A Virtue Ethics Response to Implicit Bias

Clea F. Rees

Implicit Bias and Philosophy, Vol. 2: *Moral Responsibility, Structural Injustice, and Ethics.*
Ed. Michael Brownstein and Jennifer Saul. Oxford: OUP. 2016.

Abstract

Virtue ethics faces two challenges based in 'dual-process' models of cognition. The classic situationist worry is that we just do not have reliable motivations at all. One promising response invokes an alternative model of cognition which can accommodate evidence cited in support of dual-process models without positing distinct systems for automatic and deliberative processing. The approach appeals to the potential of automatization to habituate virtuous motivations. This response is threatened by implicit bias which raises the worry that we cannot avoid habituating reliably vicious motivations. I argue that the alternative model of cognition also offers the virtue ethicist a promising response to this second challenge. In particular, the virtue ethicist can respond to the implicitly biased by counselling the habituation of egalitarian virtue, rather than merely the control of anti-egalitarian vice. Research shows both the importance of automatized individual egalitarian commitments and the potential of habituation to automatize deliberatively endorsed egalitarian goals. However, individuals' ability to sustain and implement their commitments depends crucially on hospitable environments. Communities which themselves embody egalitarian values and which encourage and support their members' egalitarian commitments are therefore essential. As Aristotle said, individual virtue requires a virtuous community.

Virtue ethics faces two challenges based on psychologists' work on the role of automatic processes in cognition. Both arise from our reliance on cognitive processing which is relatively immune to direct deliberative control and of which we are relatively unaware. These challenges threaten not only the very possibility of virtue but, more fundamentally, our conception of ourselves as rational persons (e.g. Doris 2009).

The first is the classic situationist challenge which suggests that much of our behaviour is determined by trivial and arbitrary features of situations of which we are unaware and which we would not endorse as reasons for action (e.g. Merritt, Doris and Harman 2010). Virtuous action is action done for the right reasons; behaviour cannot be virtuous if it is not motivated by reasons at all. Since virtue ethics is intended to guide the moral lives of creatures like us, this first challenge threatens not only the possibility of our realizing virtue, but the appeal of virtue ethics *qua* ethical theory.

One of the most promising responses to this first challenge argues that the influence of automatic processes on cognition facilitates, rather than threatens, rational agency and virtuous action because habituation can ensure that automatized cognition embodies the right motivations (e.g. Snow 2009; Rees and Webber 2014).

The second challenge threatens this response by suggesting that the influence of virtuous automaticity on cognition will be systematically undermined by our unwitting habituation of the wrong motivations. Virtue requires not only the habituation of virtuous motivations but the non-habituation or dehabituation of vicious ones which would otherwise undermine the connection between virtuous motivation and virtuous action. Implicit bias is an especially stark illustration of this second challenge: research suggests that we may be oblivious to the existence and behavioural influence of disturbing features of *ourselves* in the form of habituated associative biases which we have explicit reasons to reject (see the Introduction in Volume 1). While it might be disconcerting to discover a disproportionate number of aspiring **Phil**osophers named '**Phil**ippa' and '**Phil**lip' (Pelham, Mirenberg and Jones 2002), that our implicit sexism might frustrate the aspirations of the former is positively disturbing. Moreover, whereas more troubling situationist results such as Milgram's depended on carefully engineered experimental manipulations (Russell 2011), the threat to virtue posed by implicit bias requires only the reality of social prejudice. Although virtue ethicists recognize the crucial role of social support in developing and sustaining virtue, because virtue ethics purports to provide practical guidance, the pervasiveness of implicit bias rules out simply dismissing it as the product of a bad environment. Given that implicit bias occurs not only outside conscious awareness, but despite deliberative abhorrence, what counsel can the virtue ethicist possibly offer the implicitly biased?

Responses to the situationist challenge which appeal to virtuous habituation invoke a model of cognitive processing based on two areas of psychological research. The first is social psychologists' work on attitudes and attitude change in the context of an associative model of personality. The second is research concerning the automatization of goals.

In this chapter, I argue that this model also offers the virtue ethicist a promising response to the second challenge, because automatization has the potential not only to habituate virtuous motivations, but to dehabituate vicious ones. In particular, the virtue ethicist can respond to the implicitly biased by counselling the habituation of egalitarian virtue, rather than merely the control of anti-egalitarian vice. Specifically, I argue that the habituation of individual egalitarian commitments is crucial to strategies of active resistance and that communities should ensure the collective support this process requires.

Section 1 outlines dual-process models of cognition and the particular role of those which posit distinct systems for automatic and deliberative processing in accounts of the threat posed to virtue ethics by implicit bias. Section 2 explains why indirect mitigation strategies offer virtue ethicists an unsatisfying response to implicit bias given such models of cognition. Section 3 sketches the psychological structure of attitudes, as understood by social psychologists, and their role in an alternative model of cognitive processing. Section 4 explores the process by which consciously selected goals and commitments may be automatized, and explains its particular interest. Section 5 argues that a satisfactory defence of virtue ethics is supported by research on the automatization of strong egalitarian commitments. Section 6 explores some puzzling questions raised by current research and indicates how future work might seek to address them. Section 7 explores the potential of egalitarian commitments to alter our implicit biases themselves and explains some further limitations of current research. Section 8 explains an important implication of my argument for virtue ethics: individuals can effectively habituate egalitarian virtue only if their communities share their commitment to resisting implicit bias.

1. 'Dual-Process' Models and Implicit Bias

Psychologists have developed several 'dual-process' theories of cognition (Maio and Haddock 2010, 96–106). For the purposes of this chapter, what is important about such models is their common claim that deliberative and automatic cognition are distinct kinds of cognitive processing. Deliberation is explicit, conscious processing which analyses information carefully and logically, is sensitive to the content and strength of arguments, and is relatively slow and effortful. In contrast, automatic processing depends on heuristics and learnt associations, is more sensitive to the source and form of arguments, and is relatively fast and effortless. Whereas deliberation might lead you to choose unbranded paint for your home (because it was cheaper) or branded (because it was higher quality), associative processing might result in a choice of Dulux (because you liked the dog in their advertising). Whereas cost and quality comparisons require conscious attention, you might be unaware of the canine influence on your décor.

Humans could not make do with only deliberative processing; automatic cognition is essential. Moreover, bias in the broadest sense provides crucial filtering enabling us to focus limited cognitive resources on what is of greatest importance to us. At its best, implicit bias attunes parents to the particular cries and needs of their own children, allows surgeons to focus on critical features of the body in front of them, and enables examiners to assign marks informed by the features of essays of greatest disciplinary relevance. In the complete absence of such bias, every mother in the maternity ward would need to consider every cry in order to decide whether to respond to it, the surgeon would require attentional resources to ignore your appendix when removing your tonsils, and examiners would need to consciously set aside students' choice of ink as irrelevant to their knowledge of Kant's metaphysics.

Our capacity for conscious, effortful cognition is limited. Attention focused on one task cannot be simultaneously devoted to others and the expenditure of volitional resources affects their subsequent availability (Baumeister et al. 1998; Muraven, Tice and Baumeister 1998; Muraven and Baumeister 2000). Time is another limited resource. Even if the dog is irrelevant to the quality and value of Dulux paint, the canine association may facilitate a perfectly rational choice. Although appealing to the dog would undermine the rationality of a deliberative decision to buy Dulux, his appeal need not undermine the rationality of a less considered choice if the costs of more careful deliberation would outweigh the benefits of a more considered one. It can be quite irrational to expend the resources required to reach a more rational decision. Moreover, we are constrained not only by the total time available to us for all tasks, but by the time-sensitive nature of many decisions.

In general, then, it is no bad thing that we rely on associative processing and heuristic short-cuts. Unlike our relatively innocuous paint purchases, however, other learnt associations are far from harmless. 'Implicit bias' in the problematic sense refers to biases we soak up from our social environment in the form of implicit morally problematic associations with characteristics such as race, sex and sexual orientation (see introduction to volume 1). These problematic implicit biases influence cognition in ways which systematically disfavour members of non-dominant groups. Because such biases are systematic rather than arbitrary, the collective impact of individuals' implicit biases on members of non-dominant groups is likely to constitute a significant harm even when the impact of each instance is negligible (Brennan 2009). Moreover, some instances will themselves constitute significant harm. If simulated decision-making provides a reasonable indication of its effects, implicit bias reduces the chances one will be interviewed for a job if one is Arabic rather than Swedish or hired to a managerial post if one is female rather than male, and makes it more likely that one will be shot by an armed police officer ('shooter bias') and less likely that one will receive appropriate treatment for coronary heart disease if one is black rather than white (Jost et al. 2009).

The worry is that because our implicit biases are acquired and utilized outside conscious awareness, we cannot perceive or correct for their effects. Dual-process theories which explain automatic and deliberative processing by invoking distinct, relatively independent systems of cognition deepen this concern by suggesting that even educating ourselves about the problem might leave us unable to alter or control our implicit biases (introduction to volume 1).

However depressing this evidence might be from the perspective of policy makers and concerned citizens, one might think it not altogether bad news for virtue ethicists. After all, the evidence for their effects on decision-making depends on variation in the kind and degree of individuals' implicit biases. Just as one response to the situationist challenge emphasizes the rarity of virtue (e.g. Kamtekar 2004), one might argue that the prevalence of implicit bias is simply further evidence of widespread ethical deficiency.

This response is ruled out, however, by a key tenet of virtue ethics: virtue can be developed. It is true that the right sort of early education may be essential: one may not be blameworthy for one's lack of virtue if one was deprived of appropriate habituation as a child. This might be mere common sense except for the failure of much moral philosophy to acknowledge it. That virtue is as much a collective responsibility as an individual one and that the development of moral agency requires an appropriately supportive social context will come as no surprise to feminist philosophers, educators and parents (e.g. Baier 1995).

In the case of implicit bias, however, it is not at all clear what the 'right sort' of education might be. Given that implicit biases are found even in individuals engaged in efforts to actively resist prejudice, it is unclear not only how such individuals should respond to their own implicit bias, but also what they might do to reduce it in the next generation. Implicit bias is not limited to unfortunates attempting to overcome the effects of explicit encouragement to prejudice.

Moreover, most virtue ethicists argue that even a poor start can be mitigated or overcome by later efforts. Scrooge's decision to reform begins a process of rehabituation which replaces miserliness and meanness with generosity and compassion (Annas 2011, 12). Partly because Scrooge is a self-conscious miser who despises kindness in others, reflective deliberation can instigate and guide self-reform. In contrast, given that implicit bias can occur not only outside conscious awareness but despite deliberative abhorrence, what counsel can the virtue ethicist possibly offer the implicitly biased?

2. Stocking the Egalitarian's Toolbox

Indirect mitigation strategies have proven effective in combating the behavioural effects of implicit bias. The virtue ethicist might therefore recommend that individuals and institutions respond by implementing these strategies themselves, raising awareness, and encouraging others to follow their lead.

First, institutions can select from a range of mitigation strategies. For example, the representation of female musicians in top orchestras improved partly due to the introduction of screens rendering candidates audible but invisible during auditions (Goldin and Rouse 2000). Similarly, there is some evidence for a reduction in gender bias on referees' judgements when journals implement double-blind reviewing (Peters and Ceci 1982; Budden et al. 2008; but cf. Blank 1991). Ensuring that decision-makers anticipate needing to justify their decisions to an audience whose views they cannot predict can encourage more thorough scrutiny of the relevant considerations, more careful analysis of the pros and cons of various options, and reduced reliance on the automatized associations which constitute implicit bias (Lerner and Tetlock 1999, 256–8, 263). Even when the views of the anticipated audience are known, accountability may be effective in inducing more careful analysis if deliberators are motivated to base their decisions on accurate evidential evaluations (Quinn and Schlenker 2002). Although the conditions under which accountability is effective matter in reducing the effects of implicit bias on decision-making (Lerner and Tetlock 1999, 258–9, 264–6), this need not undermine its effectiveness in a range of key cases. For example, while monitoring perceived as illegitimate can actually increase the effects of bias, the legitimacy of a requirement to justify personnel or prosecutorial decisions is unlikely to be doubted.

Second, in addition to encouraging and implementing appropriate institutional practices, a number of mitigation strategies are available to individuals. Envisaging or imagining counterstereotypic exemplars, or thinking oneself into others' shoes can help to overcome the effects of implicit bias on cognitive processing (Corcoran, Hundhammer and Mussweiler 2009; Dasgupta and Greenwald 2001; Dasgupta and Asgari 2004; Blair, Ma and Lenton 2001; Galinsky and Moskowitz 2000). Forming 'implementation intentions' is another way for individuals to neutralize the influence of their implicit biases on behaviour (Webb, Sheeran and Pepper 2012). An implementation intention is a specific behavioural plan as opposed to a more general goal. 'I will study harder' is a general commitment; 'If it is 3 o'clock on a Tuesday, I will study in the library!' is an implementation intention.

Strategies which allow us to indirectly mitigate the effects of bias on our treatment and judgements of others have attracted considerable attention. Theoretical work in philosophy has recommended considering the availability and likely effectiveness of these strategies when choosing between competing normative ideals concerning racial categorization (e.g. Kelly, Machery and Mallon 2010), and leveraging them to satisfy epistemic and moral demands (e.g. Merritt 2009; Kelly and Roedder 2008). Saul has argued they should inform the philosophy REF (Research Excellence Framework) in the UK and the Philosophical Gourmet Report in the US (2012). Furthermore, institutions have begun to actively promote their use. The US National Center for State Courts has produced educational materials encouraging their use to address the effects of implicit bias on judicial decision-making (Casey et al. 2012). In the UK, the Chair of the REF Philosophy Panel has responded to Saul's concerns by ensuring that members are aware of the literature on implicit bias and of ways to reduce its impact, the Equality Challenge Unit is developing strategies to counteract its influence on recruitment decisions in higher education institutions, and Remploy offers practical ways to mitigate its effects on individuals with facial disfigurements (Saul 2012, 263–4; *Equalitylink May 2013* 2013; *Changing Perceptions with Changing Faces* 2013).

There are good reasons for these recommendations: indirect mitigation strategies are crucial if only because no momentary act of will can eradicate implicit bias. Using such strategies enables us to mitigate the behavioural effects of our biases in especially sensitive or significant situations, especially ones of which we are aware and for which we can prepare in advance. Considered in isolation, however, the solution which such strategies promise the virtue ethicist seems neither psychologically nor theoretically satisfying because it apparently offers us little hope of changing our implicit biases themselves. If automatized and deliberative processes involve distinct cognitive systems, then there is no obvious way for strategies which rely on deliberative control to alter the implicit associations whose influence they mitigate.

First, committed egalitarians who share virtue ethicists' concern with character are unlikely to find the solution psychologically satisfying. Although I would much prefer that envisaging counterstereotypical exemplars prevent my biased associations from leading me to treat a short, blind, black, female resident of Merthyr Tydfil less well than a tall, able-bodied, white, male inhabitant of Ascot, I would prefer to alter my underlying bias itself. Indeed, I would wish to be free from implicit bias even if its behavioural consequences were entirely benign. Moreover, the effectiveness of indirect mitigation strategies is limited by our epistemic and cognitive capacities. Situations arise for which nobody can be fully prepared and the number of implementation intentions one can usefully form is presumably limited.

Second, while she should surely encourage their use by both institutions and individuals, indirect mitigation strategies appear to offer the virtue ethicist at most rather cold theoretical comfort. It is not sufficient for virtue that one reflectively endorse the right values and that one's behaviour reflect those values. What matters also is that one's habituated, automatized motivations embody them. This is the grain of truth in the myth that true virtue is effortless: alleviating another's distress may require considerable effort, but being moved to do so should not. A need to rely on indirect mitigation strategies to control the effects of one's implicit biases shows that one cannot rely on one's habituated, automatized motivations and thus reflects a deficiency in virtue. Dependence on such strategies shows that vicious automaticity would otherwise interfere with the influence of virtuous automaticity on cognition. Indirect mitigation strategies cannot restore virtue if they are limited to controlling, rather than eliminating, vicious motivation. Unless the virtue ethicist can say something about how control can be habituated and implicit bias itself reduced or eliminated, she is limited to counselling the control of vice rather than the development of virtue.

Fortunately, current psychological research offers a better defence of virtue ethics based on an alternative to dual-process models which posit distinct systems for automatic and deliberative cognition. While the existing literature does not guarantee the viability of virtue ethics, it does provide grounds for cautious optimism. The virtue ethicist should therefore resist the idea that damage limitation exhausts our capacity for control. The alternative model of cognition suggests that indirect mitigation strategies aimed at short-term behavioural control may gradually reduce the underlying implicit biases themselves. Moreover, the most important items in our egalitarian toolboxes should be positive strategies aimed directly at enabling us to inhibit the cognitive effects of our biased associations rather than merely their behavioural influence. The ultimate aim should be weakening or eradicating implicit bias from response-directed processing.

Why think that automatized control is sufficient for virtue when indirect deliberative control is not? Does the habituated control of implicit bias amount to anything more than an especially effective form of mere continence? One might argue that even eliminating the influence of implicit bias on response-directed processing would be insufficient because virtue requires that the cognitive system be entirely free from such bias. However, this objection depends on an overly simplistic picture of virtue. First, as I argue in section 7, biased associations are necessary to an understanding of one's social world so long as that world is itself characterized by bias. The virtuous person cannot be altogether free from implicitly biased associations because such associations play a crucial role in understanding social situations. For example, appreciating the offensiveness of superficially complimentary remarks often depends on understanding the stereotypes they invoke. Moreover, sensitivity to social bias must be automatized if it is to guide social interactions effectively in real time. Far from being inconsistent with virtue, therefore, implicitly biased associations are crucial to the social understanding virtue requires. To deny this would commit one to the view that virtue requires a virtuous world in the strong sense that nothing would count as being virtuous in a world characterized by prejudice. Although I argue in section 8 that egalitarian virtue requires a community which shares one's egalitarian commitments, I take this to be an empirical claim concerning human psychology rather than a conceptual point about the requirements of virtue.

Second, the traditional understanding of 'continence' seems better captured in terms of dependence on a particular kind of control than on control *per se*. The difference between the virtuous person and one who is merely continent is importantly connected with the thought that virtuous motivation is effortless. Whereas the continent person must actively resist the temptation to steal, for example, 'the thought of stealing never enters the honest person's head'. This distinction is perfectly consistent with automatized control outside the honest person's conscious awareness. In general, while the virtue ethicist may be concerned with the architecture of character traits at the personal level, whether conscious or not, it is not clear why she should be committed to their having any particular structure at the sub-personal one.

Before turning to the defence of virtue ethics, I need to introduce the alternative model of cognition on which it depends. Section 3 explains this model in the context of psychological research on attitudes and section 4 outlines work on goal automaticity.

3. Attitudes in a Cognitive-Affective Personality System

The case for cautious optimism appeals to social psychologists' work on attitudes; dynamic, associative models of personality; and cognitive-affective processing. Although I cannot do justice to the literature here, this section highlights the most relevant aspects of the overall picture which emerges for the purpose of this chapter. In particular, the associative model of cognitive-affective personality can accommodate data cited in support of dual-process theory without the need to postulate separate systems for automatic and deliberative processing. Unlike models which posit distinct systems, therefore, this alternative can straightforwardly accommodate evidence that deliberation influences automatic cognition.

Maio and Haddock explain the social psychologist's conception of attitudes as complex, structured evaluations of objects with cognitive, affective and behavioural components which have functional roles in a person's psychology (2010). Attitudes are associative clusters of mental items which differ in content and strength. Their objects may be as particular and concrete as a drip of candle wax or as general and abstract as universal justice.

The *content* of one's attitude towards an object is a function of cognitive elements one associates with it such as a belief that woollen jumpers are difficult to wash; associated affective elements such as a fear of sheep; and associations with past behaviours such as the memory that one preferred wool to acrylic last time one bought a jumper. This last, behavioural factor is not so much a 'component' of the attitude as philosophers might understand it, but rather a trigger for attitude formation. In the absence of an existing, accessible attitude towards woollen jumpers, I may infer a positive attitude from my awareness of my past purchasing decisions. As work on cognitive dissonance shows, I may also alter an existing attitude as a result of attitude-incongruent behaviour (Cooper 2007). For example, my purchase of one might lead me to adjust a negative attitude towards woollen jumpers. It is important that this reduction in negativity is a change in the *content* of the attitude and not, as philosophers might be inclined to say, in its strength. How much I like or dislike an attitude object is part of the content of that attitude.

In the context of a dynamic, associative model of personality, an attitude's *strength* is a matter of the strength of the connections between its components and with other elements in the cognitive-affective personality system (CAPS), situational features, behaviours and so on. The strength of the connection between two components is a matter of how readily each affects the other's influence on cognition. As Mischel and Shoda explain it, the personality system involves five general types of 'cognitive-affective unit' (1995). First, people classify features of internal and external experience using categories such as 'philosophy' and 'penguin'. Second, individuals have beliefs about themselves and their worlds such as 'I am going to mess up this job interview' and 'oil-soaked penguins need woollen jumpers'. Third, individuals' experience is affectively laden with such things as sympathy and claustrophobia. Fourth, people value aspects of their worlds such as patience and penguins. Fifth, individuals have plans and strategies such as intentions to assuage feelings of disappointment by thinking positively and to respond to the next oil spill by knitting penguin-sized woollen jumpers.

The various cognitive-affective units in the CAPS model are part of a connectionist network which processes cognitive and affective information and which is itself modified by that processing. Internal and external inputs such as the memory of rescue workers appealing for penguin-sized woollen jumpers or the sudden discovery of an ambiguous figure slumped on the corner of Miskin Street affect the flow of information across the network in two ways. First, they induce processing aimed at an immediate response such as intending to purchase wool or dialling 999. Second, this processing strengthens network connections between activated components.[*]

Stronger attitudes are more *accessible* in the sense that they are more likely to significantly affect cognitive-affective processing, intention formation and behaviour. Attitudes are strengthened and made more accessible by activation. The more often an attitude influences cognition, the stronger the associations between its components and the stronger the connections between those components and triggering internal and external elements. Accessibility in this sense need not be conscious. Processing units can be triggered *automatically* by external and internal stimuli, feedback and associations. Processing can take place consciously or non-consciously, with or without an agent's awareness. That is, the associative model of cognitive-affective personality can accommodate data cited in support of dual-process theory without postulating distinct systems for automatized and deliberative cognitive processing. This is important because it allows the model to straightforwardly accommodate evidence for the influence of consciously endorsed commitments and deliberation on automatized cognition. For example, the associative model can more easily explain why forming an implementation intention to associate women with science or Muslims with peace especially quickly reduces bias on even implicit measures (Webb, Sheeran and Pepper 2012).

[*] The best way to understand this system is to study Mischel and Shoda's diagram (1995, 253).

4. Goals, Attitudes & Automaticity

Work on goal automaticity provides further evidence for the influence of deliberative cognition on automatized processing. Goals and commitments which are initially chosen as the result of conscious deliberation may become *automatized* if repeatedly invoked in cognitive processing, or they may result from an entirely automatic process. As with attitudes, a goal's accessibility is a matter of how readily it influences cognition, and goals are strengthened and made more accessible by activation (Bargh and Williams 2006, 2). Like attitudes, goals are understood as located in an associative cognitive system which encompasses automatized perceptual sensitivities, proactive as well as reactive goals and motives, affective processing and more (Bargh 1989, 1990, 2006, 147–148; Bargh, Gollwitzer et al. 2001, 1014; Isen and Diamond 1989).

The ability of complex, abstract goals to guide cognition automatically demonstrates the potential intelligence of automaticity. For example, temporarily raising the accessibility of the goal of cooperation outside conscious awareness caused subjects to behave as cooperatively as those explicitly asked to cooperate and significantly more cooperatively than controls (Bargh, Gollwitzer et al. 2001). Bargh's work has explored the automatic activation of, and behavioural guidance by, 'higher-order goals and motives' relevant to social interaction such as commitments to truth, justice and 'being a good mother, a high achiever, or a moral person' (Bargh 1990, 103–104, 118; Bargh and Gollwitzer 1994, 79).

Goals guide by associating features of situations with flexible and intelligent response strategies. As they are repeatedly activated, these associations become automatized. If one consciously selects the goal of cooperation sufficiently often in response to tensions over a shared resource, one will gradually associate such tensions with this response. That is, one's goal will initiate cooperation in response to such tensions without the need for conscious deliberation. The response is flexible since it must be sensitive to the details of particular cases, avoiding not only trampling others' interests, but blocking others' attempts to trample one's own. The response is intelligent since it embodies one's reflective judgement about the best way to navigate a tricky aspect of one's social world.

Although both goals and attitudes can be automatized through habituation, they differ in their relation to acts of volition: goals, but not attitudes, are potential objects of deliberative choice. Although one cannot decide to automate a goal, one can decide to consciously adopt it, potentially beginning the process of automatization if conditions are right (Bargh and Williams 2006, 2). This volitional distinction is of crucial importance to both individuals concerned about implicit bias and virtue ethicists. Since we cannot generally choose to like or dislike something by a mere act of will, even our explicit attitudes lie largely outside our direct control. If I am fond of penguins, I cannot just decide to dislike them, even though I could try to change my attitude indirectly by researching their less endearing habits. In particular, even our conscious associations are largely outside direct deliberative control. I cannot just decide to eliminate the association between penguins and winter festivities from my cognitive processing system. Since we have little direct control over even associations of which we are fully aware, there is likely to be little point in trying to eradicate implicit associations directly. Trying to 'will away' our implicit biases — or urging others to do so — is likely to be pointless at best and counterproductive at worst. In contrast, adopting a goal or making a commitment is precisely the sort of thing that acts of volition are good for.

In section 5, I examine the effects of enduring, automatized egalitarian commitments on the expression of implicit bias and argue that the automatization of egalitarian goals has a key role to play in responses to implicit bias. While the familiarity of failed resolutions is indicative of the difficulties people experience in following through on their commitments, the effectiveness of implementation intentions suggests that psychological research could guide the selection of more successful strategies. Although they are too specific to fully capture the content of most commitments, implementation intentions might be incorporated into an effective overall strategy of goal pursuit. Further research on attitudes, goals and specific cognitive strategies should enable us to better understand how to effectively habituate and maintain our commitments, enabling individuals to resist threats to goal pursuit and helping communities to encourage and sustain their commitments to egalitarianism.

5. Automatizing the Egalitarian's Toolbox

Since automatization systematically tunes the cognitive-affective processing system to reflect deliberatively endorsed values and commitments rather than working around its deficiencies, concerned individuals and virtue ethicists have good reason to be interested in the automatization of egalitarian commitments. The process of automatizing the goal of treating people fairly, for example, is precisely aimed at sensitizing the system to the right reasons and desensitizing it to the wrong ones. This is just the kind of habituation required for the development of virtue. In this section, I focus on the habituation of egalitarian virtue. In section 7, I explain why Mischel and Shoda's associative model of personality suggests that the habituation of egalitarian virtue should also dehabituate anti-egalitarian vice by decreasing implicit biases themselves.

I argue that the virtue ethicist can respond to the challenge of implicit bias by counselling the implicitly biased to habituate egalitarian virtues by adopting and pursuing egalitarian commitments. I develop this response by introducing two research programmes concerned with the effectiveness of such commitments. This research supports two claims: first, consciously chosen egalitarian commitments can be automatized; second, habituated egalitarian motivations can effectively guide automatic cognition.

Just as implicit and explicit bias are distinguished by the measures used to assess them (introduction to volume 1), so with implicit and explicit egalitarian commitments. The first research programme I discuss concerns the differential effectiveness of different explicit motivations to avoid prejudice. The second concerns the effectiveness of implicit egalitarian commitments. In section 6 I explore the differences between the psychological constructs posited by each programme, and explain why attempting to understand the automatization of egalitarian commitments in the light of both raises some puzzling questions.

I begin by outlining the different effects of two distinct kinds of explicit egalitarian commitment on expressions of prejudice. Individuals who are personally committed to not being prejudiced, as opposed to wishing to avoid appearing prejudiced, effectively avoid expressing prejudice even in ways which elude conscious control. I then outline evidence that the ability of such individuals to inhibit the influence of implicit stereotypes on cognition depends on automatization. This ability is especially significant because implicit stereotyping is more resistant to amelioration than other forms of implicit bias (Amodio, Devine and Harmon-Jones 2008, 63).

Plant and Devine's Internal and External Motivation to Respond Without Prejudice scales assess individual differences in kind and degree of egalitarian motivation (1998). 'External' motivation stems from a concern with self-presentation: the individual wishes to avoid others' disapproval of prejudiced behaviour (EMS). 'Internal' motivation stems from a concern about prejudice itself: the individual's values are inconsistent with prejudice and not being prejudiced is considered personally important (IMS). High-IMS individuals are motivated to avoid prejudice even when unobserved and their egalitarian commitments are relatively consistent across different situations (Amodio, Devine and Harmon-Jones 2008, 61). In contrast, low-IMS high-EMS individuals are motivated to respond without prejudice only in public scenarios, while low-IMS low-EMS individuals are not concerned to avoid expressions of prejudice at all.

What about differences *among* high-IMS individuals? One might think that high-IMS high-EMS individuals would demonstrate the least bias of all groups since they have not one, but two, sources of motivation. In fact, however, high-IMS low-EMS individuals show the least bias. Although relative to low-IMS individuals, all high-IMS individuals show similarly reduced bias in responses subject to deliberative control, only those low-EMS high-IMS demonstrate less bias on relatively uncontrollable implicit measures (Devine et al. 2002; Amodio, Devine and Harmon-Jones 2003).

Why should additional egalitarian motivation undermine individuals' efforts to control prejudice? Devine et al. suggest two possible explanations (2002, 846). First, high-IMS low-EMS individuals might never have acquired implicit bias whereas high-IMS high-EMS individuals might be trying to overcome biased response patterns. Second, high-IMS high-EMS individuals might be at an earlier stage in a process of overcoming bias than high-IMS low-EMS individuals. Models of internalization hypothesize external motivations as a necessary first step on the path to automatization. On this account, high-IMS low-EMS individuals are low-EMS because they no longer need the support of external motivations having more fully integrated egalitarian values into their sense of themselves. The process of internalization is one of habituating patterns of responsiveness and, in the case of egalitarian commitments, of breaking others (Amodio, Devine and Harmon-Jones 2003, 751). All high-IMS individuals are committed to this process, but those at different points in the process have different motivational mixes.

Subsequent research supports the second, developmental model. This is important because it suggests that the adoption of explicit egalitarian commitments enables individuals to change their implicit motivations by automatizing control of implicit bias. Amodio, Devine and Harmon-Jones have shown that high-IMS low-EMS, but not high-IMS high-

EMS, individuals are able to inhibit the influence of implicit stereotypes on cognition through automatized conflict-monitoring (2008). High-IMS low-EMS individuals have highly accessible, automatized egalitarian commitments which conflict with implicit stereotypes at an early enough stage of cognitive processing for the conflict-monitoring mechanism to be effective in signalling the need for increased response regulation automatically and non-consciously. In contrast, high-IMS high-EMS individuals have less accessible, less automatized egalitarian commitments which are more reliant on the deliberative control effective only later in cognitive processing. The conflict-monitoring mechanism is therefore unable to signal the need for increased response regulation because little cognitive conflict occurs at the earlier stage of processing.

Further support for the effectiveness of automatized egalitarian commitments is provided by Moskowitz et al. who showed that highly accessible, enduring egalitarian goals can inhibit the activation of stereotypes preconsciously (Moskowitz et al. 1999; Amodio, Devine and Harmon-Jones 2008, 71–2). Individuals with similarly non-prejudiced attitudes and equally accessible cultural stereotypes but stronger commitments to fairness inhibited the influence of stereotypes on cognition preconsciously.

The effectiveness of automatized egalitarian commitments not only supports a stronger defence of virtue ethics by showing that habituated egalitarian motivations can reliably guide cognitive processing without the need for ongoing deliberative control. Once automatized, egalitarian commitments also have significant practical advantages over strategies requiring conscious control. In addition to inhibiting the influence of implicit bias on more automated cognitive processing, automatized egalitarian commitments are relatively efficient in terms of cognitive resources, relatively unimpeded by the erosion of cognitive capacity which results from effortful deliberation and, therefore, relatively immune to the potential for rebound which characterizes conscious efforts to suppress the effects of stereotypes on deliberation (Park, Glaser and Knowles 2008; Glaser and Knowles 2008; Moskowitz et al. 1999, 181). For instance, cognitive depletion increases 'shooter bias' for individuals low, but not high, in implicit motivation to control prejudice (IMCP) (Park, Glaser and Knowles 2008). Furthermore, Park, Glaser and Knowles argue that the character of this particular psychological construct makes their demonstration of the effectiveness of implicit egalitarian motivations especially reliable (2008, 416). IMCP is a measure of the strength of two implicit associations: first, that between prejudice and bad; second, that between self and prejudice. Individuals who are strongly motivated by concerns about self-presentation and who have highly effective generic regulative capacities would be expected to demonstrate strong associations between prejudice and bad whether they actually had such associations or not. These individuals would also be expected to demonstrate weak associations between self and prejudice for just the same reasons, however, and so would not be assessed as high IMCP. Only individuals relatively unconcerned about self-presentation or with relatively weak generic regulative capacities would be expected to demonstrate both of the strong associations required for high IMCP. This makes it likely that the relation between high IMCP and the ability to inhibit the influence of implicit bias on cognition is a specific effect of strong implicit egalitarian motivations. This does not mean that the effectiveness of automatized egalitarian commitments requires an implicit (or explicit) belief that one is prejudiced. As Glaser and Knowles point out, the finding that strongly associating prejudice with bad is enough to inhibit the influence of implicit bias, but that strongly associating self with prejudice is not, is just what one would expect. One can be motivated by an egalitarian goal (unprejudiced behaviour) whether one thinks one is currently prejudiced or not, but merely believing that one is prejudiced will fail to motivate behavioural regulation unless one disvalues prejudice (2008, 170).

Taken together, these two research programmes are good news for the virtue ethicist. A satisfactory response to the challenge posed by implicit bias depends on two things. First, it must be possible to embed reflectively endorsed egalitarian motivations in the cognitive-affective processing system through habituation. This is supported by evidence for the developmental model of egalitarian motivation from IMS/EMS research. Second, habituated egalitarian motivations must be able to effectively guide cognition outside conscious awareness. This is supported by evidence from work on the effectiveness of implicit egalitarian commitments. Moreover, the gradual internalization and automatization of conscious egalitarian commitments, and the effectiveness of highly accessible and automatized egalitarian goals, is just what Mischel and Shoda's model of personality and work on goal automaticity predicts.

6. Puzzles About Automatization

Despite the promise of automatized egalitarian commitments, however, the picture which emerges from current research raises some puzzling questions. Plant and Devine's measures of IMS and EMS differ significantly from Glaser and Knowles's measure of IMCP. Whereas the former depend on self-report, the latter are assessed using implicit measures of association. Moreover, it is currently unclear how these constructs are related. Glaser and Knowles found neither high-IMS alone nor high-IMS low-EMS to affect the relation between strength of race-weapons stereotype and shooter bias (2008, 170–1). Similarly, Park, Glaser and Knowles found IMS and EMS to be correlated with neither race-weapons stereotype nor shooter bias, and no relation between IMCP and IMS, EMS or IMS-EMS interaction (2008, 414).

These results are puzzling in several respects. If high-IMS low-EMS individuals inhibit the influence of bias on cognition *via* automatized conflict-monitoring as the IMS/EMS research suggests, why do they not inhibit the effects of race-weapons stereotypes on their responses in the shooter task? This discrepancy cannot be explained by differences in the kinds of implicit bias studied because the discrepancy appears to affect studies specifically focused on stereotypes. Whereas IMS/EMS researchers have found high-IMS low-EMS individuals to inhibit stereotypes preconsciously (Moskowitz et al. 1999; Amodio, Devine and Harmon-Jones 2008), IMCP researchers have found no relation between high-IMS low-EMS and stereotype inhibition (Glaser and Knowles 2008; Park, Glaser and Knowles 2008). Why should high-IMS low-EMS inhibit stereotypes preconsciously in one research programme but not the other? Moreover, if high-IMS low-EMS individuals avoid biased responses even on measures which elude conscious control because they have largely succeeded in automatizing their egalitarian commitments, why do they not demonstrate a highly accessible negative association with prejudice? That is, although high-IMS low-EMS does not seem obviously predictive of a strong association between self and prejudice, it seems odd that it does not correlate with a strong association between prejudice and bad. Furthermore, the associative model of personality seems *prima facie* to rule out explaining the various results in terms of two distinct routes to bias reduction, especially since both the conflict-monitoring element of the IMS/EMS research project and work on IMCP have examined the influence of race-weapons stereotypes (cf. Glaser and Knowles 2008, 171).

As Devine et al. suggest, longitudinal studies of the development of IMS and EMS are needed to establish how high-IMS low-EMS individuals avoid bias and what might assist high-IMS high-EMS individuals to effectively pursue their egalitarian goals, as well as features of the social environment which might encourage low-IMS individuals to identify with egalitarian values (Devine et al. 2002, 846). Given the apparent discrepancies between results from work on IMS/EMS and IMCP, however, further research to clarify the relationship between the various constructs developed in the psychological literature on egalitarian motivation will be equally important. Particularly useful might be work comparing factors which support and sustain individual development of high-IMS low-EMS and high-IMCP.

One possibility is that high-IMS low-EMS individuals have strong egalitarian *goals* whereas high-IMCP individuals have strong egalitarian *attitudes* in the sense explained in section 4. Although high-IMCP is described in terms of goals in the literature, these are indirectly inferred from measures of implicit associations between prejudice and bad, and between self and prejudice.

A second possibility is that the motivation for adopting egalitarian goals is what matters: there is a difference between being motivated to pursue a personally important goal and being personally motivated to pursue an important goal. Completing this chapter might be personally important to me without my disapproving of those with no interest in pursuing academic philosophy. In contrast, kindness might be an important moral value such that I am concerned not only to be kind myself but to encourage and approve kindness in others. Perhaps high-IMCP individuals show less shooter bias than high-IMS low-EMS individuals because they are committed to egalitarianism for different reasons: whereas the former are personally motivated to avoid prejudice because they disvalue it generally, the latter may see it as a merely personal project, albeit one they happen to care strongly about. Perhaps this explains why high-IMS low-EMS does not predict a strong negative association with prejudice as such.

A third possibility is that high-IMS low-EMS alone is insufficient for the formation of an enduring, highly accessible egalitarian goal. Amodio, Devine and Harmon-Jones included an experimental manipulation to increase the accessibility of subjects' internal, but not external, motivations to avoid prejudice (2008, 63). Although the automatized egalitarian values bound up with the self-concepts of high-IMS low-EMS individuals enabled them to control their responses, it is possible that the automatization of egalitarian goals is a further step. Individuals without such automatized goals might be insufficiently sensitive to situations requiring control and so fail as a group to demonstrate less shooter bias when the need for control is not made especially salient. While one might expect the content of the shooting task to make such a need salient, perhaps this is undermined by the artificial form of the simulation.

7. Evaluating the Egalitarian's Toolbox

The defence of virtue ethics outlined so far appeals to the potential of automatized egalitarian commitments to inhibit the cognitive influence of implicit bias outside conscious awareness and without the need for deliberative control. A further advantage of habituated egalitarian commitments is their potential to eliminate or reduce our implicit biases themselves. The dynamic, associative model of personality developed by Mischel and Shoda suggests that enduring commitments to egalitarianism should decrease implicit bias itself as a long-term effect of the automatization which enables such commitments to prevent its influence on cognition.

Current research provides qualified support for this possibility. For example, Rudman, Ashmore and Gary found that students who voluntarily enrolled in diversity education demonstrated reductions in implicit, as well as explicit, bias (2001). The class was designed to increase students' awareness of racial prejudice and motivation to overcome racism in themselves, as well as providing the opportunity to make social contact with 'out-group' members in a safe and supportive atmosphere. The results suggested distinct, but mutually supportive, cognitive and affective routes to reductions in explicit and implicit bias respectively. Egalitarian commitments were plausibly partially responsible for these effects: a decision to enrol in diversity education suggests motivation to overcome prejudice (Rudman, Ashmore and Gary 2001, 866) and the class was designed to consolidate and support pursuit of this initial commitment. It is unfortunate that long-term results were not evaluated since it would be useful to know if the two routes to bias reduction would converge over time, as Mischel and Shoda's model of personality predicts. Although cognitive and affective changes might differentially support reductions in explicit and implicit bias in the short-term, this model of cognition understands the two routes as affecting a single connectionist network. Given the short-term nature of the study, however, it is unsurprising that the two routes were only weakly correlated: the model predicts that change will be gradual and that adjustments to one element of the system (e.g. a belief) will affect associated elements of the system (e.g. an affective response) only slowly as the initial change repeatedly affects the flow of information across the system. Although such research suggests that interventions might plausibly 'kick-start' the process of automatization, therefore, further research is required to substantiate this possibility.

Moreover, this picture does not yet capture the complexity of psychological reality. Research suggests that the effect of automatized egalitarian commitments depends on the *kind* of implicit bias in question. Whereas egalitarian commitments seem to reduce or eliminate implicit affective bias, for example, they seem to inhibit rather than weaken implicit stereotypes (Amodio, Devine and Harmon-Jones 2003, 2008, 63; Moskowitz et al. 1999).

Why should the kind of implicit bias matter if the cognitive-affective personality system is a connectionist network? One difference between affective bias and stereotypes is that an understanding of the social world seems to depend on the latter but not the former. Perhaps some features of the cognitive-affective system (e.g. implicit stereotypes) may be strongly connected to elements implicated in processing motivated by a need to understand, despite being weakly connected to elements implicated in processing motivated by a need to respond. So long as her society harbours stereotypes, a member of that society will need to be aware of them in order to effectively navigate her social world. The process of automatizing egalitarian goals might therefore weaken connections between stereotypes and elements of the cognitive system implicated in processing aimed at responding without weakening the stereotypes' connections with elements implicated in processing aimed at social understanding. Indeed, automatized egalitarian goals might be partially constituted by the preconscious inhibition of stereotypes in response-directed processing. If this were the whole story, the stereotypes themselves might be expected to weaken over time. Since stereotypes are crucial to understanding the social world, however, processing aimed at social understanding would continue to activate and sustain them. Because response-directed processing flows across the same network as processing aimed at understanding, the automatization of egalitarian goals would therefore tend to isolate stereotypes from response-directed processing without eliminating them.

8. Beyond Individual Commitment

I have argued that individual egalitarian commitments can play an essential part in resisting implicit bias and the challenge it presents to virtue ethics. But individuals' ability to sustain and implement their commitments depends crucially on hospitable environments. The process of automatization is designed to select and refine *successful* strategies and response patterns. We automatically adjust our strategies in response to their success or failure in enabling us to navigate social interactions. Although we can deliberatively adopt egalitarian goals independently of others, therefore, we have only limited control over our ability to automatize them, because the success of strategies aimed at achieving those goals depends on others' cooperation. While our ability to automatize intelligent and flexible responses can support individuals' pursuit of egalitarian commitments, therefore, this intelligence and flexibility also renders those commitments vulnerable to inhospitable social environments. In hospitable environments, egalitarian commitments will be strengthened by the process of modification and refinement which is essential to automatizing them, enabling individuals to habituate appropriately sensitive responses to pertinent features of their social environments. In hostile environments, however, this same process will tend to weaken and undermine individuals' egalitarian commitments, because the cognitive system will automatically modify and refine them in response to difficulties in implementing them, others' hostile responses, and unsuccessful social interactions. Community support for egalitarianism is, therefore, essential for the development and maintenance of egalitarian virtue not because there is nothing which would count as virtue in a prejudiced social environment, but because the human cognitive processing system is designed to automatically adapt our responses to whatever social environment we happen to inhabit. Communities and institutions which themselves embody egalitarian values, and which encourage more individuals to adopt egalitarian goals, are therefore crucial. Environmental and cultural interventions which foster and support strong commitments on the part of individuals, and which seek to make such commitments institutional and social norms, are thus essential to egalitarian toolboxes.

This echoes Aristotle's emphasis on the development and practice of individual virtue in the context of a supportive moral community. The process of automatizing egalitarian goals outlined here just is the habituation of appropriate values and commitments. The internalization and automatization of egalitarian motivation is a process of tuning the cognitive-affective personality system to respond appropriately to just the right features of individuals' external social and internal psychological environments. That is, the habituation of appropriate responsiveness to egalitarian reasons is part of the development and practice of practical wisdom[†]

References

Amodio, David M., Patricia G. Devine and Eddie Harmon-Jones (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* 84.4, 738–753.

— (2008). 'Individual Differences in the Regulation of Intergroup Bias: The Role of Conflict Monitoring and Neural Signals for Control'. *Journal of Personality and Social Psychology* 94.1, 60–74.

Annas, Julia (2011). *Intelligent Virtue*. Oxford and New York: Oxford University Press.

Aristotle (2002). *Nicomachean Ethics*. Trans., with a historical introd., by Christopher Rowe. Philosophical introd. and comment. by Sarah Broadie. Oxford and New York: Oxford University Press.

[†] I am grateful to all those involved in the Implicit Bias and Philosophy International Research Project for supporting and stimulating my thinking on this topic, especially participants at the July 2012 workshop who provided feedback on the ideas which formed the basis for this chapter. I am particularly grateful to Jennifer Saul for supporting my participation in the project. I would like to thank Gregory R. Maio and Geoffrey Haddock for allowing me to attend their lectures on the social psychology of attitudes and for their patience in answering a philosopher's questions, and members of Cardiff's moral psychology reading group for ongoing discussions of the implications of empirical psychology for ethical theory. I owe special thanks to Michael Brownstein, Jennifer Saul and Jonathan Webber for extensive and invaluable feedback on earlier drafts, and to the anonymous referees whose insights helped me to see where further clarification was needed.

Baier, Annette C. (1995). 'What Do Women Want in a Moral Theory?' In *Moral Prejudices: Essays on Ethics.* Cambridge, Massachusetts: Harvard University Press, 1–17. Revision of 'What Do Women Want in a Moral Theory?' *Noûs* 19.1 (Mar. 1985), 53–63.

Bargh, John A. (1989). 'Conditional Automaticity: Varieties of Automatic Influence in Social Perception and Cognition'. In *Unintended Thought.* Ed. by James S. Uleman and John A. Bargh. New York and London: Guilford. Ch. 1, 3–51.

— (1990). 'Auto-Motives: Preconscious Determinants of Social Interaction'. In *Handbook of Motivation and Cognition: Foundations of Social Behavior.* Ed. by Richard M. Sorrentino and E. Tory Higgins. Vol. 2. New York and London: Guilford. Ch. 3, 93–130.

— (2006). 'What Have We Been Priming All These Years? On the Development, Mechanisms, and Ecology of Nonconscious Social Behavior'. *European Journal of Social Psychology* 36, 147–168.

Bargh, John A. and Peter M. Gollwitzer (1994). 'Environmental Control of Goal-Directed Action: Automatic and Strategic Contingencies between Situations and Behavior'. In *Integrative Views of Motivation, Cognition and Emotion.* Ed. by William D. Spaulding. Vol. 41. Nebraska Symposium on Motivation. Lincoln and London: University of Nebraska Press, 71–124.

Bargh, John A., Peter M. Gollwitzer et al. (2001). 'The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals'. *Journal of Personality and Social Psychology* 81.6 (Dec. 2001), 1014–1027.

Bargh, John A. and Erin L. Williams (2006). 'The Automaticity of Social Life'. *Current Directions in Psychological Science* 15.1, 1–4.

Baumeister, Roy F. et al. (1998). 'Ego Depletion: Is the Active Self a Limited Resource?' *Journal of Personality and Social Psychology* 74.5, 1252–1265.

Blair, Irene V., Jennifer Ma and Alison Lenton (2001). 'Imagining Stereotypes Away: The Moderation of Implicit Stereotypes Through Mental Imagery'. *Journal of Personality and Social Psychology* 81.5, 828–841.

Blank, Rebecca M. (1991). 'The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from the American Economic Review'. *American Economic Review*, 1041–1067.

Brennan, Samantha (2009). 'Feminist Ethics and Everyday Inequalities'. *Hypatia* 24.141, 141–159.

Budden, Amber E. et al. (2008). 'Double-Blind Review Favours Increased Representation of Female Authors'. *Trends in Ecology & Evolution* 23.1, 4–6.

Casey, Pamela M. et al. (2012). *Helping Courts Address Implicit Bias: Strategies to Reduce the Influence of Implicit Bias.* Project summary. National Center for State Courts. Summary of *Helping Courts Address Implicit Bias: Resources for Education.* Project report. National Center for State Courts, 2012.

Changing Perceptions with Changing Faces (2013). Remploy. URL: http://www.remploy.co.uk/__assets/downloads/pdfs/changing-faces-the-ugly-face-of-prejudice.pdf (visited on 17/07/2013).

Cooper, Joel M. (2007). *Cognitive Dissonance: Fifty Years of a Classic Theory.* Los Angeles et al.: SAGE.

Corcoran, Katja, Tanja Hundhammer and Thomas Mussweiler (2009). 'A Tool for Thought! When Comparative Thinking Reduces Stereotyping Effects'. *Journal of Experimental Social Psychology* 45.4, 1008–1011.

Dasgupta, Nilanjana and Shaki Asgari (2004). 'Seeing Is Believing: Exposure to Counterstereotypic Women Leaders and Its Effect on the Malleability of Automatic Gender Stereotyping'. *Journal of Experimental Social Psychology* 40.5, 642–658.

Dasgupta, Nilanjana and Anthony G. Greenwald (2001). 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice With Images of Admired and Disliked Individuals'. *Journal of Personality and Social Psychology* 81.5, 800–814.

Devine, Patricia G. et al. (2002). 'The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice'. *Journal of Personality and Social Psychology* 82.5 (May 2002), 835–848.

Doris, John Michael (2009). 'Skepticism About Persons'. *Philosophical Issues* 19: *Metaethics*, 57–91.

Equalitylink May 2013 (2013). Equality Challenge Unit. URL: http://www.ecu.ac.uk/news/equalitylink/2013/05 (visited on 17/07/2013).

Galinsky, Adam and Gordon B. Moskowitz (2000). 'Perspective-Taking: Decreasing Stereotype Expression, Stereotype Accessibility, and In-Group Favoritism'. *Journal of Personality and Social Psychology* 78.4, 708–724.

Glaser, Jack and Eric D. Knowles (2008). 'Implicit Motivation to Control Prejudice'. *Journal of Experimental Social Psychology* 44.1, 164–172.

Goldin, Claudia and Cecilia Rouse (2000). 'Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians'. *American Economic Review* 90.4, 715–741.

Isen, Alice M. and Gregory Andrade Diamond (1989). 'Affect and Automaticity'. In *Unintended Thought.* Ed. by James S. Uleman and John A. Bargh. New York and London: Guilford. Ch. 4, 124–152.

Jost, John T. et al. (2009). 'The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies That No Manager Should Ignore'. *Research in Organizational Behavior* 29, 39–69.

Kamtekar, Rachana (2004). 'Situationism and Virtue Ethics on the Content of Our Character'. *Ethics* 114.3 (Apr. 2004), 458–491.

Kelly, Daniel, Edouard Machery and Ron Mallon (2010). 'Race and Racial Cognition'. In *The Moral Psychology Handbook.* Ed. by John Michael Doris and The Moral Psychology Research Group. Oxford: Oxford University Press. Ch. 13, 433–472.

Kelly, Daniel and Erica Roedder (2008). 'Racial Cognition and the Ethics of Implicit Bias'. *Philosophy Compass* 3.3, 522–540.

Lerner, Jennifer S. and Philip E. Tetlock (1999). 'Accounting for the Effects of Accountability'. *Psychological Bulletin* 125.2, 255.

Maio, Gregory R. and Geoffrey Haddock (2010). *The Psychology of Attitudes and Attitude Change.* Los Angeles et al.: SAGE.

Merritt, Maria W. (2009). 'Aristotelean Virtue and the Interpersonal Aspect of Ethical Character'. *Journal of Moral Philosophy* 6, 23–49.

Merritt, Maria W., John Michael Doris and Gilbert Harman (2010). 'Character'. In *The Moral Psychology Handbook*. Ed. by John Michael Doris and The Moral Psychology Research Group. Oxford: Oxford University Press. Ch. 11, 355–401.

Milgram, Stanley (2009). *Obedience to Authority: An Experimental View*. With an intro. by Philip G. Zimbardo. New York: HarperCollins/Harper Perennial Modern Thought. Repr.

Mischel, Walter and Yuichi Shoda (1995). 'A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure'. *Psychological Review* 102.2 (Apr. 1995), 246–268.

Moskowitz, Gordon B. et al. (1999). 'Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals'. *Journal of Personality and Social Psychology* 77.1 (July 1999), 167–184.

Muraven, Mark and Roy F. Baumeister (2000). 'Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle?' *Psychological Bulletin* 126.2, 247–259.

Muraven, Mark, Dianne M. Tice and Roy F. Baumeister (1998). 'Self-Control as Limited Resource: Regulatory Depletion Patterns'. *Journal of Personality and Social Psychology* 74.3, 774–789.

Park, Sang Hee, Jack Glaser and Eric D. Knowles (2008). 'Implicit Motivation to Control Prejudice Moderates the Effect of Cognitive Depletion on Unintended Discrimination'. *Social Cognition* 26.4 (Aug. 2008), 401–419.

Pelham, Brett, Matthew Mirenberg and John Jones (2002). 'Why Susie Sells Seashells by the Seashore: Implicit Egotism and Major Life Decisions'. *Journal of Personality and Social Psychology* 82.4, 469–487.

Peters, Douglas P. and Stephen J. Ceci (1982). 'Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again'. *Behavioral and Brain Sciences* 5.02 (June 1982), 187–195.

Plant, E. Ashby and Patricia G. Devine (1998). 'Internal and External Motivation to Respond Without Prejudice'. *Journal of Personality and Social Psychology* 75.3, 811–832.

Quinn, Andrew and Barry R. Schlenker (2002). 'Can Accountability Produce Independence? Goals as Determinants of the Impact of Accountability on Conformity'. *Personality and Social Psychology Bulletin* 28.4, 472–483.

Rees, Clea F. and Jonathan Webber (2014). 'Automaticity in Virtuous Action'. In *The Philosophy and Psychology of Virtue: An Empirical Approach to Character and Happiness*. Ed. by Nancy E. Snow and Franco V. Trivigno. New York and London: Routledge. Ch. 4, 75–90.

Rudman, Laurie A., Richard Ashmore and Melvin Gary (2001). '"Unlearning" Automatic Biases: The Malleability of Implicit Prejudice and Stereotypes'. *Journal of Personality and Social Psychology* 81.5, 856–868.

Russell, Nestar John Charles (2011). 'Milgram's Obedience to Authority Experiments: Origins and Early Evolution'. *British Journal of Social Psychology* 50.1, 140–162.

Saul, Jennifer (2012). 'Ranking Exercises in Philosophy and Implicit Bias'. *Journal of Social Philosophy* 43.3, 256–273.

Snow, Nancy E. (2009). *Virtue as Social Intelligence: An Empirically Grounded Theory.* New York: Routledge.

Webb, Thomas L., Paschal Sheeran and John Pepper (2012). 'Gaining Control Over Responses to Implicit Attitude Tests: Implementation Intentions Engender Fast Responses on Attitude-Incongruent Trials'. *British Journal of Social Psychology* 51.1, 13–32.