The Epistemic Life of Groups: Essays in the Epistemology of Collectives
Brady & Fricker eds. (OUP 2016)

Chapter 2

Fault and No-fault Responsibility for Implicit Prejudice—A Space for Epistemic 'Agent-Regret'

[I]n the story of one's life there is an authority exercised by what one has done, and not merely by what one has intentionally done. Bernard Williams[1]

How far, or in what manner, should we hold each other responsible for the inadvertent operation of prejudice in our thinking? In recent work in philosophy and psychology on 'implicit bias' this vexed question has taken on a new form and urgency. The difficulty we face in thinking about the structure of our responsibilities in this regard is a canonical one for philosophy: the puzzlement created by conflicting intuitions. On the one hand we naturally feel that any kind of prejudice that works against another group is something epistemically (and ethically) blameworthy; yet on the other hand the whole point about 'implicit' biases, including those biases we would consider prejudices, is that their influence in our judgements is so hard to detect that correcting them would seem to require supererogatory, even superhuman, levels of perceptiveness, corrective know-how, or plain time and effort. We are rightly reluctant to proclaim that we are always blameworthy for this kind of epistemic error, for

[1] Bernard Williams, *Shame and Necessity* (Berkeley, LA, London: University of California Press) p. 69.

that would surely be an unfair demand in cases where we are structurally[2] in no position to tell that we are guilty of it (a form of non-culpable ignorance), and/or where it would take heroic efforts to reliably correct it (a mere failure to perform the supererogatory). Yet we certainly should not throw up our hands and declare that since, as things stand, we cannot reasonably be said to help it, nor can we reasonably be held accountable.[3]

Other things equal, epistemic fault is culpable. Given that prejudice is an epistemic fault (I shall specify which fault in a moment), someone who falls into prejudicial thinking is a *prima facie* candidate for epistemic blame. And yet when we consider cases of *implicit* prejudice (where the subject is radically unaware of her prejudicial habit), the idea that the epistemic conduct is blameworthy starts to look pointlessly over-demanding. Implicit prejudice looks to be, at least sometimes, beyond blame. Jennifer Saul has made this point in relation to those biases of which we are unaware and/or which we are unable to control (where these inabilities are, I take it, to be conceived structurally, not as mere personal failing but rather as a product of the society we live in).[4] She adds, further, the pragmatic consideration that, in any case, going in for the expression of blame may be counter-productive. Jules Holroyd, by contrast, has argued that we are blameworthy for such biases at least some of the time, and the implication is that

[2] By 'structural' ignorance I mean ignorance caused by historical-cultural circumstances, as opposed to personal failure; I develop this idea in 'The Relativism of Blame and Williams's Relativism of Distance', Fricker (2012).
[3] I shall use the notions of accountability and responsibility as equivalent.
[4] Jennifer Saul 'Implicit Bias, Stereotype Threat, and Women in Philosophy' in *Women in Philosophy: What Needs to Change?* eds. Katrina Hutchison and Fiona Jenkins (Oxford: Oxford University Press, 2013); p. 55.

acknowledging culpability is likely to be part of raising standards.[5] I take these different claims to give expression to equally intelligible but broadly conflicting intuitions. What, then, are we to make of our mode of responsibility in these matters? My answer to this theoretical question will ultimately deliver certain first order practical imperatives for individuals, but also, and by immediate implication, for collective institutional bodies under whose auspices the individuals concerned may operate.

Even taking the idea of the 'institutional' as broadly as possible to include all corporate bodies, it is of course not the only arena of prejudice and other bias, for these are also played out in purely personal interactions. However, the institutional does preside over most dimensions of our social activity, whether it be education or employment, including commercial activity, charitable activity, regulatory activity, legislation, law enforcement, legal process, political process, social work, medicine, religion, cultural administration… In all these spheres of social operation individuals judge, deliberate, and act in role as officers or affiliates of institutional bodies of various kinds—sometimes they perform their role as an individual (the administrator, the teacher), sometimes as members of additive or 'summative' groups (the voters, the patients), sometimes as members of 'plural subjects' in Margaret Gilbert's strong collective sense where there is a joint commitment to operate 'as one' (the board, the government, or indeed any

[5] Jules Holroyd, 'Responsibility for Implicit Bias', *Social Philosophy*, Vol. 43, No. 3, Fall 2012; 274-306.

jointly committed 'we'),[6] and sometimes in looser, more easily disbanded groups united *pro tem* by intentional interdependence (the volunteers, the spectators), as we find theorised by Michael Bratman, or Christian List and Philip Pettit.[7] While these views tend to be understood as competing accounts of collective agency construed as a unified phenomenon, I regard each as capturing a different strength of collectivity.[8] And in these various collective capacities they engage in behaviours and decisions that can have major consequences for others, whether it is spontaneously perceiving someone as 'leadership material', or as carrying a weapon; or perhaps deciding whether they have one's vote, or are suitable to adopt a child; or, again, judging whether they have the mettle to stand for public office. How individuals behave, judge and deliberate in their various institutional roles is most of what goes on socially; so it is no exaggeration, I think, to describe the institutional as the most all-pervasive and consequentially extended arena for implicit prejudice. It is in relation to that enlarged context, then, that I shall be considering our epistemic accountability as regards the inadvertent operation of prejudice in our cognitive conduct.

I shall focus specifically on *epistemic* culpability, because I take it to be prior to the question of *moral* culpability, inasmuch as generally the question whether someone is morally blameworthy for an act or omission crucially depends on the epistemic question of whether there was non-culpable ignorance in play. I shall

[6] For a fuller specification of institutional bodies considered as plural subjects, see 'Can There Be Institutional Virtues?', Fricker (2010). For the classic early statement of Gilbert's view see her (1989), or more recently, (2000).
[7] See, for instance, Bratman (1999); and List and Pettit (2011).
[8] For an independent account which orchestrates a complex range of different gradations of collectivity, see Raimo Tuomela (2013).

however present a picture of epistemic responsibility that exactly mirrors a certain conception of moral responsibility—a conception according to which the domain of bad things done for which we are morally blameworthy does not exhaust the domain of bad things done for which we are morally responsible. This conception of responsibility is principally due to Bernard Williams, whose earliest explicit presentation of it is in his seminal discussion of moral luck and 'agent-regret'.[9] Our recently increased awareness of the pervasive influence of implicit prejudice and other biases in our everyday judgements needs fitting into a suitable conception of epistemic responsibility. That is the task of this paper, and my contention will be that we need a conception that finds a space for *no-fault epistemic responsibility* for certain kinds of bad judgement. A space, in effect, for a first-personal reflexive attitude of *epistemic agent-regret*.

The possibility of this responsibility status vis-à-vis prejudiced thinking promises to resolve the conflicting intuitions I rehearsed from Saul and Holroyd. We must surely start with the presumption that, at least as regards explicit prejudice, we are epistemically culpable for allowing prejudice into our thinking (though of course the question whether it is actually productive to confront each other about it remains another matter).[10] However my point will be that as we

[9] Bernard Williams, 'Moral Luck' in *Moral Luck: Philosophical Papers 1973-1980* (Cambridge: Cambridge University Press, 1082). See also the discussion of Oedipus in *Shame and Necessity* (Berkeley & Los Angeles, CA: University of California Press, 1993). We find a similarly broad conception of responsibility in the work of Raimond Gaita (see, for instance, *A Common Humanity: Thinking About Love and Truth and Justice* (London: Routledge, 1998), p. xvii; also *After Romulus* pp. 86-87.

[10] It is reasonable to worry that sometimes interpersonal confrontation will only make things worse; but on the other hand there is evidence that sometimes it can help—see Alexander M. Czopp, Margo J. Monteith, and Aimee Y. Mark (2006)

move into 'implicit' territory, where we are likely to find cases where we are not blameworthy for the reason that epistemic bad luck has played an exculpatory hand, *still* we are properly held responsible in the manner of agent-regret. Such *no-fault responsibility* stands to serve as a useful, non-confrontational (because non fault-finding) mode of holding oneself and each other accountable, and thereby pushing for collective ameliorative steps to be taken, as I will try to explain.

The kind of 'implicit bias' that is implicit prejudice

Bias is a very general category, which can span many things, from epistemically helpful heuristics to prejudice against stigmatised groups. In order to make my case for the idea that epistemic agent-regret has application, I shall focus on the idea of implicit prejudice and the kind of epistemic fault at stake in implicitly prejudiced thinking. But in order to get the relevant notions in place, let me first borrow a working definition of implicit bias from Holroyd, whose formulation does not explicitly employ the notion of prejudice but in which the notion of an *automatically* applied *negative property or stereotypic trait* seems at least to incorporate the kinds of negative prejudicial thinking on which I shall be focussing. At the very least, we might think of implicit prejudice as a dominant sub-class of implicit bias as Holroyd defines it:

'Standing Up for a Change: Reducing Bias Through Interpersonal Confrontation', *Journal of Personality and Social Psychology*, Vol. 90, No. 5; 784-803. I thank Michael Brownstein for directing me to this research.

> An individual harbors an implicit bias against some stigmatized group (G), when she has automatic cognitive or affective associations between (her concept of) G and some negative property (P) or stereotypic trait (T), which are accessible and can be operative in influencing judgment and behaviour without the conscious awareness of the agent.[11]

This definition is designed to capture the kind of biases detected not only in Implicit Association Tests (IATs) but also in tests concerning real-world activities such as the assessment of CVs depending, variously, on whether there is a male or female name at the top, or, alternatively, a name racialised as black or white.[12] Whatever circumspection one may harbour about split-second differences of click time on a mouse in pairing up, say, the word 'aggressive' with a black or a white face bearing an aggressive expression, the data about CV assessment and similar experiments involving real-world activities are worryingly impressive.

These kinds of evaluation activities structure many professional worlds—at the minimum they help determine who gets to enter the line of work in the first place, and subsequently who gets to advance in it—so that these sorts of judgements have an enormous influence in shaping the profile of productive social activity over the long term. The implication is that individuals who do not

[11] Holroyd, 'Responsibility for Implicit Bias', 2012; p. 275.
[12] Key sources in the psychology literature are Moss-Racusin et al 2012, and Bertrand & Sendhil Mullaainathan 2004. Both these experiments, and many other aspects of the issue of self-regulation, are discussed in Gendler 2014. See also the reply by Nagel 2014. There is also helpful discussion in Holroyd 2012.

have prejudiced explicit attitudes may nonetheless normally have prejudiced implicit attitudes that have an undetectable deleterious influence on their judgements and deliberations. We always knew prejudice could be stealthy, but putting this new level of undetectability together with the sheer prevalence of the phenomenon produces a whole new perspective. We now appear naively alienated from our own judgements, so that ordinary cognitive self-discipline seems more elusive than ever.

This new image of ourselves, refracted through the lens of controlled experiment, seriously compromises our conception of ourselves as cognitively authentic, or even epistemically responsible. My explicit attitudes are on the whole mine (even when they are borrowed, it is I who have borrowed them); but my implicit attitudes, it seems, are not like that. Inasmuch as they flow from me in a temporally and counter-factually stable manner, I can hardly deny them as a robust facet of my epistemic character; and yet for many of them I would disown their content utterly. This newly alienated self-image is at least as much of a shock to the system as was, in the nineteenth-century, the Freudian picture of human beings as pulled about by radically unconscious desires and fears. At least the demons of the unconscious were definitively personal to the individual psyche, their quirky forms shaped by our most intimate relations. By contrast, our implicit prejudices against stigmatised groups are peculiarly impersonal in their aetiology, for they are on the whole unwittingly absorbed from outside the spheres of intimacy—the attitudinal fall-out from a semi-toxic social environment.

How does Holroyd's working definition of implicit bias effectively incorporate the notion of prejudicial thinking? Let me focus on the idea that someone exhibits implicit bias when she *automatically associates* some *negative property or stereotypic trait* with a *stigmatised group*. Strictly speaking, such an automatic association need not quite amount to a prejudice, but it will in fact do so wherever the association is the result of any motivated failure to properly gear one's attitudes to the evidence—and this will surely be true for most automatic associations of negative properties or traits with stigmatised groups.[13] Evidential shortcoming of this sort might take various forms. It might be a matter of the subject's being guilty of some motivated resistance to counter-evidence, where 'resistance' might be a refusal even to recognise the instance as counter-evidence ('just the exception that proves the rule'), or alternatively it might consist in a failure to follow through with the requisite adjustments elsewhere in one's belief system or deliberations. Or again, evidential shortcoming might equally manifest itself in someone's being motivated to generalise from an excessively small, or unrepresentative, or contextually inappropriate sample.[14]

Let us gloss these different forms of evidential failing by saying that an attitude is prejudiced insofar as it is the product of (some significant degree of) *motivated maladjustment to the evidence*. A paradigm example of prejudiced thinking would be someone who has the prejudicial attitude that members of social group X are

[13] Perhaps an exception might be dredged up for purposes of conceptual argument: a case where someone automatically associates a negative property with a stigmatized group *without* any motivation—sheer spontaneous habit of some allegedly non-motivated kind. Insofar as this is a real possibility, it inserts a wedge between my notion of prejudice and Holroyd's notion of implicit bias, but a very thin wedge.

[14] Ishani Maitra makes this point in Maitra 2010; see pp. 206-7.

inferior to his own social group, where a significant part of the explanation why he has this attitude is some, perhaps entirely non-conscious, desire for superiority, fear of inadequacy, or perhaps simply a baseline motive to fit in with in-group attitudes. Obviously prejudice will often manifest itself by way of stereotyping, as is explicit in Holroyd's definition of implicit bias. Transferring our definition of prejudice, we can say that a prejudicial stereotype is a stereotype that is the product of some motivated maladjustment to the evidence.

Stereotyping can itself take different forms. It might take the form of an implicit generalisation ('all/most/many Xs are F'); or, alternatively, it might take the form of a 'generic' ('Xs are F') where there need be no pretension to there being a statistically significant number of instances, but merely a bald association expressing the salience of some feature—for instance what Sara-Jane Leslie calls a 'striking property' generalisation where 'striking' indicates danger or risk of harm. (One of her examples is 'Mosquitoes carry the West Nile virus' where in fact less than 1% of mosquitoes carry the West Nile virus.)[15]

For many generics, such as the one just cited, we regard them as true. As Leslie convincingly explains, this is because it is reasonable to essentialise certain natural kinds, that is, to regard them as possessing an underlying shared nature, so that properties observed in just a very few instances may reasonably be presumed to flow from that underlying nature and so generalise to the kind as a whole. In such cases the generic habit is epistemically justified, because it tends

[15] Sara-Jane Leslie, 'The Original Sin of Cognition: Fear, Prejudice and Generalization', *The Journal of Philosophy* (forthcoming).

to lead to true belief. But we can see in what fundamental respect it becomes unjustified when it is applied to social groups, for such groups are thereby treated as if they had an underlying shared essence when in fact they don't ('Muslims are terrorists' is cited as a post-9/11 example of a false generic statement of this kind—understood specifically as an essentialising over-generalisation from a minute sample). This distinctively essentialising type of over-generalisation, fuelled by fear of harm of some kind, puts on display its own distinctive form of motivated maladjustment to the evidence, and so comfortably fits the way we are conceiving of prejudicial stereotyping in general.

Having explained how Holroyd's definition of implicit bias (as including an automatic, sometimes stereotypical, negative association) might typically embody the motivated maladjustment to evidence that constitutes prejudice, let me now turn to the question of blameworthiness. I have said that the epistemic fault for which the prejudiced thinker is culpable, other things equal, is the motivated maladjustment to the evidence. Epistemic faults operating unimpeded in the individual's epistemic system are canonical cases of blameworthy epistemic conduct (other such faults might be jumping to conclusions, carelessly overlooking counter-evidence, wishful thinking, dogmatism, sloppy calculation, and so on). If, for example, someone chairing an academic appointments process were systematically, albeit unwittingly, to assess the male candidates' writing samples more highly than those of the female candidates owing to the operation

of implicit prejudice in her patterns of judgement, then other things equal we would regard her as epistemically at fault, and so blameworthy.[16]

We might not blame her very much, of course, if she were doing her well-intentioned best under difficult circumstances—pressure of time, lack of institutional support for alternative methods. These are mitigating circumstances, or excuses, and they function to reduce the degree of appropriate blame, even to zero in some cases, if we accept the possibility of fully exculpatory excuses, which we surely may. But they do not change the kind of epistemic fault that has expressed itself, which is blameworthy other things equal. This is no less so if she herself would be horrified were this to be revealed to her after the fact (as it might be if, for instance, we imagine her as a participant in a controlled psychological experiment on the operation of implicit bias). Indeed the appropriateness of blame in a case like this is actively supported by the thought that not only might the well-intentioned agent blame herself, but moreover anyone whose work had received a prejudiced assessment would surely blame her too, not only where the prejudice lowered the estimation of the work (a case of testimonial injustice[17]) but surely also where it raised it.[18]

---

[16] I have argued elsewhere that the notion of someone's being blameworthy is best understood in terms of their being *at fault*, and the notion of blaming someone as a matter of *finding fault*, where paradigmatically (though not necessarily) that judgement will be interpersonally communicated to the wrongdoer with a view to inspiring remorse. See Fricker 2014.

[17] 'Testimonial injustice' happens when prejudice deflates the credibility attributed to someone's word (see Fricker 2007, chapter 1).

[18] I have argued elsewhere that to blame is to find fault regardless of whether one benefits from the fault, so that one can blame even if one does not desire that things had gone differently (Fricker 2014).

To bolster this idea that our judgements of culpability are partly organised around whether we judge the fault to be traced to, or located in, the subject's epistemic system as opposed to being traced to someone else's (or indeed to the collective at large), let us consider a contrast case. Imagine a situation in which someone justifiedly believes the word of a speaker who confidently tells them that *p*, but who has been culpably careless with the evidence. Our hearer ends up with a false belief, but the story of epistemic fault is such that the buck is passed, and we regard the error as exclusively the fault of the original speaker—she, after all, was the one who had been careless with the evidence, whereas our hearer made no error of reasoning in believing her. The fault is not located in the hearer at all, but rather traced to the speaker. The question 'Whose fault is it?'—heard as a question about the fault's explanatorily salient location—is a very powerful organising idea in how we make judgements of blameworthiness.

Returning to our present day implicitly prejudiced assessor of writing samples, we will regard her as epistemically culpable (mitigating circumstances notwithstanding) insofar as we regard *her* epistemic system or character as the explanatorily salient source of the fault. This is so even if we imagine her to be completely unaware of having these implicit prejudices, rather as an implicitly selfish person's selfishness might systematically lead her to act selfishly even while she remains entirely unaware of this fact and (as we often say) 'cannot help it'. This idea of an implicit trait of moral character, especially a vice, makes an instructive comparison—the idea of selfishness often being non-conscious, and inaccessible through introspection yet unwittingly manifested in judgement and action, is hardly an alien one. And nor is it (here is the point of the

comparison) an alien idea to consider it nonetheless blameworthy—indeed one might take implicit selfishness as a prime case of the morally blameworthy.[19] The more general point to be extracted here is that certain kinds of moral theory tend to be forgetful of the fact that we normally—canonically, even—hold each other as blameworthy for behaviour that is expressive of bad traits or motives that are understood to be beyond our ken and control. People blame us for any faulty traits or motives they consider *ours*. Furthermore, the point is not restricted to that which is characteristic of the agent, for uncharacteristic motives and acts can still be ours in the relevant sense—features of *our epistemic system*. So whatever reason we might establish for regarding some operations of implicit prejudice as non-culpable will clearly need to invoke grounds other than that we did not know about them, or could not control them, or that they were not really expressive of our epistemic character. All these things can be true of a moment of bias, just as they may be true of a moment of selfishness, and yet I remain a perfectly proper object of blame, excuses notwithstanding, for the simple reason that the fault was mine.

It will be worth an aside at this point, I think, to emphasise the extent to which this observation about a perfectly normal and proper mode of blame is at odds with a philosophical dogma about responsibility. In a range of philosophical debates the idea has become deeply entrenched that we are only genuinely

[19] See Robert Merrihew Adams, 'Involuntary Sins', *Philosophical Review* Vol. 94 No. 1 (1985). Of the self-righteous person, for instance, Adams says 'According to the doctrine that all sin must be voluntary, it would seem that he is not to blame. And yet I think he clearly is to blame, not because of his voluntary choices but because of his self-righteous attitude. We will have a lop-sided view of the guilt in human relationships if we do not recognize this.' (p. 6).

responsible for that which is under our control. Now of course this idea does have application—*some* sorts of loss of control (such as real internal or external compulsion) clearly switch off responsibility altogether—but the idea tends to spread beyond its proper remit to figure as a quite general condition on responsibility, presenting itself as what George Sher has called 'the searchlight view'. According to the searchlight view, responsibility is limited to acts or omissions that one has chosen, or whose causal origins one has chosen.[20] The searchlight view certainly cannot be right, not least because it would rule out cases of culpable neglect and other kinds of culpable ignorance. (It would also rule out bad moral luck of course, and Sher considers a number of cases of that kind, though his discussion is not focussed on the concept of moral luck *per se*.) The view he favours, and with which I agree in broad strokes, is one according to which we are responsible for a good deal more than the searchlight view would allow, for we are responsible not only for conduct based on things we know but also for conduct based on things we *should have known* but didn't.[21] To this I would add, in the dimension of control, that we are not only responsible for those things we can control, but also for those things we ought to be able to control but can't (one's profound selfishness for instance). Obviously any

[20] See Sher (2009) *Who Knew? Responsibility Without Awareness* (Oxford: Oxford University Press).

[21] Sher, however, believes that something *additional* is needed, namely the condition that the agent's ignorance is 'explained by certain aspects of their constitutive psychology' (p. 93). Sher considers this necessary to guarantee that an agent's ignorance really is *theirs* where this is to involve the rather strong requirement that the ignorance is caused by something in their constitutive identity. But I do not see the need for this. If we are already talking, as all sides must agree we are, about an act or omission *on the part of the agent*, and of ignorance (culpable or non-culpable) *on the part of the agent*, then that is already enough.

judgements about what someone should have known must be made in relation to some reasonable standard, as Sher quite rightly says. And again I would add that any judgements about what someone should have been able to control must equally be made in relation to some reasonable standard. Indeed part of what all this suggests is that there is no escape, in judgements of culpability, from the demands of substantive normative thinking about what is a reasonable demand under the circumstances—there is no available recourse to a criterion of knowledge, or of control, that can guide our normative judgements from the outside.

This aside made, the foregoing reflections on the importance of the explanatorily salient location of the epistemic fault (in me? in her? in society as a whole?) raises the question whether there might be cases of implicit bias where the epistemic fault is *not* located in the prejudiced subject, but merely flows through her. Might there be such cases? If so, perhaps these are candidates for prejudiced thinking that is not epistemically blameworthy. In order to explore this possibility, let us now imagine a different writing sample assessor. This time she is not engaged in any motivated maladjustment to the evidence *herself*, but merely passively and *innocently* (i.e. without epistemic fault—this is important) inherits the gender biases in play from her environment. Perhaps we can stipulate that she simply has no relevant motivation for the bias; but in addition we must imagine a situation (no doubt far from that of the author or reader of this paper) in which she has no reason to suspect that she may be a conduit for gender bias (otherwise naivety on this score would already constitute epistemic

fault). In sum, we are imagining a case in which what is going on is the epistemically innocent inheritance of bad epistemic goods from the environment.

Such an imagined case looks like one of epistemically innocent error: our assessor is a faultless (and so blameless) conduit for prejudice. Her judgements are epistemically bad, but it is not her fault. The fault—the motivated maladjustment to the evidence—has already been committed off-stage by others, the epistemic collective of which she, through no fault of her own, is a member, and whose toxic influence (we have stipulated) she has no reason to suspect. We can ratchet up our subject's epistemic innocence all the more if we imagine her as justifiedly believing herself to be free of gender prejudice—perhaps she is an epistemically conscientious feminist actively trying her best not to let the gender of the writers affect her assessment of the work in any way. If so, this justified (if false) belief about her own lack of bias would constitute *counter-evidence* to the possibility (barely countenanced) that her judgements might involve gender bias ('No risk of that with me—I'm a feminist').

What should we make of such an imagined scenario in terms of our question about epistemic culpability for prejudiced thinking? Given how we have constructed the case, this particular assessor of writing samples is making a faulty judgement, yet the fault is not in her but rather in the collective. She is of course herself a member of said prejudiced collective, but mere membership is not enough to implicate her as individually blameworthy, especially given our stipulation that she has no reason to suspect she risks inheriting collective gender prejudice from it. Now, if we assume that epistemic responsibility

mirrors the common, narrow view of moral responsibility, to the effect that the only mode of epistemic responsibility for faulty judgement is epistemic blameworthiness, then we are stuck. And our innocent conduit of prejudiced judgements is simply off the hook. We find no fault in her, and the narrow view affords no residual space of responsibility once the possibility of culpability is eliminated. We shrug our shoulders and leave things as they are, for on this picture no epistemic obligations accrue to her in virtue of her serving as the innocent conduit of bad epistemic materials.

But that is a thoroughly unattractive, because conservative, conclusion to draw; and it does not explain the shame she would rightly feel if her prejudicial bias were revealed to her, or indeed if any of those who she had disadvantaged were to confront her. Surely there is something more subtle we can say about her moral status than that she made bad judgements through no fault of her own and so cannot, should not, be held accountable? There is. We can develop an analogy from an alternative conception of moral responsibility, which pictures the domain of moral responsibility for bad conduct as extending beyond the domain of the blameworthy. For this conception we can look to Bernard Williams' idea of 'agent-regret'.[22]

Agent-regret: From ethical to epistemic

[22] See Williams 1982.

Williams famously argued that the traditional picture of moral responsibility as co-extensive with potential blame delivered a falsely purified version of our normal, nuanced forms of everyday moral consciousness. Furthermore, this purified picture was the *imprimatur* of 'the morality system'—an absolutist moral outlook which remains as familiar as it is peculiar on account of the fact that it is ('incoherently', as he remarks) part of the moral consciousness of all of us. Were it not for its purifying influence, the natural place of various kinds of luck (in particular, bad luck) within moral life would be readily acknowledged. (Williams registered the peculiarity of the purified conception by reserving the word 'moral' for its special constructive purposes, and instead using the word 'ethical' for a conception of moral life from which the possibility of luck affecting one's moral status had not been theoretically expunged in advance. But I prefer not to hand over a word as resonant as *morality* to the advocates of any conception deemed false, and so will not imitate his linguistic innovation here.)

The specific fruit of Williams' approach to the place of moral luck in our lives, and in particular its implications for how we make sense of cases where people do bad things, perhaps terrible things, through no fault of their own, is something I contend not only moral philosophy but also epistemology can learn from. He argued that when we do bad things through no fault of our own (when we cause bad things through our agency sufficiently proximally to count as *having done it*, yet blamelessly) the natural and proper response is to morally *own* these aspects of our conduct. Williams did not put it in terms of 'owning', but I think it is a highly pertinent psychological trope, and not one exclusive to psychotherapeutic concerns. The schoolish notion of 'owning up' to having done

something bad is simply one of responding to the question 'Who did this?' with the admission '*I* did'. Significantly, such an admission does not entail culpability ('*I* did; but it wasn't my fault…').

This regretful owning of harm done expresses itself in a distinctively agential first-personal reflexive attitude, quite different in kind from that of a bystander, who after all has nothing to own. And it was this distinctively agential regretful expression of responsibility that Williams labelled 'agent-regret'. That agent-regret is a genuinely *moral* response—that is, an expression of moral responsibility and not simply an expression of shock or compassion of the clearly non-moral regret that might properly be felt by the bystander—is revealed by the fact that, typically, there will be moral reasons that apply to the agent which do not apply to anyone else. Reasons pertaining indeed to the owning of harm done.

In the example that Williams uses to introduce the phenomenon, the agent is a lorry driver who through no fault of his own tragically runs over a child who has stepped into the road. The response on the part of the imagined lorry driver is one of horror at what has happened while he was at the wheel, and of a (perhaps impossible) wish to make amends. Significantly, in less tragic cases, the agent *may* often be able to make amends, perhaps by offering some kind of compensation; but in a case like the imagined one, perhaps the most our driver can achieve is some kind of symbolic expression of deep sorrow at his part in the tragedy. Agent-regret, in my view, is properly considered a feeling of guilt for harm done, though clearly not of a kind entailing culpability (which is why many

would recoil from conceiving agent-regret as a form of guilt). Since the morality system has no word, indeed no conceptual space, for any kind of guilt-feeling for non-culpable harm done, where that is considered as a properly *moral* response as opposed to an understandably intensified form of bystander-regret, Williams coins a new term for us.[23]

The lorry driver case is one of 'outcome' luck: through sheer bad luck the lorry driver's entirely responsible driving turns out to cause, on this occasion, a tragic effect. Our epistemological concern with responsibilities relating to prejudiced thinking, however, does not relate to unlucky outcomes so much as unlucky inputs, and so it will be useful to present a slightly different sort of moral case in relation to which we may draw our envisaged parallel. Our epistemological concern is with a case in which an epistemic subject has blamelessly inherited bad epistemic goods from her environment (the implicit prejudice) so that the epistemic fault of motivated maladjustment to the evidence has already been committed off-stage by the collective (and we have stipulated that the subject herself has neither been culpably negligent in failing to realise this, nor does her membership of said collective sufficiently implicate her as an individual in this patch of its bad epistemic conduct). This represents what we might think of as a kind of *environmental epistemic bad luck*; one which, furthermore, *obscures from*

[23] Elianna Fetterolf has argued persuasively that a careful reading of Williams reveals an intriguing possibility implicit in his view, namely that, were our ethical thinking to free itself from the undue constraints of the morality system, there would be no continued need for the special term 'agent-regret', because the ordinary notion of remorse might at last be seen to apply to such unlucky cases, instead of remaining confined to cases of fault (Fetterolf, 2014).

her view the epistemic significance of the patterns of judgement in which she is engaged.

Let me therefore shift to a different example of moral bad luck—one that provides a better fit. Let us consider Oedipus—another case discussed by Williams, this time in *Shame and Necessity*.[24] The horrifying things Oedipus has non-culpably done are that he has killed his father and married his mother. As Williams puts it 'The terrible thing that happened to him, through no fault of his own, was that he did those things'.[25] What Oedipus was not in a position to grasp was the moral significance of killing this man or marrying this woman (in this case for the simple reason that he was not in a position to know who they were). Circumstances conspired to ensure that he could not have been expected to know these things, and so this non-culpable factual ignorance entailed non-culpable moral-epistemic ignorance of the significance of his deeds. It was without fault, then, that Oedipus committed these crimes; and yet upon their discovery he dashed out his own eyes for shame (an act which in itself symbolises the reflexive capacity of shame—the internalisation of the shaming gaze of others—a capacity for which it is Williams' chief purpose in *Shame and Necessity* to orchestrate an extended argument). The morality system can only make sense of Oedipus's moral shame as a quasi-pathological response—an expression of understandable distress, indeed trauma, but not an expression of moral responsibility itself. The morality system aside, however, Oedipus is appropriately regarded as the unfortunate subject of a cruel environmental

[24] See Williams (1993) p. 58-74.
[25] Williams (1993) p. 70.

moral bad luck, inasmuch as features of his environment obscure from his view the grave moral significance of his deeds.

How is Oedipus's moral misfortune a better moral parallel for our imagined writing sample assessor, pictured as an epistemically innocent conduit for prejudice? First, whereas the lorry driver's running over the child was a radically non-voluntary action, the deeds of Oedipus were voluntary (at least under a plain description); and so are the assessments made (at least under a similarly plain description) by our re-imagined faultless assessor of writing samples: Oedipus fought and killed a particular man, and married a particular woman; our blameless writing sample assessor read the samples and judged that these were superior to those. Both do voluntary things the significance of which, through no fault of their own, they do not grasp; and in both cases their failure to grasp it is down to their circumstances or environment. They both suffer a kind of *environmental* bad luck. For Oedipus the primary bad luck is moral, whereas for our writing sample assessor the bad luck is in the first instance epistemic (the damage to her judgement is prior to any ethical harm caused). For both, agent-regret is in order; and perhaps for both, too, some degree of shame, though that is neither here nor there as regards the present argument.

My purpose in this paper is to make available the epistemic counterpart of agent-regret, in order to illuminate the moral status of epistemic subjects in the position of our writing sample assessor—a blameless conduit of prejudice who should nonetheless be represented as accountable. What I hope the foregoing discussion reveals is that she would appropriately experience her responsibility

for what she has done in the mode of *epistemic agent-regret*. This, in turn, indicates how others might hold her to account, for although Williams himself did not develop third-personal implications of the existence of agent-regret, still the implications are inherent in the first-personal form.

Carving out the space for this emotion allows us to do two things. First, it enables us to represent her as *non-culpable and yet responsible* for her prejudiced thinking. Second, it also allows us to see specific *epistemic obligations* as applying to her, through responding to which she may better 'own' what she has done and put some reasonable effort into putting things right. (I emphasise again, even if this is not the situation of either the author or any reader of this paper, still it is surely the situation of many people inasmuch as many are still not situated so as to be exposed to information concerning the strange implicitness of implicit bias, and its heavily disguised influence on our cognitive behaviour.) The first point—that she continues to be responsible even while she is non-culpable—we have already covered. But what of the second point, that certain epistemic obligations accrue to the epistemically unlucky subject? What might these be?

The direct counterpart to the moral case will be those obligations relating to making amends—fixing things caused by one's faulty judgement. Some of these will themselves be ethical obligations, or at once ethical and epistemic obligations, such as withdrawing the initial prejudiced assessment and getting the work re-assessed by independent parties, no doubt under anonymised conditions. But these are only the most immediate obligations that accrue to the faultless conduit of prejudice. The obligation to make amends will naturally

extend to more general aspects of the case, so that we might imagine our assessor, insofar as she is epistemically virtuous, taking steps to raise greater awareness of implicit bias, at least within her workplace, with a view to changing certain procedures so that it won't happen again in the same way. In the case of our imagined assessor, a procedural change such as anonymisation might have done the trick[26], and so the idea of an obligation to push institutionally for greater anonymisation is a good candidate for an epistemic obligation rising out of her responsibility status of epistemic agent-regret. The specific epistemic obligations incurred will no doubt often be mirrored by equivalent ethical ones, and will vary from context to context, but they are naturally thought of as essentially compensatory in nature—efforts to make amends and improve the situation in some appropriate measure. As Susan Wolf has argued, there is no algorithm for this kind of thing, and furthermore the reasons or obligations will tend to be substantially indeterminate; but some kind of stepping up or 'taking

[26] I should note, however, that even in a relatively straightforward case such as this, there can be unforeseen consequences. Imagine an example where the anonymisation conceals information about certain things, such as age, whose general correlation with how long a candidate has been in the profession might make it relevant to the assessment of the quality of the work (if the writing is a little immature or over-ambitious, then perhaps that is not such a bad sign in someone who is just starting out, but a clearly negative sign in someone who has been writing for several years). Here the seemingly neutral measure of anonymisation would influence the assessment of the work in a way that was *less* sensitive to relevant factors, and it might moreover work against the minority or group to which one is especially concerned to ensure fairness—for instance, if members of that group were more likely to be early career candidates. Of course ideally we would be able to learn from such a case that the appropriate form of anonymisation in this context is one where names and institutions are removed, but not age, or not 'years since PhD' or whatever. But this learning process takes time, and in any case it is extremely hard to detect when assessments have indeed been skewed. I thank Jo Wolff for drawing my attention to this sort of possibility.

responsibility' for what one has blamelessly caused calls on a virtue without a name (Wolf 2013).[27]

The sorts of ameliorative epistemic obligations incurred by the blameless conduit of prejudice, then, will concern the taking of steps to minimise the influence of prejudice in similar processes—perhaps institutional measures such as the removal of names at the top of writing samples for long listing purposes in appointment processes. But now we encounter an important feature of the case: the individual's power to ameliorate the situation is very limited. The kinds of procedural change we are imagining might help make amends and ensure it doesn't happen again are fundamentally *institutional* changes. These are changes that individuals can push for in the institutional context (the administrator, the teacher), but if they are to be implemented then a more collective effort is called for, whether on the part of a loose association of colleagues, and/or on the part

[27] David Enoch has further developed the theme of moral luck cases involving an obligation to 'take responsibility', which he regards as expressing the grain of truth in an otherwise 'seriously flawed' discussion on Williams' part. In short Enoch rejects the idea of moral luck, and tries to explain it away with the idea that whenever one is in a situation of purported moral luck, what's really going on is that one has incurred a moral obligation to take responsibility after the fact, so that if you fail to honour that obligation, you are at fault. Thus we find ourselves back in the familiar confines of morality narrowly construed—a conception according to which moral responsibility can express itself only in success or failure vis-à-vis the demands of moral obligation. Enoch rejects the idea that the unlucky agent is *already* responsible—tragically responsible, for instance, for the death of a child who ran into the road while one was at the wheel. While I agree that failure to respond to incurred obligations to make amends would indeed typically be a matter of fault, insofar as Enoch's purpose is to *replace* Williams' claims about agent-regret with claims about 'taking responsibility' I regard it not as rescuing any isolated insight from Williams, but rather as ultimately contorting the central insight back into the straightened form Williams was aiming to leave behind. I thank David Enoch for helpful discussion of these points over which we agreed to disagree when I presented this material at the Carlsberg Institute in Copenhagen.

of a committee or board which might be more tightly knitted together by joint commitment and so operate as a plural subject.

The fact that individuals are unlikely to be able to do very much acting alone exposes the fact that the fulfilment of individual obligations in the institutional context will tend to require agitations that inspire more collective responsiveness to the ameliorative obligation. Such agitations will often be sufficient to generate new obligations on the part of relevant collective bodies in the institution, largely because raised awareness of failure tends to generate an obligation to improve where possible. And improvements *will* often be possible. For instance, our biased but blameless writing sample assessor might press for anonymisation of writing samples. If, in doing so, she makes the appointing committee as a whole aware of the problem, with or without a concrete proposal to improve things, then this will generally be enough to generate a collective obligation on the part of the committee. Given the individual is already operating in some institutional role (she assesses the writing samples as part of her job, after all), it may be that the obligations she incurs in that capacity *already* involve the institution under whose auspices she is operating. But the point here is that even when that is not the case, her raising awareness in the relevant collective body will tend to be sufficient to generate a new collective obligation to help improve the situation, whether through retrospective compensatory measures or forward-looking ameliorative steps, or both. Such collectives are thereby obliged to take responsibility after the fact.

This is significant because the lesson we should take from our example is that, when it comes to implicit prejudice, the individual is rarely the only epistemic agent that acquires the epistemic obligations arising from an individual's epistemic agent-regret. There will tend to be immediate collective normative repercussions. Ultimately, then, the take-home message here is not only that there is a space for epistemic agent-regret, but that the epistemic obligations generated by that zone of no-fault responsibility are not confined to the individual who suffers the agent-regret, but extend swiftly to relevant collective bodies too.

Let me finish by looking a little more closely at what specific form the ameliorative obligations might take: what sorts of things would in fact compensate, mitigate, or thoroughly pre-empt the operation of implicit prejudice in our practices such as assessment and selection in a competitive process? A significant part of what is needed is for institutional bodies to pursue formal procedural techniques that minimise situations in which bias can enter in to influence proceedings (e.g. through greater anonymisation). Some, however, may make an additional case for the institutional provision of de-biasing therapies of the kind being researched in psychology. It is already standard in universities, for instance, for staff to be required to attend some kind of equal opportunities awareness training prior to serving on an appointments panel. In future one imagines it might be part of such a requirement that one take certain appropriate de-biasing therapies. In recent papers Jules Holroyd (2012), Tamar

Gendler (2014) and also Alex Madva[28] have emphasised the opportunities for de-biasing oneself that these kinds of therapies seem to offer, and rightly so. But there is surely a serious proviso here: we have not yet reached a stage of stable confidence in any given set of de-biasing therapies, and perhaps may never reach such a stage. What would be needed after all is a set of well-established, cross-contextually effective and practicable de-biasing techniques, whose results are widely known to be reliable in the differing contexts into which they are introduced and across different social types. We are nowhere near that situation yet, and to urge the adoption of such therapies too soon would carry a significant risk of unintended consequences as regards bias, and also of causing resentment and backlash if they do not ultimately inspire trust. While a certain imaginative daring is needed as regards increasing purely procedural mechanisms that eradicate the possibility of bias; caution must surely remain the watchword as regards introducing de-biasing therapies in the workplace, even on an optional basis.[29]

With this cautionary note now struck, there is every reason to keep seeking solutions of all sorts. When we allow prejudice into our judgements and deliberations we are always epistemically responsible, if not culpably then non-culpably in the mode of epistemic agent-regret. The burden of this paper has been to explain how, even when individuals are non-blameworthy for implicit prejudice, they are (a) appropriately held responsible; (b) typically acquire

[28] See Alex Madva, 'Biased Against De-biasing: On the Role of (Institutionally Sponsored) Self-transformation in the Struggle Against Prejudice'. Unpublished manuscript.
[29] See the general note of caution struck by Jennifer Nagel regarding the (in)accuracy of certain de-biasing strategies (Nagel 2014, section III).

certain obligations of an ameliorative sort; and also (c) these obligations quickly extend to any collective body or bodies in the institutional structure under whose auspices they are acting, because raising awareness of a dysfunction in the occupational business of an institutional body is generally sufficient to create collective obligations for that institutional body to take ameliorative steps. This is how a defect of individual epistemic conduct may generate not only individual obligations but also collective obligations to instigate change, for it is essentially through these larger agents that we may work more effectively towards an improved epistemic environment.[30]

References

Bertrand, Marianne, and Sendhil Mullainathan. (2004). 'Are Emily and Greg More employable than Lakisha and Jamal? A Field Experiement on Labor Market Discrimination', *American Economic Review*, 94(4), 991-1013.

Bratman, Michael. (1999) *Faces of Intention: Selected Essays on Intention and Agency* (Cambridge: Cambridge University Press).

[30] I have given versions of this paper in a number of places: Carlsberg Institute, Copenhagen; University of Cambridge; University of Stirling; Dartmouth College, New Hampshire; University of Bristol; University of Manchester; University of Cape Town; and Abö Akademi, Finland—I am grateful to all those present on these occasions for helpful discussion.

Enoch, David. (2012). 'Being Responsible, Taking Responsibility, and Penumbral Agency' in *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams* eds. Ulrike Heuer and Gerald Lang (Oxford: Oxford University Press).

Fetterolf, Elianna. (2014). *Remorse: A Prospective Genealogy*. PhD University of Sheffield. Unpublished manuscript.

Fricker, Miranda. (2007). *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press).

Fricker, Miranda. (2010). 'Can There Be Institutional Virtues?', *Oxford Studies in Epistemology (Special Theme: Social Epistemology)* Vol. 3 (2010) eds. T. S. Gendler & J. Hawthorne; 235-252.

Fricker, Miranda. (2012).'The Relativism of Blame and Williams's Relativism of Distance', *Proceedings of the Aristotelian Society Supp. Vol.* LXXXIV (2010), 151-77.

Fricker, Miranda. (2014). 'What's the Point of Blame? A Paradigm Based Explanation', *Noûs*. Early view.

Gendler, Tamar Szabó. (2014). 'The Third Horse: On Unendorsed Association and Human Behaviour', *Proceedings of the Aristotelian Society Supplementary Volume* XXXVIII, 185-218.

Gilbert, Margaret. (1989). *On Social Facts* (Princeton, New Jersey: Princeton University Press).

Gilbert, Margaret. (2000). *Sociality and Responsibility: New essays in plural subject theory*. (Lanham, MD: Rowman and Littlefield).

Holroyd, Jules (2012). 'Responsibility for Implicit Bias'. *Social Philosophy*, 43(3) Fall, 274-306.

Leslie, Sara-Jane. (Forthcoming). 'The Original Sin of Cognition: Fear, Prejudice and Generalization'. *The Journal of Philosophy*.

List, Christian and Pettit, Philip. (2011). *Group Agency: The possibility, design, and status of corporate agents*. (Oxford: Oxford University Press).

Madva, Alex. 'Biased Against De-biasing: On the Role of (Institutionally Sponsored) Self-transformation in the Struggle Against Prejudice'. Unpublished manuscript.

Maitra, Ishani. (2010) 'The Nature of Epistemic Injustice', *Philosophical Books* Vol. 51 No. 4 pp. 196-211

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. (2012). 'Science Faculty's Subtle Gender Biases Favor Male

Students', *Proceedings of the National Academy of Sciences of the United States of America* 109(41), 16474-9.

Nagel, Jennifer (2014). 'Intuition, Reflection, and the Command of Knowledge', *Proceedings of the Aristotelian Society Supplementary Volume* XXXVIII, 219-241.

Saul, Jennifer (2013). 'Implicit Bias, Stereotype Threat, and Women in Philosophy'. In Katrina Hutchison and Fiona Jenkins eds. *Women in Philosophy: What Needs to Change?* (Oxford: Oxford University Press).

Sher, George (2009). *Who Knew? Responsibility Without Awareness* (Oxford: Oxford University Press).

Tuomela, Raimo (2013). *Social Ontology: Collective Intentionality and Group agents* (Oxford: Oxford University Press).

Williams, Bernard (1982). 'Moral Luck'. In *Moral Luck: Philosophical Papers 1973-1980* (Cambridge: Cambridge University Press).

Williams, Bernard (1993). *Shame and Necessity* (Berkeley, LA, London: University of California Press).

Wolf, Susan (2013). 'The Moral of Moral Luck'. *Philosophic Exchange*. Vol. 31. Issue 1.