong* and resting on *false beliefs* about the relevant principles being fundamental as opposed to mere (reliable) heuristics, we may term the attitude thereby taken by the many to the relevant heuristics as one of *epistemic fetishism*. In the meanwhile, discussions regarding the fundamental epistemic norms, and of reliabilism in particular, could continue, so long as restricted to academic journals that are unlikely to reach any particularly wide audience, to ensure that the type of 'comparative secrecy' that Sidgwick

(1981/1874: 490) had in mind for the ultimate criterion of rightness, and for the fact that the criterion is esoteric, is maintained.

It was in this context, of course, that we started out this investigation by noting that the popularity of reliabilism within the philosophical community is bad news if reliabilism is true, given the risk that its popularity will penetrate beyond academia to the population at large. Consequently, any feeling on the part of the reader that reliabilism—esoteric or not—is false is a not altogether unwelcome reaction to the present investigation. As Lazari-Radek and Singer (2014) put the point in their defense of esoteric morality, '[y]ou should be reluctant to embrace esoteric morality, and you should feel strongly that there is something wrong with our conclusion'; indeed, 'your resistance is […] the "right" response, in the sense that it is good that you should have that response' (316). The same goes for my defense of esoteric reliabilism—to a certain extent, skepticism regarding its viability is altogether appropriate.

Why only 'to a certain extent'? On account of two important qualifications:

First, compatible with the claim that widespread acceptance of reliabilism among professional epistemologists is bad news—again, given the risk of its popularity spreading to the population at large—is the observation that an across-the-board rejection within the professional epistemological community would also not be a good thing. For the reasons outlined above, there is a need for at least *some* enlightened people to embrace reliabilism, and to do what they can to promote the fetishizing of *appropriate* heuristics in society at large. As already made clear above, I have no particular theory about who those people are, beyond noting that, while all enlightened epistemologists will arguably be (esoteric) reliabilists, it's unlikely that all reliabilists will be enlightened. Still, the success of the rest of us in light of the problems outlined in this paper hangs on there being at least some reliabilists who fit the bill.

Second, for the enlightened few to at all have an impact on what norms are to be fetishized in society, there also needs to be a subset of the non-philosophical community that is both friendly to the cause of esoteric reliabilism and in a position to facilitate the relevant epistemologists influencing relevant institutions and policy. This, naturally, conjures images of Williams's 'Government House' consequentialism. However, as already argued, for those images to indicate the presence of a sound objection, we would have to (*a*) embrace an epistemic publicity principle that the reliabilist has no more reason to accept than do her utilitarian cousins; (*b*) make the case that the relevant arrangements are paternalistic *and* objection-

able; or (*c*) argue that the arrangements are duplicitous *and* objectionable. In the previous section, each of these three strategies were found wanting. For that reason, we may conclude that, while the practical blue-print for enlightened reliabilists to have the relevant influence is still to be provided, there does not seem to be any reason to believe that implementing such a blue-print would have to be objectionable, morally or otherwise.

6. Conclusion

We started out by noting that recent survey data suggesting that many philosophers are likely reliabilists is bad news if reliabilism is true. This is so on account of how a commitment to reliable belief-formation combined with a tendency for overconfidence leads to an inappropriately high frequency of second-guessing of reliable heuristics, with poor epistemic performance as a result. In that respect, widespread belief in reliabilism is likely to be epistemically detrimental by the reliabilist's own standard. The solution, I argued, is a form of two-level epistemic consequentialism, where an esoteric reliabilism will be appropriate for an enlightened few, while a type of epistemic fetishism—on which some heuristics are treated as fundamental epistemic norms—is appropriate for the rest of us.

References

Ahlstrom-Vij, K. (2015) 'The Social Virtue of Blind Deference,' *Philosophy and Phenomenological Research* 91(3): 545-582.

Ahlstrom-Vij, K. (2013*a*) 'In Defense of Veritistic Value Monism,' *Pacific Philosophical Quarterly* 94(1): 19-40.

Ahlstrom-Vij, K. (2013*b*) *Epistemic Paternalism: A Defence*, Basingstoke: Palgrave Macmillan.

Ahlstrom-Vij, K. (2013*c*) 'Why We Cannot Rely on Ourselves for Epistemic Improvement,' *Philosophical Issues* (a supplement to *Noûs*) 23: 276-296.

Alicke, M. D. (1985) 'Global Self-Evaluation as Determined by the Desirability and Controllability of Trait Adjectives.' *Journal of Personality and Social Psychology* 49.

Arkes, H. R., Christensen, C., Lai, C., and Blumer, C. (1987) 'Two Methods for Reducing Overconfidence', *Organizational Behavior and Human Decision Processes* 39: 133-44.

Arkes, H. R., Dawes, R. M., and Christensen, C. (1986) 'Factors Influencing the Use of a Decision Rule in a Probabilistic Task', *Organizational Behavior and Human Decision Processes* 37: 93-110.

Armor, D. (1999) 'The Illusion of Objectivity: A Bias in the Perception of Freedom from Bias.' *Dissertation Abstracts International: Section B: The Sciences and Engineering* 59: 5163.

Bishop, M. (2000) 'In Praise of Epistemic Irresponsibility: How Lazy and Ignorant Can You Be?' *Synthese* 122: 179-208.

Bishop, M. and Trout, J. D. (2002) '50 Years of Successful Predictive Modeling Should be Enough: Lessons for the Philosophy of Science.' *Philosophy of Science* 68 (Proceedings): S197-S208.

Bishop, M. and Trout, J. D. (2005*a*) *Epistemology and the Psychology of Human Judgment*, Oxford: Oxford University Press.

Bishop, M. and Trout, J. D. (2005*b*) 'The Pathologies of Standard Analytic Epistemology.' *Noûs* 39 (4): 696-714.

Bishop, M. and Trout, J. D. (2013) 'Diagnostic Prediction and Prognosis: Getting from Symptom to Treatment' in William Fulford (ed.), *The Oxford Handbook of Philosophy and Psychiatry*. New York: Oxford University Press.

Bourget, D. and Chalmers, D. (2014) 'What Do Philosophers Believe?' *Philosophical Studies* 170: 465-500.

Brown, J. D. (1986) 'Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments.' *Social Cognition* 4.

Camerer, C. F., and Hogarth, R. M. (1999) 'The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework', *Journal of Risk and Uncertainty* 19: 7-42.

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J., and Alibrio, J. J. (1982) 'Evaluation, Diagnosis, and Prediction in Parole Decision Making,' *Law & Society Review* 17(1): 199-228.

Dawes, R., Faust, D., and Meehl, P. (2002) 'Clinical versus Actuarial Judgment' in T. Gilovich, D. Griffin, and D. Kahneman (eds), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 716–29). Cambridge: Cambridge University Press.

DeVaul, R. A., Jervey, F., Chappell, J. A., Carver, P., Short, B., and O' Keefe, S. (1957) 'Medical School Performance of Initially Rejected Students,' *Journal of the American Medical Association* 257: 47-51.

Faust, D., and Ziskin, J. (1988) 'The Expert Witness in Psychology and Psychiatry,' *Science* 241: 1143-44.

Feldman, R. and Conee, E. (2004) *Evidentialism: Essays in Epistemology*, Oxford: Oxford University Press.

Goldman, A. (1979) 'What Is Justified Belief?' in G.S. Pappas (ed.) *Justification and Knowledge* (pp. 1-23). D. Reidel Publishing Company.

Goldman, A. (1986) *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.

Goldman, A. (1988) 'Strong and Weak Justification,' *Philosophical Perspectives* 2: 51-69.

Goldman, A. (1992) 'Epistemic Folkways and Scientific Epistemology,' in his *Liaisons: Philosophy Meets the Cognitive and Social Sciences* (pp. 156-175). Cambridge, MA: MIT Press.

Goldman, A. (2001) 'Experts: Which Ones Should You Trust?' *Philosophy and Phenomenological Research* 63(1): 85-110.

Goldman, A. (2011) 'Toward a Synthesis of Reliabilism and Evidentialism? Or: Evidentialism's Troubles, Reliabilism's Rescue Package' in T. Dougherty (ed), *Evidentialism and Its Discontent* (pp. 254-280). Oxford: Oxford University Press.

Hare, R. M. (1981) *Moral Thinking: Its Levels, Method and Point*. Oxford University Press.

Hooker, B. (2002) *Ideal Code, Real World*, Oxford: Clarendon Press.

Kitcher, P. (1990) 'The Division of Cognitive Labor,' *Journal of Philosophy* 87(1): 5-22.

Kornblith, H. (2012) *On Reflection*, Oxford: Oxford University Press.

Kruger, J. and Dunning, D. (1999) 'Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments,' *Journal of Personality and Social Psychology* 77: 1121-1134.

Langenfus, W. L. (1989) 'Implications of a Self-effacing Consequentialism,' *Southern Journal of Philosophy* 27(4): 479-493.

Lazari-Radek, K. and Singer, P. (2014) *The Point of View of the Universe: Sidgwick & Contemporary Ethics*, Oxford: Oxford University Press.

Lerner, J. S., and Tetlock, P. E. (1999) 'Accounting for the Effects of Accountability', *Psychological Bulletin* 125(2): 255-75.

Lord, C. H., Lepper, M. R., and Preston, E. (1984) 'Considering the Opposite: A Corrective Strategy for Social Judgment', *Journal of Personality and Social Psychology* 47(6): 1231-43.

Meehl, P. (1954) *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.

Mill, J. S. (1987) 'Utilitarianism' in *Utilitarianism and Other Essays*. Penguin; originally published in 1861.

Parfit, D. (1984) *Reasons and Persons*. Oxford University Press.

Pronin, E. (2007) 'Perception and Misperception of Bias in Human Judgment.' *Trends in Cognitive Science* 11.

Pronin, E., Lin, D., and Ross, L. (2002) 'The Bias Blind Spot: Perceptions of Bias in Self Versus Others', *Personality and Social Psychology Bulletin* 28: 369-81.

Railton, P. (1984) 'Alienation, Consequentialism, and the Demands of Morality,' *Philosophy and Public Affairs* 13(2).

Rawls, J. (1999) *A Theory of Justice*, Revised Edition. Cambridge, MA: Harvard University Press.

Scheffler, S. (1982) *The Rejection of Consequentialism*. Oxford: The Clarendon Press.

Sen, A. and Williams, B. (1982) 'Introduction' in *Consequentialism and Beyond* (pp. 1-22). Cambridge University Press.

Sidgwick, H. (1981) *The Methods of Ethics*, 7th edition. Hackett Publishing Company; originally published in 1874.

Sieck, W. and Arkes, H. (2005) 'The Recalcitrance of Overconfidence and its Contribution to Decision Aid Neglect,' *Journal of Behavioral Decision Making* 18.

Stillwell, W., Barron, F., and Edwards, W. (1983) 'Evaluating Credit Applications: A Validation of Multiattribute Utility Weight Elicitation Techniques,' *Organizational Behavior and Human Performance* 32: 87-108

Taylor, S. E., and Brown, J. D. (1988) 'Illusion and Well-being: A Social Psychological Perspective on Mental Health' *Psychological Bulletin* 103: 193-210.

Zagzebski, L. (2012) *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press.