Philosophical Issues
A Supplement to NOÛS

Philosophical Issues, 23, Epistemic Agency, 2013

WHY WE CANNOT RELY ON OURSELVES FOR EPISTEMIC IMPROVEMENT

Kristoffer Ahlstrom-Vij
University of Kent, Canterbury

Abstract

There is something very appealing about the idea that we are epistemic agents. One reason—if not the *main* reason—is that, while we are undoubtedly fallible creatures, us being epistemic agents that *do* things means that it might just be within our power to improve and thereby do *better*. One important way in which we would want to improve is in relation to our well-established tendency for cognitive bias. Still, the proper role of epistemic agency in us avoiding or correcting for cognitive bias is highly limited. In fact, what we know from empirical psychology—particularly with respect to our tendencies for overconfidence—suggests that we cannot rely on ourselves for epistemic improvement, and have good reason to impose significant constraints on our ability to exercise such agency in ameliorative contexts.

1. What is Epistemic Agency?

As human beings, we are not merely passive recipients of information. We interact critically with our surroundings, mull over our data, reflect on the merits of our beliefs given our evidence, collect more information when we feel that is needed, consult others who we believe to be informed on the relevant matters, and so on. In conducting inquiry thus, we are *doing* things, and are as a result properly called *agents*. More specifically, we are *epistemic* agents, in that we are doing things in pursuit of specifically epistemic goals.

This tells us something about what epistemic agency is. It tells us that the domain of epistemic agency encompasses all the things that we do in the pursuit of specifically epistemic goals. At the same time, describing epistemic

agency at this level of abstraction leaves unanswered several substantive questions. Two questions are particularly relevant here:

First, what are the *epistemic goals* we are pursuing as epistemic agents? This question has rightly received a lot of attention in recent epistemology.[1] For present purposes, I will assume that the answer is as follows: There is one and only one epistemic goal, and that is the dual goal of attaining true belief and avoiding false belief. This is by no means an uncontroversial answer. However, since I have defended it at length elsewhere, and rehearsing the relevant arguments here would take us too far afield, I will simply assume that this is the correct answer for present purposes.[2]

Second, what are the *actions* relevant to epistemic agency? One common answer understands them in relation to our ability to *reflect* on our first-order beliefs. For example, Richard Moran suggests that our abilities for self-knowledge should be explained with reference to how deliberating on our first-order beliefs 'involves a sense of agency and authority'.[3] This is so on account of how such deliberations involve 'stepping back'[4] from a belief for the purpose of making up one's mind by answering the question 'Shall I believe?' through 'a decision or commitment of oneself'.[5] While not primarily concerned with self-knowledge, Ernest Sosa, too, sees a connection between agency and reflection, on account of how 'reflection aids agency, control of conduct by the whole person, not just by peripheral modules', by providing a way to 'holistically [. . .] strike a balance' through 'an assessment of the respective weights of pros and cons' in cases where reasons are in conflict.[6]

The idea that reflection is central to epistemic agency has recently come under fire by Hilary Kornblith.[7] According to Kornblith, once we factor in the absence of a difference between first-order and higher-order belief-formation as far as our abilities for doxastic control are concerned, as well as what we know from psychology about the limits of our reflective capacities, the appeal to epistemic agency seems 'nothing more than a bit of mythology', since a 'demystified view of belief acquisition leaves no room for its operation'.[8] However, it seems perfectly possible for someone attracted to the idea of epistemic agency to grant that reflection might not be as central to epistemic agency as some philosophers have suggested, while denying that this implies that there is no such thing as epistemic agency. If we return to the rough characterization of the kind of activities in virtue of which we are epistemic agents that we started out with, then it is not clear that the appeal to reflection is a necessary component of the case for epistemic agency. We do many things as epistemic agents, and reflect on our beliefs is one, but by no means the only (let alone the most important) thing that we do.

In light of this, there is one type of account of epistemic agency that remains largely unaffected by Kornblith's critique, namely a type of account that takes epistemic agency to encompass the full range of things that we do in our pursuit of epistemic goals, including collecting and evaluating information, choosing among different methods of investigation, and so on.

This is the type of account we will be concerned with in the below. Moreover, in considering the merits of this kind of account, it is important to keep in mind why we bother with epistemic agency in the first place. There is something very appealing about the idea that we are epistemic agents. One—if not the *main*—reason is that, while we are undoubtedly fallible creatures, us being epistemic agents that *do* things means that it is within our power to improve, e.g., by thinking longer and harder, collecting more information, re-considering our methods, and so on. That is, epistemic agency enables us to do *better*—or so we hope.

Differently put, one important thing we want from epistemic agency is *epistemic improvement*. But improvement with respect to what? Our fallibility manifests itself in a variety of ways. In so far as we want to improve, however, our main focus should be on the ways in which we fail *systematically*, rather than accidentally. Consequently, one central and important way in which we should want to improve epistemically is by avoiding and correcting for *cognitive bias*, systematic and by now well-established tendencies to form inaccurate beliefs.[9] In other words, while we undoubtedly are subject to cognitive bias, us being epistemic agents means that there might just be things that we, *qua* agents, can do to avoid or correct for bias.

It is the purpose of the present paper to evaluate the prospects for this idea of epistemic agency as central to our attempts to do something about our tendencies for bias. The plan will be to consider two approaches to de-biasing that rely on things that the agent can do herself, namely correct for bias that has occurred (§2), or try to prevent bias from occurring in the first place (§3). It will be argued that both approaches face significant challenges, and that our best bet when it comes to avoiding bias is to impose external constraints on our freedom to conduct inquiry in whatever way we see fit (§4). This, moreover, suggests that epistemic agency is not something we have reason to consider particularly valuable in relation to our attempts to improve epistemically (§5). It also suggests that the problems with epistemic agency go beyond those identified by Kornblith in relation to reflection. In contexts of epistemic improvement, the problem with epistemic agency is not reflection *per se*, but the far more general problem that we cannot rely on ourselves for epistemic improvement, be it through reflective means or otherwise.

2. On the Limits of Self-Correction

The worries raised by the fact that we often reason in biased ways would be greatly diminished if it were easy for us to mobilize our epistemic agency for the purpose of either correcting for bias after the fact, or preventing bias from occurring in the first place. After all, it might be that all we need to do about our biased ways is to be more careful and vigilant in going about our

epistemic business. The present section will consider one suggestion as to how to do this, focusing on the prospects for the agent correcting for bias on her own accord. It will be argued that the prospects for this suggestion are dim.

2.1. The Problem of Motivation

It is a well-established psychological fact that—depressed people aside[10]—most of us tend to rate ourselves as above average on desirable traits,[11] especially in contexts of abstract targets of comparison (for example, the average colleague, compatriot, college student, and so on) and with respect to traits that leave room for ambiguity in interpretation.[12] The extent to which one deems oneself to be more objective, insightful, and less biased than one's peers is no exception on this score. For example, in a series of five studies, each involving a different measure of objectivity, David Armor found that approximately 85 per cent of the participants rated themselves as more objective than the average member of the group from which they were drawn.[13] Similarly, in a series of three studies by Emily Pronin, Daniel Lin, and Lee Ross, college students rated themselves as less susceptible to each of a number of described biases compared both to the average American and to their peers in a seminar class.[14] To rule out that the relevant effect was an artifact of the possible arrogance of students at prestigious universities, Pronin and colleagues replicated the results at an international airport. As Pronin notes in a recent overview, the upshot of the data collected on this 'bias blind spot' is that 'people tend to recognize (and even overestimate) the operation of bias in human judgment—except when that bias is their own'.[15]

The obvious problem with the beliefs thereby guiding people's conception of the relative merits of their competence over that of others is, of course, that we cannot all be above average. Consequently, a significant proportion of us must be mistaken about our own relative insusceptibility to bias, suggesting that the relevant self-other asymmetry in fact reveals a tendency for *overconfidence* in the accuracy of our judgments. Indeed, the relevant kind of overconfidence has been independently revealed in calibration studies, investigating the extent to which our degrees of confidence track our actual abilities to get the relevant matters right. As it turns out, most of us are not very well-calibrated, in that we have a tendency to express a greater degree of confidence in our answers than is warranted by the extent to which we actually tend to get the relevant kind of questions right.[16] Interestingly, this tendency can be found not only among lay people, but also among scientists, where the relevant phenomenon manifests itself in a tendency to underestimate the likelihood of errors.[17]

As a result, the main problem facing any attempt to bring about epistemic improvement through the epistemic agent taking certain corrective measures is *not* that there are no corrective measures to take.[18] The problem

is that, while we might see the point of others committing themselves to corrective measures, each and every one of us will tend not to see the point of doing so ourselves. In this respect, the above results on our tendencies for overconfidence in our epistemic capabilities present a *motivational* problem for the idea of self-correction.

Can this motivational problem be overcome? Emily Pronin and Matthew Kugler have provided some reason to believe that one of the primary factors driving our tendency to see others as more susceptible to bias than ourselves is our penchant for relying on introspective information when determining whether or not we are subject to bias.[19] Since the processes that give rise to bias tend to operate on a sub-personal level, outside the scope of our introspective gaze, our search tends to come up empty. From the fact that our search comes up empty, we then infer an absence of bias—despite the fact that such a search is more or less *guaranteed* to come up empty, given the inaccessibility of the relevant mechanisms. Pronin and Kugler also found, however, that subjects who, prior to evaluating the extent to which they were susceptible to a variety of biases compared to others, were asked to read an article highlighting the introspective inaccessibility of large parts of our mental lives, ceased to claim that they were less susceptible to bias than their peers.

When evaluating whether this presents a way to get around the problem of motivation, however, we need to take a closer look at a crucial ambiguity in Pronin and Kugler's results. In the studies demonstrating a self-other asymmetry, subjects are typically presented with a presentation of a certain tendency, and then asked to what extent they believe that they show that tendency on a nine-point scale (1 = *not at all*, 5 = *somewhat*, 9 = *strongly*), and to what extent they believe that the average member of some relevant group (for example, the average American or class mate) shows that tendency on the same scale. In cases of bias blind spots, their evaluation of themselves tends to fall around five, while their evaluation of others tends to fall around seven. In the study where educating people about the limits of introspection was shown to remove their bias blind spot, however, the subjects were asked to rate themselves *relative to the average member of the relevant reference class* on an 11-point scale (for example, 1 = *much less than the average student*, 6 = *same as the average student*, 11 = *much more than the average student*). Those who had been educated about the limits of introspection prior to being asked the relevant question tended to answer 'six'.

What this result leaves open, however, is if the relevant subjects on average rating themselves as being equally susceptible to the relevant biases as the average student would rate themselves with a seven on the nine-point scale—that is, move from a fairly optimistic to a more pessimistic evaluation of themselves—or rate others as five—that is, move from a fairly pessimistic evaluation to a fairly optimistic evaluation of others. If the latter, then the relevant results would go no way towards identifying a way to avoid the problem of motivation. On this reading, the effect in question would *not*

involve people becoming more worried and as a result more motivated when it comes to counteracting bias in themselves; it would simply involve people becoming less cynical about bias in others. In fact, on this reading of the results, it seems that implementing a practice of educating people about the limits of introspection would give rise to an even more serious problem of motivation, since the relevant subjects would go from not being particularly worried about bias in their own beliefs, to not being particularly worried about bias in *anyone's* beliefs!

If the relevant educational effect leads to a more pessimistic evaluation of oneself, however, it might stand a better chance of presenting a way around the problem of motivation. More specifically, assuming that an increased concern for bias in one's own beliefs results in an increased motivation to do something about such bias, the relevant results might just be suggesting a way to make people more motivated to think about ways to avoid or correct for bias. Consequently, if the self-correction approach under consideration is modified in a manner that combines (*a*) an emphasis on the self-correction of bias with (*b*) an acknowledgement that agents might need to be subjected to externally imposed educational programs regarding the limits of introspection to at all become sufficiently motivated to correct thus, it might be able to avoid the problem of motivation. As we shall see in the next section, however, this falls short of showing that self-correction constitutes a promising approach to de-biasing, which brings us to the second problem facing that approach.

2.2. The Problem of Proper Correction

Let us assume that there are ways to bring agents to worry enough about bias in their own beliefs for them to become motivated to mobilize their epistemic agency by taking steps to correct for such bias. Of course, merely being *motivated* to correct for bias is not enough. Additionally, we need to be able to do so *successfully*. In order to correct successfully, however, we need to do (at least) two things. First, we need to correct for bias when and only when we are biased. In so doing, we face a challenge of bias *identification*. This challenge, in turn, consists in avoiding both *necessary but neglected corrections* and *unnecessary corrections*. Second, we need to correct to and only to the extent necessary to counteract the relevant bias. In so doing, we face a challenge of *correction*, namely that of avoiding both *insufficient corrections* and *overcorrections*.[20] Let us refer to the problem associated with meeting both of these challenges as *the problem of proper correction*, and consider each challenge in turn.

As for the challenge of bias identification, consider what kind of data the agent has available to her as to whether her beliefs are biased, and as such are in need of correction. One thing she may do is to *look inwards*, that

is, to introspect and then reflect upon the relevant beliefs and the manner in which they were formed. As we have already noted above, however, an immediate problem with looking inwards to identify bias is that we typically do not have introspective access to the sub-personal processes giving rise to our beliefs, including biased belief, nor consequently any ability to reflect on the merits of the relevant processes.[21] What we do have introspective access to is often simply the *outputs* of those processes, that is, the beliefs themselves. As pointed out by Timothy Wilson and Nancy Brekke, however, bad judgments, unlike bad food, do not smell.[22] Consequently, any introspective search for indications of bias will in many cases come up empty. Acting on this information alone might, of course, result in the agent failing to identify biases that have actually occurred, and thereby also failing to meet the challenge of necessary but neglected correction.

That said, a person who is educated about the limits of introspection, in accordance with the results by Pronin and Kugler above, might of course refrain from taking the absence of introspective signs of bias as reliable evidence on the issue of whether the relevant beliefs actually are biased. Indeed, humbled by her new insights about the limits of introspection, the agent may instead *look outwards* to determine whether her beliefs might be biased, and pay particular attention to others' warnings about the possibility of bias. Heeding such warnings, she might avoid the mistake of failing to neglect for bias when bias has occurred. But, in so doing, she faces a different challenge, namely that of not also 'correcting' beliefs that were not biased to start with—an outcome observed by Richard Petty and Duane Wegener in a study on people's success in correcting for contrast effects.[23] Prompted by a de-biasing cue, the relevant subjects 'corrected' for a bias that they had not exhibited, thereby failing to meet the challenge of unnecessary connection.

Let us assume, however, that our self-correcting agent manages to meet the challenge of bias identification, and thereby corrects for bias in all and only in cases where the beliefs in question are actually in need of correction. While an impressive feat in itself, it unfortunately carries no guarantee that she will have rid her beliefs of bias. As noted above, ridding our beliefs from bias requires not only correcting all and only biased beliefs, but also correcting to and only to the degree needed. Here, it is interesting to note that Petty and Wegener also found evidence of *insufficient* correction, in cases where a more subtle de-biasing cue was used. The subjects in this subtle-cue condition did, indeed, correct. However, they did not correct to an extent that served to rid their judgments of the bias induced by the contrast effect. In other words, they failed to meet the insufficient correction challenge.

In attempting to avoid insufficient correction, it is of course possible to go too far in the other direction by correcting too much, and thereby biasing one's beliefs in the opposite direction. This would amount to failing the fourth and final challenge, namely that of overcorrection. Evidence for overcorrection has been found in studies on priming effects, where people's

ratings of others shift in the direction of the primed category. For example, making the category 'kind' particularly salient has people rate others as kinder than they otherwise would have done. Studies have shown that making people aware of that their judgments have been primed by arbitrary categories sometimes has them adjust their judgments. However, the adjustment far from always removes the priming effect—indeed, in many cases the result is an overcorrection amounting to a contrast effect.[24] For example, if the prime is 'kind', the 'correction' involves subjects ending up rating the evaluated person as *less* kind than they would have in the absence of the prime, and thereby failing the fourth and final challenge, that of overcorrection.

None of what has been argued above is to suggest that it is *impossible* for people to meet the challenges of proper correction. After all, the above examples of ways in which people have failed to meet the relevant challenges do not go to show that the relevant challenges cannot be met—of course it is possible for people to meet them. But when we think about the merits of different approaches to dealing with bias, we are not concerned so much with the (merely) possible as with the probable, that is, with what not only *could* work but also is *likely* to work. The burden of the above has been to show that, even if we assume that the relevant agents are at all motivated to engage in bias correction—and as we saw in the previous section, this is far from a trivial assumption—there are substantial challenges they need to meet when it comes to doing so successfully.

Moreover, it is important to note that this is not simply a problem for an approach along the educational lines considered at the end of the previous section. To the contrary, it is a problem for any approach that attempts to solve the problem of bias simply by addressing the problem of motivation. For example, we might attempt to increase people's accuracy of judgment, not by having them worry directly about cognitive bias as such, but by instead making them feel socially accountable for their judgments, on the assumption that the motivational effects of social benefits will make them invest greater cognitive effort in the relevant judgment tasks. However, as Jennifer Lerner and Philip Tetlock note in an overview of two decades of research on the relation between 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'.[25] And in light of the above, this should come as no surprise: When it comes to our attempts to successfully rid ourselves of bias, the fact that there are more ways to get it wrong than right suggests that motivation alone cannot be assumed to make for proper correction.

The same goes for approaches that attempt to address the problem of bias by way of financial incentives. For example, there is some evidence that incentives improve performance on simple clerical and memorization tasks.[26] The problem is, of course, that many judgment tasks are not simple

clerical tasks. Consequently, Colin Camerer and Robin Hogarth sum up the available evidence on the relation between incentives and performance by noting that 'incentives sometimes improve performance, but often don't'.[27] And no wonder. Again, there are several more ways to get things wrong than right when it comes to successful correction. To get things right, we must not only manage to correct when and only when we are actually biased; additionally, in those cases where we are actually biased, we must also correct exactly to the extent needed to undo the effect of the relevant bias, neither more nor less. As illustrated by the examples discussed above, this is quite a tall order—in fact, such a tall order that it seems a fair bet to suggest that self-correction is not likely to provide a particularly promising recipe for the correction of bias.

3. On the Limits of Self-binding

The previous section called into question one way of mobilizing our epistemic agency for the purpose of doing something about our biased ways, namely by trying to correct for bias that has already occurred. But perhaps there is some other, more promising way to avoid bias through exercising one's epistemic agency. After all, one way to avoid the problem of proper correction is to focus on the *prevention* rather than on the correction of bias. Indeed, this is the conclusion drawn by Timothy Wilson, David Centerbar, and Nancy Brekke. Having discussed some of the same problems of self-correction discussed above, and noted that 'just because people attempt to correct a judgment they perceive to be biased [that] is no guarantee that their result will be a more accurate judgment', they conclude that '[t]he best way to avoid biased judgments and emotions is exposure control'.[28]

The idea of dealing with corrupting influences, not by undoing their influence when the damage is already done, but by avoiding the relevant influences in the first place, has been pursued with great sophistication by Jon Elster.[29] The poetic metaphor invoked by Elster to illustrate the relevant approach is that of Ulysses, tying himself to the mast of his ship in anticipation of the sirens that will lure him to steer his ship into the rocks, unless he renders himself unable to fall to the temptation of heeding their songs. Indeed, Elster makes an intriguing case for the idea that we may predict and prevent undue influence of passion, self-interest, and so on, by making commitments that remove or make more costly tempting yet potentially detrimental future options. Following Elster, let us refer to the taking of measures such as this one as *self-binding*.

Importing Elster's framework of self-binding into the epistemic domain, we get the suggestion that the epistemic agent, on her own accord, may deal with bias by making commitments that either remove the option of engaging in activities known to prompt biased reasoning or, failing that,

make engaging in the relevant activities significantly more costly than not doing so. Let us consider an example, to make matters more concrete.

3.1. Self-binding in the Epistemic Domain: Prediction Models

It is well known that clinicians are susceptible to a variety of biases when making clinical judgments on the basis of clinical intuition. At the same time, there is a large body of research demonstrating that clinicians and other professionals can avoid bias and thereby significantly increase their reliability by relying on so-called *prediction models*. A prediction model is typically developed by (*a*) running a regression analysis over large sets of data, (*b*) picking out the factors identified by such an analysis as predictive, and (*c*) incorporating those factors in a simple algorithm. For example, a linear prediction model would look like this:

$$V = w_1c_1 + w_2c_2 + \cdots w_nc_n$$

V represents the predicted value of a target property, c_1 through c_n a set of cues, and w_1 through w_n the weights assigned to those cues. The first evidence of the superiority of judgments made on the basis of such surprisingly simple models compared to unaided clinical judgments came when Paul Meehl famously reviewed 22 studies comparing the judgments of expert psychologists and psychiatrists with judgments based on nothing but the outputs of linear models.[30] In all studies, those making judgments solely on the basis of the outputs of linear models either performed equally well or outperformed the clinicians. Following up on Meehl's study twelve years later, Jack Sawyer reviewed 45 studies comparing clinical and statistical predictions via linear models.[31] Again, not in a single study were the former superior to the latter. The studies conducted since Meehl and Sawyer's reviews come out equally in favor of prediction models. As noted by Robyn Dawes, David Faust, and Paul Meehl, there are now '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'.[32]

This is, of course, excellent news for every clinician interested in making accurate diagnoses and prognoses. Indeed, in light of the prevalence of bias, the limits of self-correction—that is, of correcting for bias after the fact, in the manner discussed in the previous section—and the presence of highly reliable prediction models, one self-binding measure that clinicians could take is that of committing to using prediction models, whenever such models are available for the relevant clinical judgments. There are two questions that we need to ask about such self-binding, however, the first one being: Is it *possible* for clinicians to self-bind in this manner, for the purpose of protecting

their clinical judgments from bias? Here, the answer is clearly 'yes'. While it is, arguably, not possible to literally make unavailable the option of relying on one's clinical intuition, the relevant commitment could take the form of clinicians not only agreeing to but—since this is supposed to be an instance of *self*-binding by motivated epistemic agents—*insisting on* being subject to certain sanctions if they fail to rely on the relevant models in making their judgments. (I will not speculate on what particular sanctions would be the most effective ones here.)

As noted in the previous section, however, when we think about the merits of different approaches to dealing with bias, we are not concerned so much with what is possible as with what is probable, which brings us to the second question: Is it *likely* that clinicians will self-bind in this manner, for the purpose of protecting their clinical judgments from bias? The next section suggests that the answer is 'no'. Yet again, the problem will turn out to be a *motivational* one: Relevant psychological research suggests that most of us will tend not to see the point of self-binding, on account of being overconfident in the accuracy of our judgment. To return to Elster's metaphor, unlike the superhuman king of Ithaca, most of us would never make it to the mast, let alone see the point of having ourselves be tied to it.

3.2. Another Problem of Motivation

Since its inception, the main challenge of predictive modeling has been, not the development of reliable prediction models—given a sufficiently large set of data on which to run a regression analysis, that can be done fairly easily—but getting people to actually utilize them. Researchers talk about the so-called 'broken leg problem'.[33] The problem is illustrated with reference to an imagined prediction model that is highly reliable in predicting a person's weekly attendance at a movie, 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 models one happens to be relying on. The problem is that people tend to see far more broken legs than there really are, and, thereby, also defect from reliable models far more often than they should, from an epistemic point of view.[34]

Moreover, evidence suggests that the major culprit behind such defection is a phenomenon we have already discussed in the above, namely that we are systematically overconfident about the accuracy of our judgments. While several studies have suggested that overconfidence is a very recalcitrant phenomenon, typically mitigated neither by accuracy incentives nor by simple motivational declarations,[35] what *have* been shown to reduce overconfidence to some degree, however, are rigorous schemes of feedback.[36] Building upon this fact, Winston Sieck and Hal Arkes found that such feedback not only

lowered people's overconfidence in their judgments in so far as they did not rely on the statistical models offered, but also led to greater reliance on those models as well as improved performance, as compared to a control group.[37] This gives us reason to believe that overconfidence is an important, albeit not necessarily the only, cause of defection.

Moreover, the fact that we defect on account of overconfidence has important implications for the viability of the self-binding approach. Granted, if the extent of the evidence was that we are prone to defection, that might not have presented a problem for the idea of avoiding bias by mobilizing our epistemic agency through self-binding. On the contrary, it could arguably have provided an argument *for* self-binding, that is, for taking steps to bind oneself in ways that renders ineffective one's predictable tendencies to defect from the relevant models when it comes to employing them. But if the source of defection is as widespread and general a psychological phenomenon as overconfidence in the accuracy of our judgment, then it is not clear that we can expect individual agents to have sufficient motivation to self-bind in the first place. To see why, consider the following:

First, if defection is the result of such a *widespread* psychological phenomenon as overconfidence, we can expect that a great majority of us—not just clinicians—will be prone to defection. Second, if defection is the result of such a *general* psychological phenomenon as overconfidence in the accuracy of our judgment, it seems reasonable to infer that people are not defecting on account of any feature peculiar to prediction models. Rather, available data give us reason to believe that people are defecting simply because they think that they can do better without relying on the models. That is to say that, anyone who is likely to defect on account of overconfidence, is *also* likely to be such that she would lack the motivation to commit to the relevant models in the first place. Taking the two points together, we see that, if overconfidence causes defection—as Sieck and Arkes' results suggest—overconfidence is likely to prevent a great majority of us, and not just clinicians, from self-binding in a manner that commits us to taking protective measures in the first place, be it by relying on prediction models or otherwise.

In other words, much like in the case of the self-correction approach to bias, the self-binding approach faces a problem of motivation: Since we have a tendency for overconfidence in the accuracy of our judgments, we are unlikely to see the point of self-binding. As in the case of self-correction, however, it might also be suggested that the problem can be overcome. More specifically, it might be suggested that one way to come to terms with problems of defection is by focusing our attention on strategies that temper our often inflated confidence in our own accuracy. For example, it might be suggested that one way around problems of defection is to self-bind in two steps, so to speak. In the first step, one commits to being subjected to feedback on the accuracy (or inaccuracy) of previous judgments. In the second step, after having been sufficiently motivated by grasping one's lack of calibration, one

then commits to using the relevant preventive strategies. Let us refer to this as *the sophisticated self-binding approach*.

There are two problems with this approach. First, in non-experimental cases, we might not have available any data on the previous track record of the subjects in question, and will therefore not be able to provide any feedback on previous successes or failures of judgment. Second, 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 that has been shown to reduce overconfidence is what Sieck and Arkes refer to as 'enhanced calibration feedback'. Such feedback involves having subjects (*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 the subjects understand the feedback information. In other words, while the relevant experiments show that 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. This counts against the sophisticated self-binding approach, and suggests that we need to look elsewhere for a more promising approach to the correction or prevention of bias.

4. Bias Prevention Through External Constraints

Let us take stock. We started out by noting that what is appealing about the idea of epistemic agency—be it understood in terms of reflection or not—is that it calls our attention to the possibility of epistemic improvement through the proper mobilization of one's agency. Then, we noted that one central and important respect in which we would want to do better is with respect to our tendencies for cognitive bias, and considered two approaches to the problem of bias framed in terms of things that the individual agent can do by way of mobilizing her epistemic agency. The first approach—that of *self-correction*—involved the agent trying to correct for bias that has already occurred. We found two problems with this approach: the problem of *motivation* and the problem of *proper* correction. On the second approach, agents do not try to correct for bias, but rather try to avoid bias in the first place. We framed this approach in terms of an epistemic version of Elster's notion of *self-binding*, understood in terms of self-imposed commitments that either remove the option of engaging in biasing activities, or make choosing such options significantly more costly than not choosing them. The problem for the self-binding approach was shown to be another *motivational* one.

In closing, a different approach will be outlined.[38] The approach involves imposing *external* constraints on the agent for the purpose of protecting her from bias by either shielding her from biasing information or reducing the risk that biasing information will affect her by mandating that she conducts her inquiry in a particular way. If the metaphor for self-binding was that of Ulysses having himself tied to the mast to avoid the tempting song of the sirens, the relevant metaphor for the external constraints approach is that of Ulysses putting wax in the ears of his sailors. By focusing on protecting people from bias rather than correcting bias when it has occurred, the approach avoids the problem of proper correction that afflicts the self-correction approach. Moreover, since the relevant constraints are imposed externally, and independently of whether the agents constrained are motivated or even willing to be constrained thus, the external constraints approach does not fall prey to the problems of motivation that affected both the self-correction and the self-binding approach.

4.1. External Constraints on Information Access

What, then, would the relevant kind of external constraints look like? A very straightforward kind of external constraint is a constraint on our *access* to information that is likely to bias us. For example, consider the practice on the part of U.S. judges to withhold certain kinds of information from the jurors, such as character evidence or evidence about past crimes, on the assumption that the jurors will systematically overestimate the probative value of such information.[39] Consequently, according to the U.S. *Federal Rules of Evidence*, the mere fact that a piece of evidence is relevant, in that it makes the relevant hypothesis about guilt more or less likely, is not a sufficient condition for presenting it to a jury. Relevant information can do more epistemic harm than good if jurors give it greater weight than it actually has. For this reason, legal practice and regulation takes into account not only relevance but also whether jurors are able to *gauge* that relevance properly. If not, the presiding judge may withhold the relevant information from the jury.

This practice on the part of U.S. judges has been invoked by Alvin Goldman as an instance of what he refers to as *communication control*.[40] Someone is involved in communication control *vis-à-vis* someone else, according to Goldman, when the former is exercising control over what evidence or, more generally, information is available to the latter. Another example of communication control discussed by Goldman is the common practice of teachers to withhold certain kinds of information from their students. As Goldman notes, students *not* being exposed to all possible viewpoints is probably good from an epistemic point of view, that is, from the point of view of having them form true beliefs and avoid forming false ones. For one thing, if we focus on 'palpably false or indefensible' theories and viewpoints, withholding

the relevant information can be expected to minimize the risk that the students will accept the theories in question, and thereby form false beliefs.

However, in analogy with the legal case above, we may even imagine that teachers could justifiably withhold true and perfectly accurate theories, on the grounds that those theories are nevertheless such that they would have a tendency to confuse the students, and have them draw the wrong conclusions. Consider classes on the health risks of illegal drugs, for example. Some drugs are more addictive than others, and some drugs have more detrimental health effects than others. Moreover, some legal drugs (for example, tobacco and alcohol) may have more detrimental effects than some illegal drugs (for example, cannabis). Consequently, a completely accurate account of the risks and benefits of drugs would have to be fairly complex, on account of making several distinctions and qualifications. As a consequence, such an account might also be more likely to lead students to draw inaccurate conclusions than a less sophisticated—and strictly speaking less accurate—account.

Whether such a less sophisticated account would, in fact, have students form more true beliefs and less false ones, compared to a more accurate account, is of course an empirical question. Settling that empirical question might have been necessary for determining whether a practice of favoring a less sophisticated account over a more sophisticated account in the relevant settings would be *justified*. However, for the purposes of whether some practice would qualify as a form of information control, it does not matter whether the practice in question would actually have the intended effects. What matters is whether it would involve controlling the evidence or information available to the students by selectively withholding some information—which it clearly would.

Moreover, the idea of communication control serves to illustrate one relevant kind of external constraint, namely *an external constraint on information access*. Constraints on information access serve to restrict the information that the agent has available to her, and thereby also the choices she can make when expressing her agency in relation to the question of what information to bring to bear on whatever matter she happens to be considering. The particular use of such restrictions that we are concerned with presently is that of restricting access to information that is likely to bias the recipient, for example, in the manner that current legal practice assumes that character evidence or information about past crimes might do with respect to jurors, or that educational practice assumes that certain information might do with respect to students.

For present purposes, we do not need to assume that the relevant legal and educational practices are successful on this score. It suffices to note that they are the *kind* of practices that would avoid the problems identified above for the self-correction and self-binding approaches. The kind of practices discussed in this section avoids these problems by imposing external constraints on the epistemic agency of the agent, rather than relying on the agent for

correcting or avoiding bias on her own accord. By invoking external constraints, the relevant practices do not require that the agents constrained are motivated to avoid bias. And by involving constraints specifically on biasing information, and thereby protecting people from bias rather than attempting to correct for bias after the fact, the practices in question do not face the challenges associated with proper correction.

4.2. External Constraints on Information Collection

A second kind of external constraint can be illustrated in terms of the practice of experimental randomization in the sciences, and in the medical sciences in particular. Albeit a fairly recent practice, randomization is today standard procedure in many of the sciences. Moreover, it is not hard to identify an *epistemic* rationale for why this is an excellent way to proceed, in light of common biases. After all, it reduces the risk that the agent will fail to spot confounding factors, and commit the common fallacy of taking a mere correlation to constitute a causal relation. Hence, the practice on the part of the U.S. Food and Drug Administration (FDA) to require the use of randomized experimental design in medical research pursuing causal hypotheses regarding the efficacy and safety of drugs, in so far as those drugs are to be marketed to the public.[41]

The word 'require' might be considered too strong here. It might be suggested that what we are dealing with here is not a requirement but rather a certain structure of *incentives*. However, barring legitimate ethical worries about the use of randomization in certain contexts, the relevant structure is such that medical research that is *not* performed in accordance with the randomization paradigm, and as such does not stand a good chance of satisfying FDA regulations, typically will not be deemed worthy of grant money, the researchers involved not be considered suitable for research positions, and so on. Given that the practice thereby is so intimately connected with the very livelihood of the researchers, it seems somewhat cynical to describe the practice in terms of incentives. As pointed out by Ian Hacking, 'the broad mass of routine empirical experiments take randomized design for granted and suppose that their employers would fire them if they did not'.[42]

In light of this, the practice of having medical researchers pursue causal hypotheses through randomization seems properly described as nothing short of a requirement. Moreover, much like restrictions on information access, such a requirement imposes a constraint on the choices an epistemic agent can make when it comes to what information to bring to bear on whatever matter she happens to be considering. However, unlike restrictions on information access, the kind of requirement under consideration here does this *not* by way of restrictions on what information the agent has available to her, but by constraining her choices regarding how to go about *collecting*

information. The particular requirement relevant to experimental randomization in the medical sciences does this by imposing restrictions specifically on choices regarding information collection as it pertains to the evaluation of causal hypotheses, by mandating that such information be collected through (and only through) experiments with a randomized design. As such, the mandate on randomized experimental design exemplifies a second kind of external constraint, namely *an external constraint on information collection*.

In analogy with what was noted in relation to the above examples of external constraints on information access, we do not need to assume for present purposes that the practice of requiring medical researchers to collect information by way of randomized controlled designs is one that is actually successful in protecting such researchers from bias. It suffices to note that such a practice constitutes a kind of practice that avoids the challenges identified for the practices relying on self-correcting or self-binding agents. As above, a practice involving external constraints on information collection avoids these problems by being in the business of protecting people from bias rather than correcting bias, and not assuming that the agents constrained are motivated to avoid bias.

5. Where Does This Leave Epistemic Agency?

We have considered three approaches to the question of how to come to terms with our biased ways. The first two were framed in terms of the kinds of things that the individual agent can do herself by mobilizing her epistemic agency, either for the purpose of correcting biased beliefs, or making commitments that will either remove or make more costly the option of engaging in activities known to lead to biased reasoning. If either constituted a plausible approach, we would have had reason to believe that there was an important role for our epistemic agency to play in relation to our desire for overcoming cognitive bias. However, both approaches faced significant challenges—challenges met by our third approach, framed in terms of external constraints imposed on the agent for the purpose of protecting her from bias by either shielding her from biasing information or reducing the risk that biasing information will affect her by mandating that she conducts her inquiry in a particular way.

Where does this leave epistemic agency? As noted at the outset of our investigation, what is appealing about epistemic agency is the fact that us being the kind of epistemic creatures that actively *do* things means that we also might be able to do *better*. However, what the above suggests is that, in so far as there is anything of philosophical substance to this intuitive appeal of epistemic agency, we find little evidence of it when we consider the role of epistemic agency in our attempts to improve epistemically in response to our well-known susceptibility to cognitive bias. If the above is on the right track,

it seems we simply cannot rely on ourselves for epistemic improvement. This, moreover, suggests that the problem with epistemic agency goes beyond that identified by Kornblith in relation to the idea that exercising one's epistemic agency is a matter of reflecting on one's first-order beliefs. The problem with epistemic agency is not limited to that arising in relation to epistemic agents reflecting thus; it extends to the substantive challenges associated with it being up to the agent to attempt to epistemically improve, be it by reflecting on her beliefs or otherwise.[43]

Notes

1. See Haddock, Millar, and Pritchard (2009) for a recent anthology.
2. See Ahlstrom-Vij (forthcoming *a*) and Ahlstrom-Vij (2013).
3. Moran (2001: 114).
4. Moran (2001: 144).
5. Moran (2001: 145).
6. Sosa (2004: 292).
7. See Kornblith (2012).
8. Kornblith (2012: 90).
9. In the psychological literature, biases are sometimes identified with instances of heuristical reasoning that flout principles gleaned from logic, statistics, and probability theory in a systematic manner. See, for example, Gilovich and Griffin (2002: 4). However, since some heuristics identified in the literature as biases may very well be adaptive (see, for example, Gigerenzer *et al*. 1999 and Cosmides and Tooby 1996), it is better to understand what it is to be a bias simply in terms of systematic tendencies to form inaccurate beliefs, be it on account of flouting formal principles, or due to differences between modern cognitive challenges and those that faced our evolutionary ancestors.
10. See Taylor and Brown (1988).
11. See, for example, Alicke (1985) and Brown (1986).
12. See Alicke *et al*. (1995) and Dunning *et al*. (1989).
13. See Armor (1999).
14. See Pronin *et al.* (2002).
15. Pronin (2007: 37).
16. See, for example, Fischhoff *et al*. (1977).
17. See Henrion and Fischhoff (1986).
18. For example, research shows that agents that take care to consider arguments for positions inconsistent with the one they themselves are sympathetic to, or to list reasons why some particular anchor value might be inappropriate, are less susceptible to biased assimilation of evidence (Lord, Lepper, and Preston 1984) and anchoring effects (Mussweiler, Strack, and Pfeiffer 2000).
19. See Pronin and Kugler (2007).
20. The terms 'unnecessary correction', 'insufficient correction', and 'overcorrection' are borrowed from Wilson *et al*. (2002).
21. See Wilson (2002) for an extended discussion of the introspective inaccessibility of large parts of our mental lives.

22. See Wilson and Brekke (1994).
23. Petty and Wegener (1993).
24. See, for example, Lombardi, Higgins, and Bargh (1987), Martin (1986), and Martin, Seta, and Crelia (1990).
25. Lerner and Tetlock (1999: 270).
26. See Camerer and Hogarth (1999) for an overview of relevant research.
27. Camerer and Hogarth (1999: 34).
28. Wilson *et al*. (2002: 192 and 195).
29. See Elster (2000).
30. See Meehl (1954).
31. See Sawyer (1966).
32. Dawes *et al*. (2002: 719).
33. Meehl (1954).
34. See Dawes *et al*. (2002).
35. See, for example, Arkes *et al*. (1986) and Lord *et al*. (1984).
36. See Arkes *et al*. (1987).
37. Sieck and Arkes (2005).
38. I develop this approach in greater detail in Ahlstrom-Vij (forthcoming *b*) as part of a general defence of epistemic paternalism.
39. See, e.g., *Michelson v. US*, 335 U. S., at 475–6, 1948, for one legal statement to this effect.
40. See Goldman (1991).
41. See Meldrum (2000) for a history of randomized controlled trials, and a discussion of the development of the FDA's policy on such trials in relation to their requirement for 'substantial evidence' about efficacy and safety.
42. Hacking (1988: 430).
43. This paper is a significantly reworked version of the first chapter of Ahlstrom-Vij (forthcoming *b*), and I am grateful to Palgrave Macmillan for their permission to reproduce some sections of that chapter here. Thanks also to Duncan Pritchard and Allan Hazlett for an invitation to present at the University of Edinburgh's *Epistemology Research Group*, where I received particularly helpful comments from Duncan and Allan, as well as from Nick Treanor, Robin McKenna, and Shane Ryan.

Bibliography

Ahlstrom-Vij, K. (2013) 'Meno and the Monist', *Metaphilosophy*, 44(1–2), 157–170.

Ahlstrom-Vij, K. (forthcoming *a*) 'In Defense of Veritistic Value Monism', *Pacific Philosophical Quarterly*.

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

Alicke, M. D. (1985) 'Global Self-Evaluation as Determined by the Desirability and Controllability of Trait Adjectives', *Journal of Personality and Social Psychology*, 49, 1621–1630.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., and Vredenburg, D. S. (1995) 'Personal Contact, Individuation, and the Better-Than-Average Effect', *Journal of Personality and Social Psychology*, 68, 804–825.

Arkes, H. R., Christensen, C., Lai, C., and Blumer, C. (1987) 'Two Methods for Reducing Overconfidence', *Organizational Behavior and Human Decision Processes*, 39, 133–144.

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*, 59, 5163.

Brown, J. D. (1986) 'Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments', *Social Cognition*, 4, 353–375.

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

Cosmides, L., and Tooby, J., (1996) 'Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment under Uncertainty', *Cognition* 58, 1–73.

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. 716–729). Cambridge: Cambridge University Press.

Dunning, D., Meyerowitz, J. A., and Holzberg, A. D. (1989) 'Ambiguity and Self-Evaluation: The Role of Idiosyncratic Trait Definitions in Self-Serving Assessments of Ability', *Journal of Personality and Social Psychology*, 57, 1082–1090.

Elster, J. (2000) *Ulysses Unbound*. Cambridge: Cambridge University Press.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977) 'Knowing with Certainty: The Appropriateness of Extreme Confidence', *Journal of Experimental Psychology: Human Perception & Performance*, 3 (4), 552–564.

Gigerenzer, G., Todd, P. M., and the ABC Research Group (eds.) (1999) *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press.

Gilovich, T., and Griffin, D. (2002) 'Introduction—Heuristics and Biases: Then and Now', in T. Gilovich, D. Griffin, and D. Kahneman (eds.) *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 1–18). Cambridge: Cambridge University Press.

Goldman, A. (1991) 'Epistemic Paternalism: Communication Control in Law and Society', *The Journal of Philosophy*, 88 (3), 113–131.

Hacking, I. (1988) 'Origins of Randomization in Experimental Design', *Isis*, 79 (3), 427–451.

Haddock, A., Millar, A., and Pritchard, D. (eds.) (2009) *Epistemic Value*. Oxford: Oxford University Press.

Henrion, M. and Fischoff, B. (1986) 'Assessing Uncertainty in Physical Constants', *American Journal of Physics*, 54, 791–798.

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

Lerner, J. S., and Tetlock, P. E. (1999) 'Accounting for the Effects of Accountability', *Psychological Bulletin*, 125 (2), 255–275.

Lombardi, W. J., Higgins, E. T, and Bargh, J. A. (1987) 'The Role of Consciousness in Priming Effects on Categorization: Assimilation versus Contrast as a Function of Awareness of the Priming Task', *Personality and Social Psychology Bulletin*, 13, 411–429.

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), 1231–1243.

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), 1231–1243.

Martin, L. L. (1986) 'Set/Reset: Use and Disuse of Concepts in Impression Formation', *Journal of Personality and Social Psychology*, 51, 493–504.

Martin, L. L., Seta, J. J., and Crelia, R. A. (1990) 'Assimilation and Contrast as a Function of People's Willingness and Ability to Expend Effort in Forming an Impression', *Journal of Personality and Social Psychology*, 59, 27–37.

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

Meldrum, M. L. (2000) 'A Brief History of The Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard', *Hematology/Oncology Clinics of North America*, 14 (4), 745–60.

Moran, R. (2001) *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, NJ: Princeton University Press.

Mussweiler, T., Strack, F., and Pfeiffer, T. (2000) 'Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility', *Personality and Social Psychology*, 26 (9), 1142–1150.

Petty, R., and Wegener, D. (1993) 'Flexible Correction Processes in Social Judgment: Correcting for Context-Induced Contrast', *Journal of Experimental Social Psychology*, 29, 137–165.

Pronin, E. (2007) 'Perception and Misperception of Bias in Human Judgment', *Trends in Cognitive Science*, 11 (1), 37–43.

Pronin, E., and Kugler, M. (2007) 'Valuing Thoughts, Ignoring Behavior: The Introspection Illusion as a Source of the Bias Blind Spot', *Journal of Experimental Social Psychology*, 43, 565–578.

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, 369–381.

Sawyer, J. (1966) 'Measurement and Prediction, Clinical and Statistical', *Psychological Bulletin*, 1, 54–87.

Sieck, W. R., and Arkes, H. R. (2005) 'The Recalcitrance of Overconfidence and its Contribution to Decision Aid Neglect', *Journal of Behavioral Decision Making*, 18, 29–53.

Sosa, E. (2004) 'Replies', in J. Greco (ed.) *Ernest Sosa and His Critics* (pp. 275–325). Malden, MA: Blackwell Publishing.

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

Wilson, T. D. (2002) *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.

Wilson, T. D., and Brekke, N. (1994) 'Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations', *Psychological Bulletin*, 116 (1), 117–142.

Wilson, T. D., Centerbar, D. B., and Brekke, N. (2002) 'Mental Contamination and the Debiasing Problem', in T. Gilovich, D. Griffin, and D. Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 185–200). Cambridge: Cambridge University Press.