Jules.Holroyd@nottingham.ac.uk

Responsibility for Implicit Bias

Philosophers who have written about implicit bias have claimed or implied that individuals are not responsible, and therefore not blameworthy, for their implicit biases, and that this is a function of the nature of implicit bias as *implicit*: below the radar of conscious reflection, out of the control of the deliberating agent, and not rationally revisable in the way many of our reflective beliefs are. [1]

On this way of thinking about bias and responsibility individuals may be responsible for responding to information that they are, most likely, beset by implicit biases; they may be blameworthy if they fail to put in place strategies for preventing their biases from having an effect on decisions (such as anonymising CVs or essays). In other words, individuals may be blameworthy for failing to take responsibility for implicit biases once they are aware that they are likely to be influenced by them; but otherwise, individuals are not blameworthy for *being* biased, or for being *influenced* by implicit bias.

I argue that close attention to the findings of empirical psychology, and to the conditions for blameworthiness, does not support these claims. I suggest that the arguments for the claim that individuals are not liable for blame are invalid, and that there is some reason to suppose that individuals are, at least sometimes, liable to blame for the extent to which they are influenced in behaviour and judgment by implicit biases. I also argue against the claim that it is counter-productive to see bias as something for which individuals are blameworthy; rather, understanding implicit bias as something for which we are (sometimes) liable to blame could be constructive.

In section 1, I set up what is meant by implicit bias, and the concerns that some philosophers have expressed about treating persons as liable to blame for their implicit biases. In 2, I consider these arguments in detail, giving consideration to empirical studies and to the necessary conditions for responsibility posited. In doing so, we gain a clearer view on some of the kinds of factors that can influence the extent to which bias is manifested. Having concluded that we should reject the arguments canvassed for the conclusion that we are not liable to blame for biases, in section 3 I then go on to consider, and respond to, the concern that blaming individuals for implicit biases can be counter-productive.

1. Implicit bias: pervasive and blameless?

What is meant by 'implicit bias'? Let us start with a working definition that draws on the findings of empirical psychologists' studies over the last two decades.

An individual harbours an implicit bias against some stigmatised 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 judgement and behaviour without the conscious awareness of the agent.[2]

1 Brief discussion of responsibility, accountability and exculpation can be found in Machery, Faucher & Kelly, 'On the Alleged Inadequacy of Psychological Explanations of Racism', *The Monist*, 93(2), pp.228-255 (see esp. pp.246-249). See also Susanne Pohlman, (ms.) 'Accountability and Underpinning Attitudes of Biased Beliefs'. However, Pohlman focuses rather on the extent to which individuals are accountable for implicit bias, which she argues comes apart from responsibility in the sense of liability for blame.

2 Three points to note: i) whilst I am aware of disagreements between psychologists as to how best to capture the phenomena of implicit attitudes or bias (e.g. Fazio et al's MODE model, versus Dovidio et al's 'dual attitudes' approach), I think this characterisation is neutral between the these different understandings; ii) some people think that the IAT tracks in-group/out-group associations, rather than more stable associations between members of

There are three noteworthy features of this understanding of implicit bias. First, the *implicit* part of the bias pertains to the association, rather than to concepts implicitly held or to any implicit belief-like states. (Psychologists often refer to implicit associations as 'implicit attitudes'.) For example, an agent could explicitly entertain non-prejudiced thoughts about a member of a stigmatised group whilst unconsciously making cognitive associations with negative evaluations or stereotypic traits; she might then be described as having implicit negative attitudes or biases (I'm here using these terms interchangeably). Secondly, there are two parts to this understanding: the implicit associations (*having* the bias); and their influence on behaviour or judgement (*manifesting* the bias). This distinction corresponds to that made in some of the empirical literature between stereotype activation (the presence of accessible associations) and application (the use or influence of those in decision making and action).[3] It will be important when we come to consider *for what* individuals might be liable to blame. Finally, the associations in question are *automatic*, occurring without the instigation of the process being consciously directed or undertaken, and not directly subject to rational revision in the way our explicit beliefs are.[4]

How do we know that an individual harbours certain biases, for example, about black people and about women? This isn't something known by introspection, reflection, or self-report on one's motives: else the empirical data on implicit bias would not be surprising (or perhaps needed). But the presence and influence of such biases in the cognitive structures of individuals has shown up in a number of studies over the past two decades. One of the most well known and widely written about studies is the Implicit Association Test: a test undertaken on a computer which monitors the time it takes for the subject to pair up terms (e.g. positive or negative) and faces (e.g. black or white) or names (e.g. which sound stereotypically black or white). The pairing tasks are to be executed very quickly (in mere milliseconds), so as not to allow time for conscious reflection to guide responses. For many individuals, the time taken to pair up black faces with positive terms is greater than that for black faces and negative terms, or for white faces and positive terms (mutatis mutandis for black or white sounding names).[5] So for an individual who is *slower* to pair up black faces or words and positive terms (than pairing up white and positive terms), and *faster* to pair up black faces and words and negative terms (that white and negative terms), it is concluded that she has more accessible, and therefore other things being equal stronger, associations between black people and negative evaluations.

Another test is the 'evaluative priming measure', which primes individuals with some concept, and then measures how quickly or slowly they are able to recognise and categorise positive or negative terms. The idea is that if the individual has negative associations with the prime (e.g. a black face), she will be *faster* to categorise the negative terms or items as bad than the positive ones as good (because negative items will

---

stigmatised groups and negative or stereotypic traits. But there's reason to suppose at least some cases of biases are concerned with stigmatised, rather than just out-groups (because members of the stigmatised group themselves harbour biases); iii) by 'accessible' I do not mean 'accessible to introspection'. Rather accessibility is understood in terms of the speed and ease of associations made. For discussion of disagreement over precisely how we should understand 'association', see Greenwald, A, Nosek, B, Banaji M., Klauer, K (2005) Validity of the Salience asymmetry interpretation of the IAT: Comment on Rothermund and Wentura (2004) *Journal of Experimental Psychology: General* 134: 420-425.

3 See Gilbert & Hixon, 1991. The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*, 509–517.

4 Because the implicit biases appear to be best understood as kinds of accessible cognitive association, it is inappropriate to think of them as analogues of explicit belief; they will not have propositional content, for example.

5 Dovidio, Kawakami, Johnson & Johnson, (1997) On the nature of prejudice: Automatic and controlled processes *Journal of Experimental Social* Psychology, 33, 510-540; Nosek, B. A., A. G. Greenwald, and M. R. Banaji. (2007) 'The Implicit Association Test at Age 7: A Methodological and Conceptual Review' *Automatic Processes in Social Thinking and Behavior.* Ed. J. A. Bargh. Philadelphia, PA: Psychology Press.

have been made accessible by the black face prime).[6] Importantly, individuals who show such implicit biases often explicitly endorse non-biased, non-racist beliefs – although in one respect this is unsurprising, given the widespread norm that prejudiced expressions are unacceptable.[7] Thus these tests use implicit measures, to detect implicit biases of which the individual may not be aware, rather than asking individuals to simply report on their attitudes.

One might think that such experiments reveal associations that, discerned in the lab, are unlikely to play much role in the real world. But a range of studies suggest that various implicit negative associations, held by people who may disavow explicit racism, have an effect on judgement and behaviour outside the lab: the greater readiness to identify an indeterminate object in the hands of a black (rather than white) male as a gun; the less positive evaluation of the same CV when it they bears a woman's name rather than a man's; the differential hiring recommendations (more positive for the white applicants) made for black and white candidates when their qualifications were equally moderately good, but not wholly decisive.[8] Such studies are regarded as evidence that negative associations about stigmatised groups can be more specific than those identified in some IAT tests, concerning quite particular stereotypical and negative associations, and can influence judgement and behaviour 'in the field', beyond response times in IATs conducted in the lab.

1.1 Are we liable for blame for these implicit biases?

Many published studies have supported the hypothesis that individuals' behaviours are influenced by implicit biases.[9] This is cause for significant concern, not least if such biases are part of the explanation for patterns of subtle differential treatment, with effects such as the continuing under-representation of women and minority groups in a range of professions and educational environments.[10] Is it also cause for blame? Are individuals responsible and liable to blame for the implicit biases that they have? This question is important, as our answer to it has implications for how we might treat others and regard ourselves, for being influenced by these biases. Should we feel guilty for manifesting implicit biases? Should we expect others to do so? Is it reasonable to challenge and blame each other if implicit biases are manifested?

A number of philosophers have touched upon this question.[11] In her influential paper 'Unconscious Influences and Women in Philosophy,' Saul writes that we should *not* regard people as responsible for their biases:

> I think it is also important to abandon the view that all biases against stigmatised groups are *blameworthy.* My first reason for abandoning this view is its falsehood. A person should not be blamed for an implicit bias that they are completely unaware of, which results solely from

6 See Petty, Fazio, Brinol 2009, pp.1-9 for an overview of the IAT and evaluative priming measures. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 3-18). New York, NY: Psychology Press.
Fazio also notes that the term 'implicit' is sometimes used to refer to the *measure* as well as or rather than the *attitude* (what I have here referred to as the cognitive associations).

7 Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1 – 52). San Diego, CA: Academic Press; Pearson, Dovidio & Gaertner, (2009) The Nature of Contemporary Prejudice: Insights from Aversive Racism *Social and Personality Psychology Compass* 3, pp.314-338.

8 Payne, B. K. (2005) Conceptualizing Control in Social Cognition: The Role of Automatic and Controlled Processes in Misperceiving a Weapon. *Journal of Personality Social Psychology* 81: 181–92; Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science,* 11, 319-323; Valian, V, 1999 *Why So Slow? The Advancement of Women* MIT Press.

9 For an excellent overview, see also Jost et al 2009.

10 See Anderson (2011) *The Imperative of Integration* for discussion of the problematic effects of under-representation and (de facto) segregation.

11 See footnote 1 for other authors who have started to address this question.

> the fact that they live in a sexist culture. Even once they become aware that they are likely to have implicit biases, they do not instantly become able to control their biases, and so they should not be blamed for them. [12]

On this line of thought, it is simply a mistake to suppose that individuals are blameworthy for biases they harbour, insofar as their biases have the three characteristics, which plausibly exempt from blame, above. We can say that biases are *problematic*, and can encourage people to take steps to get rid of, or mitigate the influences their biases might exert on behaviour or judgement - but we shouldn't hold people to be blameworthy for the implicit associations or biases they are influenced by, insofar as the relevant awareness, control and causal etiology conditions are not met.

Scepticism about the extent to which individuals are responsible for implicit bias is also expressed in Kelly & Roedders' rich and highly suggestive exploration of the ethical implications of empirical findings about implicit bias.[13] They write that:

> Particularly in the case of implicit attitudes, it is salient that their acquisition may be rapid, automatic, and uncontrollable. These features, it might be thought, are related to features that establish blameworthiness – such as identification (Frankfurt) or reasons-responsiveness (Fischer and Ravizza). For instance, it might be said that the implicitly racist person doesn't identify with his implicit attitude, or that the attitude isn't responsive to reasons; thus we cannot hold a person fully accountable for those implicit attitudes. If this is right, one might say that such attitudes are morally wrong – and condemnable – but that the person himself cannot be blamed for having them (2008, p.532).

The thought here is that implicit biases might be such that they do not meet the conditions for moral responsibility. It is worth noting that Kelly and Roedder are tentative about this thought, acknowledging that whether the conditions for responsibility are met with respect to implicit bias will depend upon empirical findings about the nature of implicit biases:

> We are reluctant to embrace this solution wholeheartedly – it may turn out, for instance, that narrow-mindedness partially explains the acquisition of implicit racism (ibid).

I will suggest that the empirical findings do indeed suggest that there may sometimes be grounds for holding individuals liable to blame for their implicit biases. But what, precisely, might individuals be liable to blame for?

1.2. Responsibility *for what*?

It is worth distinguishing between three different things for which one might be liable to blame:

1. *Having* an implicit bias: simply having the cognitive or affective associations between groups and traits.
2. *Being influenced* in behaviour or judgement by these biases: manifesting it in behaviour or judgement.
3. *Responding to knowledge* that one is (or is likely to be) biased.

Saul points to the distinction between responsibility for being biased and for responding to the knowledge that one is likely to be biased, in her remarks that:

12 Saul, J. Unconscious Influences and Women in Philosophy forthcoming in *Women in Philosophy: What Needs to Change?* Edited by Fiona Jenkins and Katrina Hutchison

13 Kelly, D. & Roedder, E. (2008). Racial Cognition and the Ethics of Implicit Bias, *Philosophy Compass*, Vol. 3, No. 3: 522-540.

> [Individuals] may, however, be blamed if they fail to act properly on the knowledge that they are likely to be biased— e.g. by investigating and implementing remedies to deal with their biases .

It is important to distinguish between the backward-looking question of whether individuals are blameworthy for implicit bias, and the forward-looking question of when and how individuals should take responsibility for addressing their implicit biases. I am primarily concerned with the former kind of claim, about blameworthiness for implicit bias. However, I want also to distinguish between two things for which one might be blameworthy: *having* some implicit biases, and *being influenced* by them (as mentioned previously, between the stereotype or negative association being accessible, and the operation of this stereotype or negative attitude in deliberation and action)**.** This distinction is important for when we come to consider the arguments in detail.

In the remarks from Saul and from Kelly and Roedder, above, are considerations that provide the premises for five arguments:

a. an argument from the causal etiology of biases,
b. an argument from control (or lack thereof) with respect to biases,
c. an argument from lack of awareness,
d. an argument from lack of reasons-responsiveness, and
e. an argument from lack of identification.

I'm going to set aside the argument from lack of identification. This is because identification conditions for responsibility - roughly, the conditions that an individual must identify with, or endorse (perhaps wholeheartedly) the motives on which she acts - are presented by central proponents as *sufficient*, rather than necessary, for moral responsibility.[14] As such, even if an individual does not meet an identification condition for responsibility, there may be other conditions with respect to which it is appropriate to regard her as responsible (and hence liable to blame, in the absence of excusing conditions). In the next section, I consider each of the other arguments in turn.

2. Considering the Arguments Against Responsibility

We can say that an individual is responsible for some action or mental state or process when it reflects on her as an agent, in a way that makes it appropriate (absent excusing conditions) to hold her liable to blame for it.[15] One might claim that having and being influenced by implicit bias is not something that reflects on the individual as an agent either because it is a 'rogue' element in her cognitive structures (not subject to control or responsive to reasons, only there due to cultural influence), or because there are excusing conditions (she is unaware that bias is operating).

Each reconstructed argument is characterised by a claim about the conditions for responsibility, and an empirical claim about the nature of implicit bias. So, in assessing these arguments, we will want to ascertain whether these empirical premises are true, and whether the conditions for responsibility are to be accepted. I'll consider each argument in turn, before drawing preliminary conclusions about the legitimacy of treating individuals as liable to blame for biases.

14 Frankfurt, (1971), Freedom of the Will and the Concept of a Person, *Journal of Philosophy,* 68(1), pp.5-20.

15 Clearly this brief remark does not do justice to the wealth of literature on moral responsibility. But this thought is one that is shared by those compatibilists who focus on quality of will or rational capacities and incompatibilists who focus on control or avoidability alike. My commitments lie with compatibilism, but that makes little difference to the discussion here (although hard determinists who deny that individuals are ever morally responsible will not accept my starting assumption that individuals are, at least sometimes, morally responsible).

2.1 Argument from causal etiology.

One of the arguments suggested by Saul's claims is as follows:

1. Individuals cannot be held responsible for cognitive states or processes whose causal etiology lies wholly in factors out of their control.
2. Living in a sexist and racist culture is out of an individual's control.
3. Having implicit biases (against e.g. women, black people) results solely from living in a sexist and racist culture.
4. Therefore, individuals cannot be held responsible for their implicit biases.

Note that two different interpretations of premise 3 are possible:

3a. Having implicit bias (i.e. having certain cognitive associations) results solely from living in a sexist and racist culture.

3b. Being influenced by implicit biases (i.e. manifesting them in behaviour and judgement) results solely from living in a sexist and racist culture.

Thus we might conclude that individuals are not blameworthy for *having* biases, or negative cognitive associations; or that individuals are not blameworthy for *being influenced* by these biases in behaviour and action.

Let's grant the plausible premise 1, which (insofar as it pertains to the persistence, rather than just the presence, of a cognitive state) picks out a necessary condition for moral responsibility. More controversial in this argument is premise 3 (and its disambiguations, 3a and 3b), according to which having, or being influenced by, implicit biases, results solely from living in a sexist culture.

2.1.1 Variations in implicit biases.

The thought that implicit biases are solely the result of cultural determinants seems to be accompanied by two other assumptions, usually tacit. The first assumption is that it is likely that we all have implicit biases.[16] The second is that the extent to which we are implicitly biased is pretty much the same.[17] But empirical data suggests otherwise. A number of studies have shown that there appears to be considerable variation in the degrees of implicit bias that individuals display in experimental tests such as the IAT (Devine 1996)[18]. Just as individuals vary in the extent to which they are explicitly prejudiced, and in the reasons for and extent to which they care about not being prejudiced (Plant & Devine, 1998), individuals vary quite significantly in the extent to which implicit biases show up in, e.g. their response times to IAT (Devine et al 2002).[19]

This variation is not incompatible with the claim that implicit biases result solely from living in cultures that are sexist or racist: there are obviously variations in the kinds of experiences and cultural norms we're exposed to, so the varying degrees of implicit bias might correspond to how fortunate (or not) we've been in to what we've been exposed. However, various findings do provide a challenge to the claim that the implicit biases are the result solely of the culture we live in: the extent to which we manifest biases may rather be a function of other cognitive states we have, and over which we plausibly have control.

16 This thought is at work in Kelly & Roeder, 2008, who consider what to do given that it is likely that we are biased.

17 There are some exceptions to this: Jolls & Sunstein note that certain bias insulating policies might be problematic because they are not sensitive to the different extents to which individuals might be biased. See their (2006) 'The Law of Implicit Bias' *California Law Review*, 94, pp.969-996.

18 Devine 1996 Breaking the Prejudice Habit

19 Devine, P., et al. The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice. *Journal of Personality and Social Psychology* 82.5 (2002): 835–48.

2.1.2 What affects variation in degree of implicit biases?

A point of contention amongst empirical psychologists has been whether implicit measures reveal *personal* attitudes (attitudes of the agent which may be related to an individual's beliefs and evaluative stance); or whether the measures are rather reflective of the individual's awareness of social stereotypes (the content of which they repudiate). Increasingly, experimental findings have supported the claim that variations in the manifestation of implicit biases is the result of variations in personal attitudes, rather than in general knowledge of social perceptions. I detail three relevant findings below.

1) Some authors have suggested that there might be variation in the extent to which individuals *have* implicit biases. For example, Fazio et al (1995) write that their studies on the relationship between measures of explicit and implicit attitudes about race imply that:

> some individuals do not experience the automatic activation of any negative evaluation ... on encountering a Black person . (1025)

Fazio et al identify truly non-prejudiced individuals as subjects who demonstrated prejudice neither at the level of explicit beliefs, or implicit associations. Of course, this does not show that having such explicit beliefs influences the extent to which one has implicit biases. Were this the case, then implicit biases would appear to result from one's environment *and* one's explicit beliefs and values. But we could think it just as likely that implicit prejudice influences explicit beliefs as that the explicit beliefs affect the implicit prejudice. So it would be a mistake to infer, at this stage, that the extent to which one is implicitly biased is likely to be *influenced by* one's explicit beliefs and values. Were we to make this inference, it would invalidate the argument from causal etiology against responsibility for bias).

2) In her early work on race bias (and contra Fazio's claims above), Devine (1989) argued that individuals are likely to hold the same cultural stereotypes - so are likely to *have* the same implicit associations. In her later work, she identified different variables that might influence the extent to which individuals manifest these biases in behaviour and action. One important finding concerns the difference in bias manifested between individuals who see behaving in a non-prejudiced way as important in itself, and those who see it as (also) important because of social norms prohibiting it, or social sanctions that one might face for so acting (and those who see non-prejudiced behaviour as of little importance). Devine et al (2002) present evidence for the claim that those individuals who hold non-prejudiced behaviour to be important in itself appear prone to display less bias in the kinds of tests (specifically, tests for associations between black and white face or name primes and positive and negative word associations) described in section 1. The suggestion is that the negative or stereotypic associations may be weaker in these individuals, or they may have more effective mechanisms for regulating the manifestation of bias. Crucially, though, there is no suggestion here that this is done consciously or with effort, or indeed that this regulation is in response to an awareness of implicit bias. Rather the hypothesis is that preconscious regulatory systems are at work: a sub-personal, or automatic inhibitory system may prevent the influence of negative associations on behaviour and judgement. If so, then the manifestation of bias would appear to be a function of the agents attitudes, values and beliefs, rather than solely the culture they live in: rather, having certain kinds of non-prejudiced explicit beliefs features as part of the story in claims regarding the extent to which individuals *manifest* implicit biases (focusing now on premise 3b).

3) Finally, in more recent work on how we might understand the relationship between implicit and explicit beliefs, Nosek (2005) claims that the manifestation of bias does vary with self-reported preferences, such that it is plausible to suppose that there explicit and implicit attitudes are related. However, this relationship might be moderated by various additional considerations such as the strength of one's preference for the

target object, one's concerns about self-presentation, and the extent to which one's preferences are in step with social norms. Because of the automatic nature of implicit attitudes, it is unsurprising that with respect to e.g. negative race biases, when an individual is concerned with self-presentation, explicit and implicit attitudes diverge (because explicit attitudes might be shaped by self-presentation concerns, whereas implicit attitudes are less malleable).
One of the striking findings from Nosek is that the relationship between implicit and explicit attitudes was stronger (more convergent) when the individual's evaluative attitudes were perceived to be distinct from the perceived (by the subject) social norms. This makes it difficult to maintain the claim that, at least with respect to evaluative attitudes, IAT results indicate culturally absorbed knowledge or stereotypes, rather than personal attitudes. Nosek concludes that 'this effect is difficult to reconcile with the hypothesis that ... cultural knowledge influence[s] IAT performance' (p.579).[20]

If we accept these findings, then we should reject at least one version of the argument from causal etiology against responsibility for bias. On the one hand, it seems genuinely an open question as to whether individuals differ in the degree to which they *have* biases; so it may well be that the presence of certain strong negative or stereotypic associations is due to the result of (sexist, racist) cultures, and that it is inappropriate to hold individuals responsible for *having* biases (the argument may be valid, with premise 3a). But there is reason to suppose that *being influenced* by bias is not solely the result of the sexist or racist culture in which one lives, and that premise 3b of the argument is false. Being influenced by implicit bias appears also to be an effect of the kinds of explicit beliefs and evaluations individuals make, as well as the strength of the agent's commitment to these values. So we cannot conclude, on this basis, that individuals are not blameworthy for being influenced by bias. The factors which influence the manifestation of bias are important in relation to the argument from control, to which I now turn.

2.2 Argument from control (lack thereof).
Suppose we grant that the influence of bias is not something that results solely from cultural influence, but is affected also by the explicit and consciously held beliefs and values of an individual. We might still think that it is inappropriate to hold an individual liable for blame for the influence of the implicit associations on behaviour, because neither the presence of these associations, nor their influence, is under an individual's direct or immediate control. This is the second argument we took from Saul's remarks:

1. Individuals cannot be held responsible for cognitive states over which they do not have immediate and direct control.
2. Implicit biases are not under an agent's immediate and direct control.
3. Therefore, individuals cannot be held responsible for implicit biases they harbour.

Again, this argument could run in two different ways, depending upon our interpretation of premise 2:

2a. Having implicit biases is not under an agent's immediate and direct control.
3a. Therefore, individuals cannot be held responsible for implicit biases that they have.

2b. Manifesting - being influenced in behaviour and judgement by - implicit biases is not under an agent's immediate and direct control.
3b. Therefore, individuals cannot be held responsible for the influence of implicit biases on behaviour and judgement.

20 See also De Houwer, et al (2009) at p.353 for discussion of this dispute, and the conclusion that 'few arguments remain to support the claim that IAT effects are causally influenced by extrapersonal views [i.e. knowledge of cultural stereotypes].' However, I note in the concluding remarks of this essay that there may be reason to suppose that this may differ according to what kind of association the IAT is testing in any one study.

There are two questions that need to be addressed: first, what kind of control, if any, do individuals have with respect to implicit biases? Secondly, is it right to suppose that direct and immediate control over cognitive states is required for responsibility?

2.2.1 Direct Control over biases

Do individuals have direct and immediate control over biases? It is common for authors to contrast implicit attitudes and biases with controlled processes (see Dovidio & Gaertner, 2004; Gaertner & Dovidio 2000, Pearson et al 2009).[21] Because these processes are automatic, and because the agents are typically not aware of their operation, it is assumed, they are not under the agent's control.

I'll return to consider the argument concerning awareness shortly. However, it is worth noting at this stage that some findings suggest that individuals *do*, at least sometimes, appear to be able to exercise direct and immediate control over the extent to which implicit biases influence behaviour. For example, some studies have lead researchers to be optimistic about individuals' abilities to exert conscious control in attempting to suppress any negative or stereotypic associations. When instructed *not* to display, e.g. sexism or racism, individuals were recorded as in fact showing less biased responses in experimental tests for implicit attitudes (Monteith et al 1998b).

However, serious concerns have been raised about these studies. Whilst individuals can, for a limited period, suppress the influence of implicit bias, many studies have shown that this tends to result in a 'rebound effect' – whereby having tried to suppress negative or stereotypical biases, the influence of such biases are more strongly shown when this effortful suppression is not maintained. One hypothesis that might explain this rebound effect is that the unconscious 'monitoring' process for the presence of (for example) the stereotype means that it is made more accessible to later activation (Galinsky & Moskowitz, 2000; Macrae, Nodenhausen, Mine & Jetten 1994).[22]

Whilst strictly speaking, then, some empirical studies support the claim that individuals at least sometimes have direct control (so that premise 2b is false), it is not clear that we should want to place too much emphasis on these studies. Firstly, the effectiveness of conscious suppression seems at best limited. Secondly, given the likelihood of the rebound effects, it is problematic to hold individuals responsible for not trying to exercise such control. If one cares about not being biased, then conscious suppression would be a risky strategy which would appear to later increase the influence of bias.[23]

---

21 Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999 *Psychological Science* 11: 319–323; Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances inexperimental social psychology* (Vol. 36, pp. 1–51). San Diego, CA: Academic Press; Pearson, Dovidio and Samuel L. Gaertner, 2009, The Nature of Contemporary Prejudice: Insights from Aversive Racism, *Social and Personality Psychology Compass* 3. pp.1-25

22 Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, andin-group favoritism. *Journal of Personality & Social Psychology, 78*(4), 708–724; Macrae, N., Bodenhausen, GV., Milne, AB. & Jetten, J. (1994). 'Out of Mind but Back in Sight – Stereotypes on the Rebound'. *Journal of Personality and Social Psychology*, vol 67, no. 5, pp. 808-817

23 Monteith, Spicer & Tooman, 1998 Consequences of Stereotype Suppression: Stereotypes on AND Not on the Rebound, *Journal of Experimental Social Psychology* Volume 34 (4) pp. 355-377. Monteith et al have argued that the propensity to experience any such rebound effect will depend on other features of the individual: those who are low-prejudiced (on explicit measures) are less susceptible to it (Monteith et al, 1998a). They propose other factors that may decrease susceptibility to implicit biases, and conclude that conscious suppression may, in the presence of these features be an effective method of regulation. See Monteith, Sherman & Devine (1998b) Suppression as a Stereotype control strategy. *Personality and Social Psychology Review,* vol.2(1), pp62-83. These findings remain controversial, however, and for the reasons above we should hesitate to hold individuals

2.2.2 Is direct control a necessary condition for moral responsibility?
Whether or not individuals are able to directly control the influence of biases will matter a great deal to the issue of responsibility if we accept premise 1 of the argument from control, above:

> 1. Individuals cannot be held responsible for cognitive states over which they do not have immediate and direct control.

The idea that control is necessary for responsibility is a plausible one. In the literature on free will, the idea that a necessary condition for free will and responsibility for some action, A, is that one is able to do not-A (and so has direct control over whether one does A), garners considerable support.[24] But many have rejected this condition, and indeed we might doubt whether this premise is true.[25] I here set out some of the considerations pertinent to our present concern with responsibility for bias, which speak in favour of rejecting the direct control requirement.

Lots of the things we are able to do are the result of the exercise of long range, rather than direct and immediate, control: the ability to sustain concentration, play the piano well, lose or gain weight, speak a second language. The doing of these things is not under an individual's direct and immediate control, but rather the result of 'long range' control. We are able to exercise (direct) control over a series of intermediate steps (placing hands on a keyboard, increasing food intake), such that we have non-immediate control over whether we are able to do those things. Many of these activities are, in the usual run of things, morally neutral, but it is nonetheless appropriate to regard each other as responsible agents when we engage in such activities, or exercise the long range control necessary for their pursuit. Moreover, long range control may be important to some morally relevant activities. The cultivation of virtue, such that one is able to act generously, is something that results from long range control, according to some virtue ethicists.

But these are examples of skills or *activities* for which, whilst we cannot do them at will, we are plausibly responsible. What about responsibility for cognitive states? The question of whether direct (or 'voluntary') control is a necessary condition for responsibility arises in debates about epistemic obligation more generally. It has been asked whether individuals ought (or are permitted, or have a right) to hold certain beliefs - and can be held responsible (epistemically or morally) for failing to do so. On the assumption that ought implies can, it is difficult to see how individuals could be obliged to believe p unless it is the case that they can - voluntarily, 'at will' or by deciding to do so - believe p. And believing p doesn't seem to be the kind of thing we can do voluntarily, or at will – we don't have direct control over our beliefs in this way.

Nonetheless, it has been argued that this lack of direct control does not confound the claim that individuals have epistemic obligations, and are responsible for meeting these. Two lines of argument are pertinent here.

First, consider the indirect kind of long range control we might have over our beliefs (Feldman, 2000;

responsible for failing to attempt direct control.

24 See e.g. Van Inwagen, P. 1975. The Incompatibility of Free Will and Determinism, *Philosophical Studies*, 27: 185–99; Ginet, C. 1996, In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing, *Philosophical Perspectives*, 10: 403–17; Wideker, D. 1995. Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities, *Philosophical Review*, 104: 247–61.

25 Those who have rejected this condition for responsibility include Frankfurt, H. 1971. Freedom of the Will and the Concept of a Person, *Journal of Philosophy*, 68: 5–20; Wolf, S. 1980. Asymmetrical Freedom, *Journal of Philosophy*, 77: 157–66; Fischer J. & Ravizza, M. 1998. *Responsibility and Control: An Essay on Moral Responsibility*, Cambridge: Cambridge University Press.

Hieronymi, 2008).[26] We cannot simply decide to believe that p (e.g. that the Ozone layer is depleting). But we could undertake enquiry and take into account various sources of evidence so that we form appropriate beliefs about the Ozone layer (that it is, indeed, depleting). We could even undertake an enquiry that sets out to ensure we have the belief that the Ozone layer is depleting, by being selective about the evidence we consider – although doing so is epistemically (and perhaps even morally) problematic. Precisely because we have this long-range control, we can be held responsible (and sometimes blameworthy) for our beliefs.[27] The same can be said for cognitive states that involve evaluation: Murdoch describes the attentional processes by which a woman comes to revise her initially negative view of her daughter-in-law.[28] It is plausible that we have this kind of indirect control not only over our beliefs, but also over our affective responses also. These cases are ones in which the agent has intentionally exercised long range control.

Secondly, we might think it appropriate to hold individuals responsible where their cognitive states are reflective of their evaluative stance, or their 'take on the world', so to speak. Support for this view is found in the work of both Hieronymi, and Arpaly. On Hieronymi's view, it is appropriate to hold people responsible for their beliefs *even in the absence of direct control, and in the presence of only limited indirect control*. She writes: '[b]ecause these attitudes [beliefs] embody our take on the world, on what is or is not true or important or worthwhile in it, we control them by thinking about the world, about what is or is not true or important or worthwhile in it' (2008, p.371). Whilst we cannot simply believe whatever we want, or think it good to believe, we can (perhaps non-intentionally or unconsciously) be selective about what evidence we look for, and how we interpret or weigh it (given the constraints on time and effort, such selectivity will not always be problematic). So, even if we cannot directly control our beliefs, by believing at will, the fact that they are responsive to evidence means that they are reflective of certain evaluative stances that we might take with respect to seeking and weighing evidence. We might say, then, that we are *indirectly* responsible for these beliefs insofar as they are causally and rationally related to other things that we *are* directly responsible for.

Further, Arpaly's remarks (on the nature of moral worth) elaborate on this way in which we might be indirectly responsible for our beliefs. For most people, exposed to the vicissitudes of daily life, certain false sexist beliefs, or false beliefs about a Jewish conspiracy, she suggests, cannot be properly thought of as *honest mistakes*. Rather, such beliefs, she hypothesises, are likely the result of motivated irrationality, for such beliefs have to be maintained in the face of what should be recognised as adequate countervailing evidence. Insofar as it is reasonable to hold individuals responsible for the ill will or moral indifference, which sustains these beliefs, we can hold them responsible for their false and prejudiced beliefs (over which they do not have *direct* or voluntary control).[29] These beliefs are a function of other states over which individuals do have control – so individuals can be held liable to blame for these beliefs.

There are two central thoughts in the claims sketched (albeit briefly) above: firstly, that direct and immediate control is not necessary for moral responsibility. This is plausibly the case with respect to both

26 Hieronymi, P. Responsibility for Believing, *Synthese* 161, no. 3 (April 2008): 357–373; Feldman, R. (2008). Modest Deontologism in Epistemology**.** *Synthese* 161 (3):339 - 355.

27 It is worth noting Hieronymi's emphasis on the point that even this kind of indirect influence is significantly limited: even in exercising indirect influence, an individual cannot believe *for just any reason*. She can properly hold some belief only for reasons bearing on its truth.

28 Murdoch, I. 1970[1985] *The Sovereignty of Good*, (Routeldge, London) 32-35.

29 A contrast case Arpaly provides: an individual who holds racist or sexist beliefs due to a very insular life in which he has been exposed to no counter-evidence which could persuade him of the falsity of his beliefs (e.g. beliefs that black people are of inferior intelligence, that women cannot hold leadership roles). If such an individual's beliefs really are the result of unfortunate exposure to partial evidence, rather than irrationally motivated, we would expect her to revise her beliefs upon e.g. leaving her closeted existence for work or university and encountering intelligent female black colleagues or lecturers. See Arpaly, 2003, *Unprincipled Virtue* OUP.

actions, and beliefs, both factual and evaluative. So, premise 1 of the argument from control should be rejected. The second important point here is that this kind of (constrained) indirect influence is plausibly sufficient for moral responsibility with respect to some cognitive state, insofar as these cognitive states reflect the agent's stance, or take on the world (by reflecting her partial or ill-motivated attention to the available evidence, for example).

We are now in a position to ask whether individuals exercise indirect influence over implicit biases, either intentionally or otherwise.

2.2.3. Indirect Control over biases

Whilst we have seen serious concerns with about the claim that individuals have, and ought to exercise, direct control over their biases, there is considerably more data on the extent to which individuals have *indirect* and *non-immediate* control over the influence of implicit biases. I'll here briefly mention three ways in which it appears that individuals have control, albeit indirect and non-immediate, over the manifestation of biases in behaviour and action.

Note that the first two kinds of control I discuss require intentional undertaking, so will only be relevant to cases in which individuals *know* they are biased and seek to mitigate this. This is an important kind of control, but recognising this is consonant with Saul's remarks about individuals being blameworthy for failing to take steps to remedy bias once they are aware of it. This kind of control is not one that will be useful to those who are not aware that they are biased. But it is nonetheless important to identify the kinds of control individuals can exercise over biases, because this helps us to understand in more detail the operation of implicit biases, and to scrutinise the experimental findings in relation to these kinds of control. These are considerations that are pertinent to the question of what individuals may do in order to *take* responsibility for implicit bias.

However, I will also suggest that the third kind of control (indirect) can make it appropriate to hold individuals responsible and sometimes blameworthy for the influence of bias on behaviour, even in the absence of awareness of bias (and this takes us to face squarely the argument from awareness).

i. Intentional long-range control I: exposure to counter-stereotypical exemplars or members of stigmatised groups

Some things that *are* under our voluntary control appear to have indirect influence on the extent to which biases influence behaviour and judgement. One of those is exposure to members of the stigmatised group, or to counter-stereotypical exemplars. We might not be able to rid ourselves of, or limit the influence of biases at will. But just as we are able to seek out evidence which influences our beliefs, empirical findings suggest that we are able to undertake steps which mitigate the influence of biases.

Blair (2002)[30] reports on studies in which participants who were exposed to pictures of counter-stereotypical exemplars showed less bias than individuals in control conditions. It appears that even simply thinking about individuals who are counter-stereotypical (e.g. an admired black person) can limit the manifestation of negative implicit bias against black people (the importance of counter-stereotypical exemplars is emphasised in Saul, ms).[31]

30 Blair, I. 2002. The Malleability of Automatic Stereotypes and Prejudice, *Personality and Social Psychology Review*, 3: 242-261. See p.249.

31 Heidi Howkins Lockwood Counterstereotypical and Uncanny Exemplars: Moving Beyond the Mere Maximisation of Smartness (ms.) raises some important concerns for this strategy also: namely, are there any constraints on the kinds of exemplars that are effective? Should we be seeking the *most* counter-stereotypical of exemplars, or rather any individuals who do not meet the stereotyped role?

In further studies, simply having contact with the stereotyped individual served to decrease the influence of bias. On completing an IAT, those individuals who did so in the presence of a black experimenter displayed less race bias than those who completed the test in the presence of a white experimenter (Lowery et al 2001). One hypothesis advanced is that automatic processes - such as implicit associations - are sensitive to the social context. Another is that they are sensitive to evidence - and so the associations are weakened when presented with clear counterexamples (see Lowery et al for discussion).[32]

So, one indirect way of controlling the influence of bias is to intentionally increase one's exposure to members of stigmatised groups, and to have present to mind counter-stereotypical exemplars. This seems to suggest that individuals have the kind of long range control, at least over the manifestation of biases, discussed in the previous subsection.[33]

ii. Intentional Long-range control II: Implementation intentions

Recent studies have revealed another strategy with which individuals may be able to intentionally exercise long range control over their biases. Studies in empirical psychology have argued for the efficacy of 'implementation intentions' in bringing about changes in responses guided by implicit biases (see Gollwitzer, Bayer & McCulloch, 2005 and Webb & Sheeran 2007, for overview).[34] Implementation intentions differ from (straightforward) intentions, either in having a built in conditional ('If I'm in condition C, I'll do X'), or being tied to particular environmental cues ('When I arrive at D/at E o'clock/when I see F, I'll do X'). In studies where participants attempt to change behaviour (often addictive or habitual behaviour), those in the implementation intention condition are reliably more successful in achieving the specified behaviour change than those who just form general intentions ('I'll do X'). The hypothesis is that the agent's goals are, with implementation intentions, sensitised to environmental cues, such that the goals (to do X) are automatically activated in certain contexts.

Recently, it has been suggested that implicit biases might be effectively regulated with implementation intentions.[35] For example, if one harbours biases about Muslim people (e.g. cognitive associations between Muslims and terrorism), the influence of these biases could be controlled by forming an implementation intention for one's responses in the IAT's: 'If Muslim names and peace are at the top of the screen, then I respond especially fast to Muslim words and peace words!' (Webb, Scheeran & Pepper, 2010, p.11). The formulation of such implementation intentions resulted in faster response times to *Muslim* and *peace* prompts than participants in other conditions. Indeed, with the implementation intention, these response

32 Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology, 81*, 842–855.

33 See also Kang & Banaji 2006, Fair Measures: A Behavioral Realist Revision of "Affirmative Action" *California Law Review* 94, 1063-1118. See esp pp.1101-1108 for discussion of the social contact hypothesis and the de-biasing effects of counter-stereotypical exemplars.

34 Gollwitzer, P. M., Bayer, U. C., & McCulloch, K. C. (2005). The control of the unwanted. In J. A. Bargh, J. Uleman, & R. Hassin (Eds.), *The New Unconscious* (pp. 485–515). Oxford: Oxford University Press; Webb, T.L. & Sheeran, P. (2007). How do implementation intentions promote goal attainment? A test of component processes. Journal of Experimental Social Psychology 43, 295-302.

See also Gollwitzer, P. M., & Schaal, B. (1998). Metacognition in action: The importance of implementation intentions. *Personality and SocialPsychology Review, 2*, 124–136

35 Webb, Sheeran and Pepper 2010 Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials *British Journal of Social Psychology*, DOI:10.1348/014466610X532192;
Stewart & Payne, Bringing Automatic Stereotyping Under Control: Implementation Intentions as Efficient Means of Thought Control *Personality and Social Psychology Bulletin* 2008 34: 1332-1345.

times were commensurate with the association between *Scottish* and *peace* targets, suggesting that the formulation of an implementation intention mitigated almost entirely the manifestation of biases.

However, I think there are further questions about these results in terms of whether they can be generalised to the manifestation of biases in other contexts: where the implementation intention is specifically attuned to the completion of an IAT, considerable attention is needed to whether similar intentions could be applicable for non-laboratory conditions. In particular, the form of the implementation intention outside of the lab might need to be considerably different: a 'fast response' will not be appropriate (or even coherent!) in many every day contexts. However , implementation intentions have recently been shown to effectively shape behaviour, such as for example, influencing the seating distance between experimental participants and members of a stigmatised group (those with implementation intentions to show warmth as soon as they had the chance to do so, sat nearer).[36] One might suppose that alternative, more general, implementation intentions could be formulated: 'If I see a Muslim, I will think 'peace''. But it is worth noting that the structure of these intentions is quite different from those tested by Webb et al in the lab, focusing on the implicit association itself, rather than the behavioural response to certain targets. This is clearly fertile territory for more empirical work.

However, findings about the efficacy of implementation intentions even in the limited context of the laboratory tests have implications for our concerns here. For this research indicates once again that individuals can exercise long range control over the manifestation of bias in behaviour, at least in some contexts.

iii. Unintentional Indirect Control: Influence of explicit beliefs and values.

The previous two sub-sections outline long-range control strategies that individuals might intentionally undertake to limit or remove the influence of implicit biases. The control condition for responsibility, then, may be met (at least in some contexts) when individuals are aware that these kinds of long range control over biases are required. But it is worth returning to some of the studies mentioned earlier, pertaining to the unintentional influence of explicit beliefs and values on an individual's implicit biases: namely, the manifestation of implicit bias might be influenced by the explicit beliefs, values and goals that individuals hold.

Indirect control in relation to beliefs

These considerations support the claim that individuals can indirectly and unintentionally control the manifestation of biases even when they are unaware of the possibility of influence. As detailed above, in studying race bias, Devine et al. (2002) found patterns of responses which indicate that individuals who are highly committed to responding without prejudice, for its own sake (and not rather or also for reasons of social pressure or norms), manifest significantly less negative race bias across a range of tests for implicit biases.

Importantly, they argue that their results cannot be explained by such individuals exerting conscious control (in at least one of their studies, individuals completed the tests under a heavy cognitive load, to try to prevent any attempts at conscious control). Rather, they suggest that strong commitments to avoid prejudice may weaken any implicit negative associations, or may limit the activation of such associations in the production of behaviour and judgement (2002, 845-847).

The claim isn't simply that having non-prejudiced beliefs makes one less likely to manifest bias, so that

36 Tidswell, K., Sheeran, P. & Webb, T. L. (2012). Self-regulation of the impact of implicit attitudes on behavior. Unpublished manuscript. University of Sheffield.

individuals ought to make sure they reject explicit racist beliefs. Rather, important differences showed up with respect to different anti-racist beliefs. Individuals who endorsed non-prejudiced behaviour for its own sake (e.g. 'I attempt to act in nonprejudiced ways toward Black people because it is personally important to me') rather than for instrumental reasons (e.g. 'If I acted prejudiced toward Black people, I would be concerned that others would be angry with me') manifested less bias in experimental conditions. Those who endorsed such instrumental reasons, or endorsed instrumental reasons *in addition to* the non-instrumental reasons for avoiding prejudiced responses, displayed greater bias (that is, the difference between response times to congruent and incongruent pairs was larger). The reasons for which one cares about non-prejudiced behaviour then – and in particular, a commitment to it for its own sake – is indirectly related to the degree to which individuals manifest bias.

Indirect control in relation to goals and attitudes

Explicit beliefs and values may affect the influence of bias in another way. Moskowitz & Li (2011) draw attention to the importance of having the goal of treating people non-prejudicially, and argue that individuals who are committed to this goal are less likely to manifest bias in responses to tests for implicit biases.[37] Crucially, they draw attention to the importance of *activating* the relevant goals (rather than merely having them), where a goal's being active means that it is operative in the production of action.[38] We have all sorts of goals, and they can't all be at work in producing action at once. The idea, broadly speaking, is that only some of our many goals are active at any one time, and their activation blocks the activation of others – else all of our goals will be active all of the time, and this might hinder the effective pursuit of any one of them.

Importantly, Moskowitz & Li set out evidence in support of the hypothesis that the activation of certain goals (such as the goal to treat people fairly) can, at the sub-personal level regulate the manifestation of biases. When the goal to treat people non-prejudicially was active, less biased responses were recorded. In contrast, they found that in the experimental condition where prior success in relation to the goal to treat individuals non-prejudicially was contemplated – and therefore that goal deactivated, because achieved - then the inhibition of biases is diminished. This is because other goals, which implicit biases may serve (such as efficiency), are no longer then inhibited. Importantly, the non-prejudice related goal may only inhibit the influence of bias when it is active. These findings are consonant with those of Park et al (2008), who argue that individuals with automatized goal to behave in a non-prejudiced manner displayed less race bias on a number of tests for implicit attitudes. [39]

For our purposes, this brings to light the following important distinction: distinction between having a particular goal (of treating people non-prejudicially) and that goal being 'active' in the production of action. This suggests that not only caring strongly about treating people non-prejudicially for its own sake, but having those goals activated, is important in mitigating the influence of negative biases on action. Whether or not we have a particular goal is something that is under our control. But more importantly, the key modulating consideration in this experiment was whether the individual dwelt on a case of prior success in treating people fairly, or a case of failure. Reflecting on a case of success, they claim, deactivated the goal such that it no longer played a bias-blocking role in their responses.

37 Moskowitz, G.B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. Journal of Experimental Social Psychology.

38 Psychologists have identified various ways of discerning when a goal is active: these include increased value attached to the means to one's goal; decrease in value attached to means when goal is secured; increased value attached to, and influence of, the goal as the distance from goal attainment decreases; increased influence of the goal as the probability of achievement increases; decreased influence of competing goals, etc. (Forster, Seven Principles of Goal Priming 2007.)

39 Park, S. H., Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination. *Social Cognition* **26**:401–419

This suggests that even whilst holding the goal of treating people fairly, being easily satisfied or confident that one has not and does not treat others prejudicially may lead to being *more* biased. If such an attitude means the goal of treating people fairly is not activated, then an individual with these beliefs is preventing an effective automatic bias limiting strategy from taking effect. The importance of the goal remaining active counsels against self-satisfaction with one's efforts to treat others fairly; a certain humility would appear to be important in ensuring that the goals to behave non-prejudicially remain active.

Again, there is significant further work to be done here, including on the question of how one might best 'activate' or 'trigger', a particular goal. Moskowitz & Li asked individuals who had the relevant goal to dwell on a case in which they had failed to live up to it. It is important to note their emphasis that a goal can be activated (by priming) and operative, such that it is effective in blocking stereotype activation, 'even if [the subject] is not consciously aware of the goal' (2011, p.114). Given that we cannot always introspectively access whether the relevant goals are activated, finding out more about the specific strategies for ensuring one's bias blocking goal is activated, would be fruitful.

How do these findings inform our understanding of an individual's responsibility, and liability to blame, for bias? We have seen that increased influence of negative implicit associations are related to - caused by - states that are under our control, such as the explicit beliefs about the reasons for treating individuals fairly, or a ready satisfaction that one has succeeded in doing so. In these cases, we might think it appropriate to hold individuals responsible for not caring enough, or in the right way, about non-prejudiced behaviour, or being unduly satisfied with one's treatment of others as fair. In such cases, whilst the implicit bias is not itself under the direct control of the agent, it is indirectly a function of these explicit beliefs or attitudes. I think that in such cases it is appropriate to hold individuals responsible for the increased manifestation of biases.

I want to provide additional support for this claim by considering how we might judge the following two examples: first, consider an individual who has prejudiced beliefs about black people, and as a result experiences certain affective and physiological states of discomfort whilst in the presence of black people. Even if she tries to suppress her explicit prejudice (by, say, speaking in a polite tone) she cannot but give off subtly different cues which negatively affect her interactions (empirical studies suggest that subtle signs of discomfort, such as increased blinking, more infrequent eye contact are often detected in interactions).[40] Such affective and physiological states are not under the agent's direct and voluntary control; nor may she be aware of them or their effects on her interaction. But they are causally related to the explicit beliefs that she has, and over which she has long range control.

Next, consider an individual who believes it is important not to treat people in a prejudiced way, because doing so would cause general tension and anger, and is confident that she has in the past, and will continue to in fact treat people fairly. She acts in accordance with these beliefs, but nonetheless does give off subtly different cues (increased blinking, more infrequent eye contact), as a result of implicit negative associations she harbours, and these affect the quality of her interaction. These responses are not under her direct and voluntary control. But they are causally related to the extent and strength of her commitment to caring about treating people fairly for its own sake. Were she to care more about treating people fairly for its own sake, or were she to be more scrupulous about whether she meets her ideals - and less easily satisfied that she does - there is reason to suppose that less bias would affect her interactions.

40 Amodio, Harmon-Jones & Devine, 2003, Individual Differences in the Activation and Control of Affective Race Bias as Assessed by Startle Eyeblink Response and Self-Report, *Journal of Personality and Social Psychology,* Vol. 84, No. 4, 738–753

I find it plausible to hold both of these individuals responsible and liable to blame for the kind of automatic responses that affect the quality of their interactions, insofar as they are causally related to their explicit beliefs and attitudes, and so something over which they have indirect control. The two cases are structurally analogous. So if one is to deny that individuals are responsible for the manifestation of implicit bias due to the lack of direct control, then one ought also to deny that individuals are responsible for the negative physiological responses that affect the quality of interactions in the first example (though, of course, the extent to which one may be blameworthy may vary with the blameworthiness of the explicit attitudes). This seems to me an implausible denial: rather the relationship between the physiological responses (or biases) and explicit values or beliefs (over which agents do have control) makes it appropriate to hold individuals responsible for these automatic responses.

iv. summary

The main argument of this section has been to reject the requirement for direct and immediate control as necessary for responsibility. I suggested that there are good reasons for supposing indirect control or long range control is sufficient for moral responsibility. I then presented evidence supportive of Saul's remark that individuals can be held responsible for failing to respond to the knowledge they are biased; this is pertinent to the issue of what individuals can do to take responsibility for implicit bias. But further, in relation to the question of blame for being influenced by implicit bias, I argued that individuals might also be reasonably held responsible for the manifestation of biases where this is causally related to explicit beliefs or attitudes individual holds.

At this point, the following objection might be raised: it isn't reasonable to hold someone liable for blame for failing to undertake long range control strategies to mitigate bias if they are not aware of the existence of biases in their cognitive processes. Nor is it reasonable to hold individuals responsible for the biases that are causally related to their explicit beliefs if they are unaware of these causal corollaries of their beliefs, values or goals. This brings us directly to the argument from unawareness, so I'll turn to consider that in more detail now.

2.3 Argument from lack of awareness

Recall the third argument that we identified earlier, for the conclusion that individuals are not responsible for cognitive states or features of which they are unaware. This argument must proceed as follows:

> 1. Individuals can only be held responsible for cognitive features that they are aware they possess.
> 2. Individuals are not aware of cognitive features such as implicit bias
> 3. Therefore, individuals cannot be held responsible for their implicit biases.

Again, we can identify the two versions of premise 2:

> 2a. Individuals are not aware of the presence of cognitive features such as implicit bias
> 2b. Individuals are not aware of the influence on their decisions and actions of cognitive features such as implicit bias.[41]

Premise 2a, which pertains to the existence of implicit biases, seems plausible: the presence of cognitive

41 DeHouwer et al (2009) also identify two further dimensions of awareness that are relevant: awareness of the stimuli in the experimental tests, and awareness of the origins of the attitude being tested (see p.357). De Houwer, Teige-Mocigemba, Spruty & Moors, (2009) Implicit Measures: A Normative Analysis and Review, *Psychological Bulletin*, 135(3) pp.347-368.

associations of the sort discerned on implicit association tests is most likely something we are unaware of. Were we able to detect implicit biases by means of introspection, we would not need such sophisticated indirect measures to discern the presence of such biases.

What of premise 2b, which speaks to the manifestation of biases in behaviour? Interestingly, some research has shown that individuals are sometimes aware of the discrepancies between how they would act and how they believe they ought to, thus demonstrating an awareness of their proneness to being influenced by bias (if not an introspective awareness of the bias itself operating). For example, Monteith & Voils (1998) found that individuals reported different degrees of discrepancy between how they believed they *should* act in a given situation, and how they *would* act (this was true of both high and low-prejudiced individuals). With low-prejudiced individuals, these reported discrepancies reliably correlated with the degrees of implicit bias manifested on tests for implicit attitudes, indicating that individuals were accurately tracking the extent to which their actions were biased and fell short of their non-prejudicial normative standards.[42] In later studies, findings suggest that sometimes (64% of the participants in the race IAT), individuals are aware of their discrepant responses on IATs, and some of those individuals (37%) attribute those discrepancies to negative attitudes they suppose they harbour (Monteith et al 2001).[43] How to make sense, then, of the cases in which individuals report surprise – and shock – at the implicit biases revealed in experimental contexts? There are two possible explanations: first, these individuals may be in the group of individuals who are *not* aware of the discrepancies in action. Second, it might be that occasionally factors other than implicit attitudes affect IAT results – so the individuals may be right to be surprised if the results are attributed to negative implicit attitudes.[44]

Two points are worth emphasising here, though: firstly, these findings about individuals' awareness of discrepancies were limited to the highly artificial experimental context (the discrepancy report exercise and the bias manifested on the evaluative task set), and it isn't clear whether we can generalise to other contexts. Secondly, when assessing the discrepancies, very few people reported a small discrepancy,[45] and even those who reported small discrepancies between how they would and should act still displayed *some* bias, albeit less than those who reported greater expected discrepancies in how they would and should act. What this means is that whilst some individuals appear to be aware of the extent to which they are disposed to fall short of their non-prejudiced ideals there is no evidence to suggest that the belief 'I don't behave in a way discrepant with my ideals' would predict non-biased behaviour. Indeed, given the small proportion of individuals who appeared to show little discrepancy between ideals and behaviour, it is likely that we are in the portion of the population who manifest greater discrepancies. And, if we think we do not act in discrepant ways at all, it is likely we are guilty of self-deception. Thus Monteith et al (2001) emphasise that: 'we were unable to garner convincing evidence that people's self-reports of being free of discrepant responses correspond to an underlying lack of racial bias at the implicit level' (p.409).

Let's now reconsider the argument in full:

42 See also Devine et al, 1991, Prejudice With and Without Compunction *Journal of Personality and Social Psychology*, Vol. 60, No. 6, 817-830

43 Monteith, Voils, Ashburn-Nardo, (2001) Taking a Look Underground: Detecting, Interpreting and Reacting to Implicit Racial Biases, *Social Cognition* 19(4) pp395-417.

44 De Houwer et al (2009) suggest that 'salience asymmetries' – namely, differences in the extent to which the participants are familiar with some of the categories in the test (e.g. white participants may be more familiar with white names) – might account for some results on the IAT. See p.353-357.

45 13% had a small 'discrepancy score'. These scores were computed by subtracting the 'should' score from the 'would' score – these, in turn, had been assigned according to reports of how individuals should act or feel (from a set of descriptors, e.g. 'I should not feel uncomfortable about having a Black roommate'), and would act or feel (from a set of descriptions, e.g. 'I would feel uncomfortable if I was assigned a Black roommate'). The scales were constructed so that the higher the score, the greater the prejudice.

1b. Individuals can only be held responsible for their decisions and actions if they are aware of the influences of various cognitive states on those decisions or actions.
2b. Individuals are not aware of the influence on their decisions and actions of cognitive features such as implicit bias.
3b. Therefore, individuals cannot be held responsible for the manifestation of their implicit biases (the actions that are influenced by implicit biases).

I have suggested that there are good reasons to accept the argument for the conclusion that we are not responsible for the *presence* of implicit biases in our cognitive states. This is the conclusion emphasised by Saul, in her remarks that 'a person should not be blamed for an implicit bias that they are completely unaware of' (p.29). But this doesn't settle the question of responsibility. For it does not follow from this that individuals are not liable for blame for their actions which manifest implicit biases, nor that they are not responsible for being influenced by biases.

Firstly, this is because premise 2b appears to be impugnable: if some individuals are aware that there is a discrepancy between their actions and their normative expectations, then (whilst not aware of the processes that account for this) they are to a certain extent aware of their biased behaviour; that *something* is influencing their actions such that they fall short of their ideals (in the studies discussed in Monteith 2001, 37% of those who reported they were aware of discrepant responses attributed these disrepancies to negative associations). This awareness will be important in enabling individuals to undertake the long-range control strategies outlined above.

But more importantly, there is also good reason to reject premise 1b. It is not a necessary condition for responsibility that individuals are aware of the influence of certain cognitive states on their decisions and actions. One reason for rejecting this condition for responsibility is that it is unreasonably demanding, and if accepted would lead to the kind of global scepticism about responsibility that has been recently been endorsed by some. Doris (2009) has extrapolated from empirical findings on unexpected influences on behaviour that we do not display the kind of reflective self-direction arguably necessary for moral responsibility.[46] We might for various reasons think that Doris is mistaken in endorsing this scepticism. However, the important point for present purposes is simply that this kind of scepticism *cannot be avoided* if we endorse premise 1 of this argument. It entails that very many (all?) of our actions – not only automatic actions, guided by habit, but also actions guided by well executed reflective deliberation – are not ones for which we are responsible. Some may be content with this kind of sceptical conclusion. In this context, however, the argument from unawareness is working to draw a contrast between those actions for which we are responsible and those actions which, because influenced by bias, are not ones for which we are liable for blame. If we want to retain this contrast, then we should not accept premise 1 of the argument from unawareness.

That an individual lacks awareness of biases that influence her action and judgements, then, does not in itself provide reason for concluding that individuals are not responsible for being influenced by implicit biases.

2.4 Argument from lack of reasons-responsiveness

Finally, let's turn to the argument from lack of reasons-responsiveness, which can be reconstructed as follows:

1. Individuals cannot be held responsible for traits that are not responsive to reasons

46 Doris, (2009) Skepticism About Persons. *Philosophical Issues* 19 (1):57-91

2. Implicit biases are not responsive to reasons
3. Therefore, individuals cannot be held responsible for implicit biases.

Let's distinguish, once again, between:

2a. The harbouring of implicit bias is not responsive to reasons
2b. The manifestation of biases (the automatic processes by which biases operate in the production of action) are not reasons responsive.

There is not here space to do justice to the complexities of the literature on reasons-responsiveness conditions for moral responsibility. We can very roughly characterise those views as follows:

> An individual is responsible for her actions if the action issues from mental processes which are regularly responsive to reasons (including some moral reasons), where this means regularly recognising such reasons, and at least sometimes acting upon them.[47]

2.4.1 Are implicit biases reasons-responsive?

This condition for responsibility most naturally points us towards consideration of the *manifestation* of biases in action. However, it is worth noting that, insofar as the implicit associations are culturally influenced (tracking to some extent cultural stereotypes) there is reason to suppose that the processes that lead to the presence of an implicit bias in an individuals' cognitive structure are not wholly unresponsive to reason. Were there not a stereotype prevalent in the US that connects black males with guns, it is unlikely that an implicit association between 'black male' and 'guns' would show up in tests for weapons bias. This is not to say that the processes by which biases are entrenched are fully rational; any responsiveness to reasons is at best partial and limited. Indeed, being responsive to such stereotypes is being responsive to *bad* reasons. However, such a pattern of response might be sufficient for meeting reasons-responsiveness conditions: we hold people responsible for responding to bad reasons as much as for good ones. It is also worth noting again Lowery et al's (2001) suggestion that the more limited influence of biases when the subjects have contact with members of the stigmatised group indicates that such associations are sensitive to evidence – and thus might sometimes be responsive to good reasons also.

2.4.2 Are the processes that lead to the manifestation of bias reasons-responsive?

What of the processes by which bias is manifested in action – do these processes meet the reasons-responsive condition for responsibility? One might think that, insofar as the influence of implicit bias on action is automatic, and so outside of awareness, and not under direct control, it is not responsive to reason. So even if these two conditions (awareness, direct control) are not themselves necessary for responsibility (as I have argued above), they may be relevant to whether an individual's action producing processes are reasons-responsive.

However, it would be a mistake to suppose that actions which are produced by automatic processes in general are not responsive to reason. Consider the kind of automatic processes involved in the production of an excellent shot by an accomplished tennis player (cf. Arpaly, 2003, p.52). Such actions are clearly not the result of reflective deliberation on the reasons for action – and the shot would be worse were it so. Such automatic processes involved in the production of the excellent shot are highly responsive to reason – reason to move from the baseline and approach the ball, to play it with topspin and so on. So, that implicit biases function automatically does not itself entail that individuals are not responsible for being influenced by them in action.[48]

47 See Fischer & Ravizza (1999) for a full and much more detailed articulation of the view. See also Wolf, 1990.
48 See also Snow N. (2009) Virtue as Social Intelligence: An Empirically Grounded Theory, Routledge.

Moreover, in a recent overview of empirical findings on the reliance of individuals upon stereotypes, Uhlmann et al attend to cases in which the influence of stereotypic implicit associations on action is motivated by the need to enhance self-esteem or the motive to rationalise inequality.[49] (Their claim is that the reliance on stereotypes is therefore epistemically irrational.) If this is the case, then the operation of bias – its influence on action – might well be understood as meeting the reasons-responsive condition. The production of action that manifests bias might involve processes that are sensitive to certain reasons (the need to enhance self-esteem, say), and at least sometimes responsive to those reasons (by producing behaviours or judgements influenced by bias, which under-evaluate a black or female colleague, say, and so enhances self-esteem). Once again, these are *bad* reasons, and other reasons – such as reasons of accuracy, and respect, and fair attention to individual qualities – would demonstrate greater sensitivity to reasons. But that the action producing processes are sensitive to *bad* reasons, rather than good ones, is not sufficient to exempt an agent from responsibility for those actions.

Finally, one might think that being sensitive to these bad reasons is not sufficient for responsibility, if individuals are not able, at least sometimes, to be responsive to better reasons (such as reasons of respect). But some of the considerations I have detailed above, pertaining to the kinds of control that individuals may have over the manifestation of implicit bias – long range control, or indirect control – suggest that individuals at least sometimes are able to manipulate the action producing processes so as to make them *more* responsive to such good reasons, by regulating or mitigating the effects of biases.

These considerations suggest that premises 2a and 2b of the argument presented above are false; the harbouring and manifestation of bias do not always fail to meet a reasons-responsive condition. Accepting a reasons-responsiveness condition for moral responsibility, then, does not entail that individuals are not liable to blame for bias.

2.5 Summary

I have considered in some detail the arguments against holding individuals responsible either for harbouring biases, or for the manifestation of biases in action or behaviour. I have argued that the arguments from causal etiology, control, awareness and reasons-responsiveness cannot establish that individuals are not responsible for the influence of bias upon action. I argued that we should not accept these arguments – either because they rely on false empirical premises, or because they posit conditions for responsibility which are not necessary. So whilst (if we accept premise 1 of the argument from causal etiology) it might be inappropriate to regard individuals as responsible for *having* implicit biases, we cannot conclude that we ought not to regard individuals as liable for blame for *being influenced by* implicit bias.

This does not show that individuals *are* liable for blame for the manifestation of bias in action and judgement – there may yet be other conditions for responsibility that I have not considered here, and that actions influenced by implicit bias do not meet. However, the considerations picked out by philosophers as most salient to responsibility in the context of implicit bias do not support the conclusion that individuals are never responsible for being influenced by bias. Moreover, the considerations I raised with respect to long range and indirect control over biases lays the foundations for a more detailed exploration of whether such conditions are sufficient for responsibility for the manifestation of implicit bias.

I have thus addressed, in as much detail as is presently possible, the question of the truth of the claim that individuals are liable for blame for implicit bias. However, we might be left the following pressing question:

49 Uhlmann, Brescoll, Machery, (2010) The Motives Underlying Stereotype-Based Discrimination Against Members of Stigmatized Groups, *Social Justice Research* 23:1-16.

even if individuals are sometimes liable for blame for the manifestation of bias, how can we discern *when* this is the case? What use is it to maintain that people can be liable to blame for manifesting bias in their actions, if we are unable to ascertain when it would ever be appropriate to blame an individual for bias. The objection might continue: 'You say that individuals may be responsible for the influence of bias where this is a function of their explicit beliefs and values – but that this is the case can only be detected, surely, in the kind of laboratory conditions in which empirical psychologists were able to draw out of their data the statistical analyses that support this claim. And we have no ready means of doing so in our daily interactions with others. So the claim that it is false that people are not responsible for bias, and the suggestion about the conditions under which they may be liable for blame for bias, are inert and cannot be incorporated into our practices of holding each other responsible'.

This worry is an important one, and in the next section, I address it, together with the practical concern raised by Saul; namely, that holding people responsible for their biases will not help to motivate individuals to try to alleviate the presence and influence of implicit bias. In doing so, I hope to provide additional motivation for the claim that we could justifiably regard individuals as liable for blame for implicit bias, but also articulate further qualifications on when and whether it may be appropriate to do so.

3. The practice of holding responsible for implicit bias

The concern set out above asks whether it is at all useful to maintain that individuals are responsible for implicit bias. Saul presents a further challenge: rather than simply being unhelpful, it might actually be damaging to maintain that individuals are liable for blame for such biases:

> What we need is an acknowledgement that we are all likely to be implicitly biased—only this can provide the motivation for what needs to be done. If acknowledging that one is biased means declaring oneself to be one of those bad racist or sexist people, we cannot realistically expect the widespread acknowledgement that is required. Instead, we'll get defensiveness and hostility (p.22)

In this section, I address this worry, arguing that the focus on blaming is unduly narrow, and that a more comprehensive picture of our practices of holding responsible can show why this practical concern may be misguided. However, as the claims are about empirical matters, I make some proposals for future research, and elaborate on the different debates this research might fruitfully inform.

3.1. Blaming and holding responsible

We have two questions here. First, what is the point of regarding people as liable for blame if we cannot identify when they might be blameworthy for implicitly biased actions? Second, do we only have two options: deny that individuals are responsible, or maintain they are bad sexist and racist people? On this latter question, I think we should deny that these are the only two options. But Saul's concern hones in on a tendency we might well worry about: that people may tend to suppose that claiming that individuals are implicitly biased means that they are being accused of being racist and sexist. But if maintaining that individuals are liable to blame for the manifestation of such bias does not entail that they are bad racist and sexist people, then it is important to try to combat this tendency.

In order to mitigate any such tendencies, it will be important to resist any leaps from the claim that an individual harbours and is influenced by implicit bias to the claim that they are therefore racist or sexist. I think there are clear ways in which that leap can indeed be resisted. Compare the case of an individual who, despite being concerned to treat people fairly and respectfully, holds an explicitly sexist and racist belief (e.g. that women are not as good at philosophy as men, or that black men are more aggressive than white

men). Before leaping to the conclusion that she is a 'bad sexist or racist', we might say that she has gone wrong somewhere – she has a false belief, which, if she really does profess to care about treating people fairly and respectfully, she should revise. We can of course say that their beliefs are racist and sexist without thereby assuming that the person is racist or sexist, if by that we mean that the individual harbours ill will or hatred towards women and racial minorities (Garcia, 1996), or that the individual endorses a system of beliefs about the inferiority of one race or gender amounting to an ideology (Shelby, 2002).[50]

Likewise with implicit biases: before concluding that an individual who is influenced by implicit bias is sexist and racist, we might say that she has just gone wrong – some aspects of her cognitive and motivational structures are such that, if she really cares about treating people respectfully and fairly, she should work to get rid of or limit the influence of. Only where such false beliefs or negative implicit biases are maintained, rationalised and defended might we be more inclined to think that the individual involved really is sexist and racist. To suppose that an individual who has some racist belief or attitude or implicit bias is *racist* is to perform the kind of unhelpful 'categorical drift' that Blum warns against, arguing that when different failings (being a racist person, having a racist attitude) are not adequately distinguished, this serves only to 'diminish their usefulness and force as concepts expressing moral reproach' (2002, p.13). Rather, he argues, 'in the interest of accuracy and of facilitating communication about these vexing matters, we would do well to recognise such complexity' (p.29). Further, we can deny that the individual is a sexist and racist person, without denying that she is liable for blame for her sexist and racist belief or implicit bias.

It may still be that pointing out to individuals that they have false beliefs, or implicit biases, for which they are liable for blame, might sometimes lead them to suppose that they are being accused of being racist and sexist. But this is a reason to be careful about the way such claims are presented, rather than not to make such claims at all.

What about *blame* though? It is one thing to hold that, theoretically speaking, individuals can be justifiably held responsible and regarded as blameworthy for their implicit biases, and quite another thing to in fact blame them for being biased. This brings us to some of the issues raised by the first concern: that given our epistemic position with respect to whether and why individuals are influenced by implicit bias, we ought not to in fact blame one another for manifesting implicit bias.

I think that we can agree with these remarks about the difficulties of knowing when an individual is blameworthy for being biased, and the impropriety, therefore, of in fact blaming each other, whilst nonetheless maintaining that it is important to regard individuals as responsible (and so potentially liable to blame) for implicit bias. To suppose that the only purpose of regarding each other as liable for blame is so that we are in a position to blame each other when required is to take an unduly narrow view on the point of our practices of responsibility.

Here are two aspects of our practices of regarding each other as responsible that are overlooked if we focus only on blaming and on identifying people as 'bad', and which show the importance of emphasising liability for blame even in the absence of clarity about when to apportion blame:

3.1.1 Holding ourselves responsible

We might not feel warranted in actually blaming others for being influenced by implicit biases. But this does not mean that we are not able to blame ourselves for being influenced by such biases. As mentioned above, some evidence indicates that individuals are aware of the discrepancies between how they *should* act and how they *would* (given the influence of implicit biases) act. With this awareness, individuals are well placed

50 J. L. A. Garcia (1996). The Heart of Racism. *Journal of Social Philosophy* 27 (1):5-46;Tommie Shelby (2002). Is Racism in the "Heart"? Journal of Social Philosophy 33 (3):411–420.

to blame themselves for being influenced by bias, whilst it might be inappropriate for others to do so - perhaps because others do not or cannot know the extent to which my biased behaviour results from my attitudes towards treating others non-prejudicially, or whether I have done enough to activate my goal of treating others non-prejudicially.

There is some evidence which suggests that holding oneself responsible might be particularly effective a way to mitigate the effects of implicit bias. Amodio et al (2007) present evidence that suggests that guilt is a particularly useful affective response in regulating implicit bias.[51] And we have seen Mozkowitz and Li's (2011) claims to the effect that focusing on one's failures to live up to ideals can activate the goal of treating others non-prejudicially, thereby inhibiting bias related goals. If this is right, then holding oneself responsible, and in particular feeling guilt when one fails to live up to one's non-prejudicial ideals, could play an important role in responding to implicit biases.

3.1.2 Shaping expectations and behaviour

Facts about whether or not a certain activity or behaviour is one for which individuals are held responsible and therefore are liable to blame, can alter norms and expectations about how individuals ought to behave, even if we rarely in fact identify individuals who are blameworthy for so acting. Classifying certain actions as prohibited, for which individuals are liable to blame, can have numerous important effects, including: strengthening norms against so acting; encouraging individuals to self-monitor; leading us to change our expectations of the steps others might take in monitoring their own behaviour. These changes are all important corollaries of regarding some form of behaviour as something for which individuals are properly held responsible, and for which they are liable to blame. But note that they do not depend upon us being able to in fact engage in blaming, although some of them might encourage us to challenge others' decisions and provide careful justification for them.

Of course, these practices do not only follow from regarding a certain kind of behaviour as appropriately within the remit of responsible action; but insofar as they are bound up with our practices of holding responsible, attending to these considerations helps us to see why it can be important to regard ourselves and others as responsible for being influenced by implicit bias, even if we think that we will rarely be in an epistemic position to in fact blame another for being influenced by implicit bias.

3.2 Further empirical work

I have in this final section tried to suggest that there might be ways of addressing the concerns about the practical efficacy of regarding individuals as liable for blame for being influenced by implicit bias. However, whether holding each other responsible, and blameworthy, for being influenced by bias is likely to have an effect on the extent to which individuals in future manifest bias is in large part an empirical matter.[52] Empirical investigation into the effects of whether individuals are more or less likely to display implicit bias, or more or less motivated to mitigate the influence of implicit bias, when told they are responsible for being influenced by such biases, could shed light on how best to structure our interactions with each other in addressing the influence of bias. One reason for which one might worry about the effects of holding each other responsible is that some empirical data indicates that individuals tend to respond badly to negative feedback (Blair et al 2002, pp.244-247). However, it is not clear that blame can be understood

51 Amodio, Devine, and Harmon-Jones (2007) 'A Dynamic Model of Guilt: Implications for Motivation and Self-Regulation in the Context of Prejudice, *Psychological Science*, vol 18 (6).

52 Note also that whether treating each other as liable to blame inhibits the manifestation of bias is a separate question from whether doing so is efficacious in getting people to take steps in e.g. introducing policies to limit the possible influence of bias (anonymising reviews), or taking other practical steps. Saul reports (in conversation) that the strategy of emphasising non-responsibility has been important in bringing about change. We could agree with this whilst insisting that whether treating each other as responsible (or not) affects the manifestation of bias directly (rather than by bias-blocking policies) is an empirical question not yet answered.

straightforwardly as a form of negative feedback. And the studies that support the hypothesis that guilt has an important role in inhibiting the influence of implicit bias (Amodio et al 2007) might lead us to suppose that blaming could have *some* constructive role in mitigating the influence of bias. Further empirical study might help us both in addressing our responses to bias, but additionally in helping us to consider whether and how blame differs from other forms of negative feedback.

3.3 Implications for philosophical methodology: the heterogeneity of 'implicit biases'

My discussion throughout has for the most part focused on negative implicit racial biases. Much of the experimental work I have considered has pointed to the ways in which the manifestation of these implicit biases is related to individuals' beliefs and values, might be something of which individuals can be aware, and can be seen to be, in a limited and defective way, reasons-responsive. Before concluding, it is worth noting that many of these findings about the effects of attitudes on implicit race biases do not seem to generalise to all kinds of implicit biases. Regarding the relationship between implicit biases and explicit values and beliefs in particular, Banaji & Hardin (1996) found no difference in the implicit biases regarding gender with respect to individuals who, on self-reports, measured high or low in sexist beliefs (p.139).[53] One feature worth noting in relation to their study, however, is that the associations involved tested for speed of association between gendered primes (nurse, mechanic) and recognition of pronouns (he, she). It may be that individuals who measured low on sexism share these strong gender stereotypical associations because of their familiarity with, if not endorsement of, these stereotypes. Because they are not inherently negative or evaluative, we might not expect to see the same inhibitory mechanisms at work in low-sexism subjects as in the low-prejudice subjects in the race/negative association studies.

This raises important issues about the extent to which we can talk about 'implicit biases' per se, as philosophers have tended to do. This term, rather, should be considered to cover a heterogeneous set of cognitive associations, including negative evaluative or affective associations, semantic associations, negative stereotypes, and neutral stereotypes. When talking about implicit biases, and when making recommendations about how to alleviate or inhibit the influence of these associations on action, we ought to be alert to the particular kind of bias at issue, and be cautious in generalising from claims about the operation of other kinds of biases, which may not be regulated or manipulated in the same ways.

4. Concluding remarks

I have argued that we should reject the arguments for the claim that individuals are not responsible for being influenced by implicit biases. Individuals might sometimes meet sufficient conditions for responsibility, when they have long range control and so can take responsibility for mitigating implicit biases; or when they are blameworthy given the indirect influence, via reflective level beliefs and attitudes, over whether their actions manifest implicit bias. In considering in more detail the relevant empirical findings, we have been able to draw out some methodological implications for how philosophers might deal with the heterogeneity of 'implicit biases'. Finally, the practical considerations taken to speak against holding each other responsible take an unduly narrow view both on what it is to regard someone as liable for blame, and on the nature of our practices of holding each other, and ourselves, responsible. But discerning whether treating each other as responsible is or is not an effective means to mitigating implicit bias is something that could be tested in future empirical research.[54]

53 Banaji, M., & Hardin, C. (1996). Automatic stereotyping. *Psychological Science, 7,* 136-141.

54 This paper has greatly benefited from fruitful discussions and illuminating feedback from Jenny Saul, Komarine Romdehn-Romluc, Dan Kelly, Joseph Sweetman, Tom Stafford, Clea Rees, Jonathan Webber, Peter Kirwin, the editors of this volume and an anonymous reviewer for JSP.