Instilling Virtue

Jonathan Webber

– Working draft – comments welcome (email via www.jonathanwebber.co.uk) –

Final version will appear in
From Personality to Virtue: Essays in the Psychology and Ethics of Character
edited by Alberto Masala and Jonathan Webber

Two debates in contemporary philosophical moral psychology have so far been conducted almost entirely in isolation from one another despite their structural similarity. One is the debate over the importance for virtue ethics of the results of situational manipulation experiments in social psychology. The other is the debate over the ethical implications of experiments that reveal gender and race biases in social cognition. In both cases, the ethical problem posed cannot be identified without first clarifying the cognitive structures underlying the problematic phenomena. In this chapter, I argue that the two kinds of phenomena share a basic cognitive structure, which is well articulated by the findings of the empirical psychology of attitudes, especially if these findings are understood in the context of the cognitive-affective system theory of personality. On the basis of this joint construal of situationism and implicit bias, I argue that the negative programme of ethical improvement that many philosophers recommend in response to one or other problem is unrealistic. Instead, we should consider more seriously the prospects of the positive programme of ethical improvement recommended by Aristotle, the direct aim of which is to instil deeply in ourselves the values at the heart of each of the virtues.

1. Situational Manipulation and Implicit Bias

The richest and most robust demonstration of situational manipulation in social psychology remains Stanley Milgram's investigation into the ease with which people can be persuaded to inflict what appear to be potentially lethal electric shocks on what appears to be a fellow volunteer. By varying the experimental setup, Milgram showed that subtle situational differences made a significant difference to the degree to which subjects did as they are asked (Milgram 1974). Indeed, a recent analysis of Milgram's personal archive makes clear that Milgram employed such situational manipulation to design his experiment in the first place. Through a series of pilot studies, he refined the instructions given to the subjects, their sensory access to the effects of the shocks on their victim, the design of the shock generator itself, and various other details of the experiment, with the express aim of finding a surprising headline result. Having found a structure that would produce such a surprising result, Milgram published this version of the experiment first, subsequently referring to it as the 'baseline condition' when publishing results of other versions of the experiment (Russell 2011).

Does it matter, scientifically, that Milgram refined his experiment until it achieved his desired results? Does it matter that he designated one version the 'baseline condition' purely because its results were most likely to attract attention? It depends on the lesson that one wants to draw from the data. The results of the 'baseline condition' should not be taken in isolation. Taken together, the variations of the experiment might provide important evidence concerning the details of human motivation. There is certainly one general truth, however, that they and the pilot studies clearly reveal: that the subjects' response to the morally most important aspect of the situation, the requests to inflict high levels of electric shock on another person, vary significantly with the other aspects of the situation, many of which seem to be of no moral importance at all.

Research into implicit bias probes more deeply into the cognitive architecture that generates this situational variation of behaviour. One such experiment found subjects to hold much stronger

cognitive associations between white people and positive evaluation than between black people and positive evaluation. A series of words appeared on a screen and subjects were asked to press one of two keys to classify each word into one of two categories. They were asked to make their judgments as quickly as possible, but not so fast as to allow mistakes. Some of the words (such as 'crash', 'happy', 'peace', 'rotten') were to be categorised according to whether they are pleasant or unpleasant. These were mixed with names (such as 'Ebony', 'Jed', 'Katie', 'Lamar') that were to be categorised as typically names of black people or typically names of white people. The subjects themselves were white. Their reaction times were much faster when the button for indicating pleasant words was also the one used to indicate that a name is typically of a white person than when the button for indicating pleasant words was also used to indicate that a name is typical of a black person (Greenwald et al 1998: Experiment 3).

Moreover, a subsequent replication of the experiment found these cognitive associations to correlate strongly with biases in the way subjects behaved towards a black experimenter and a white experimenter. In particular, the bias in cognitive associations correlated with differences in the amount of time the subjects spent talking to each experimenter, the proportion of that time the subjects spent actually facing the experimenter, and the physical distance they set between themselves and each experimenter. These behavioural biases were evident to the experimenters in the discussion as well as to external observers (McConnell and Leibold 2001).

For both kinds of experiment, it has been shown that subjects would explicitly disavow the attitudes implied by their behaviour. Of the many subjects who had obeyed his experimenter to a high level of electric shock, Milgram found that very few subsequently claimed that they had done the right thing in the circumstances (Milgram 1974: ch. 5). When the 'baseline' version of the experiment was explained to diverse groups of people and they were asked to predict how they themselves would behave, most said they would stop the experiment when the shock levels were still very low (Milgram 1974: ch. 3). The original experiment into cognitive associations of names of black people and white people with positive evaluation asked the same subjects explicit

questions about their attitudes to black people and white people, about the causes of discrimination, and about the value of multiculturalism. Many subjects' responses indicated no racial preference or a mild preference in favour of black people even though the implicit association test had indicated a preference for white people in those same subjects. The experimenters concluded that the results should be taken as 'indicating the pervasiveness of unconscious forms of prejudice' (Greenwald et al 1998: 1475).

2. Evaluative Judgments and the Cognitive-Affective Personality System

How should we understand the disparity evident in both the Milgram experiment and the implicit association experiment between the attitudes indicated by the subjects' behaviour and the attitudes indicated by their explicit judgments? What is the moral problem that this disparity poses? One answer to these questions distinguishes between a person's evaluative beliefs and their behavioural dispositions. Evaluative beliefs, on this account, are reflectively held and are reported in explicit avowals of belief and in conscious judgments about what one should do, but our behaviour is governed by these beliefs only to the extent that it results from conscious deliberation. The more intuitive and automatic aspects of our behaviour manifest our behavioural dispositions. Given this account of action, the moral task presented by the disparity between behaviour and evaluative judgment is the task of training one's behavioural dispositions to bring them into line with one's evaluative beliefs (Besser-Jones 2008).

An alternative account denies that people have evaluative beliefs that remain consistent across contexts. Not only behavioural responses are dependent on seemingly irrelevant aspects of the context in which they are made, on this view. The same is true of explicit, conscious, deliberative evaluative judgments. The influence these details have over judgment, as with their influence over behavioural responses, need not be noticed by the subject and might not be endorsed if brought to the subject's attention. One form of this account rests on the idea that mental states each have a degree of accessibility, measured by the time it takes to be brought to bear on

cognition. The more rapidly one makes a judgment, the fewer relevant considerations are going to be taken into account. The more slowly and effortfully one deliberates, the more one brings into play relevant beliefs and desires that have lower degrees of accessibility. On this view, the moral task is to ensure that one makes well considered judgments when it matters, perhaps by adopting the strategy of imagining justifying one's judgment to an audience whose values are unknown to oneself (Merritt 2009).

Both accounts seem consistent with the cognitive-affective system theory of personality, which is the dominant theory of the cognition generating behaviour across situations. This theory was developed to account for the stability in an individual's behaviour across repetitions of the same situation as well as the variation in that individual's behaviour across situations that differ in subtle details, a stability and variation which together make the individual's 'behavioural signature that reflects personality coherence' (Mischel and Shoda 1995: 251). It is not just an individual's set of mental states themselves that determine their behaviour, since 'it is the organisation of the relationships among them that forms the core of the personality structure and that guides and constrains their impact' (Mischel and Shoda 1995: 253). Each mental state is associated with many others and these connections vary in strength. They are formed through experience and strengthened through use. The cognition that generates behaviour is a flow of activity across this network of cognitive and affective states, constrained by the stimuli presented by the environment. Because the system develops only slowly, the resulting behaviour is likely to be the same on two occasions where the situation is the same. But where a detail of the situation is different, this may result in a different flow of activity through the personality system and thus a different behavioural outcome.

This theory was developed to account for behavioural patterns. But if this is the right picture of personality generally, rather than simply of behaviour, then we ought to be able to understand evaluative judgments in terms of it as well. There seem to be two ways in which this personality system might generate evaluative judgments. One corresponds to the view that such judgments

remain consistent across contexts. If this is right, then the judgments in question manifest stable evaluative beliefs, such as the belief that people of different ethnicities are equal. Such a belief would be a mental state within the personality system. Alternatively, we might understand evaluative judgments to be generated by the personality system in much the same way as behaviour. Since the number and strengths of a mental state's associations determine the speed with which it influences cognition, a given evaluative judgment will depend on the amount of time devoted to seriously deliberating about the issue.

The cognitive-affective personality system was proposed as a 'framework within which to conceptualise and conduct research to understand the intra-individual dynamics of personality and their expression' (Shoda and Mischel 1996: 415). One way to develop this research is to consider which of the two accounts of evaluative judgment is correct. Do evaluative judgments express beliefs that are themselves stable units within the personality system, or are they situationally variable products of the personality system? We will see that this question can be answered by augmenting the personality system theory with the findings of attitude psychology.

3. Moral Choice Blindness

Current research into 'moral choice blindness' casts doubt on the idea that moral beliefs are generally stable. In one experiment, subjects were asked to complete a two-page survey that asked how much they agreed with a series of moral statements by giving a score on a numerical scale. During the course of the survey, some of the statements they had responded to were switched for their negations. After the survey had been completed, subjects were asked to work through each statement and justify the response they had given. Experimenters were interested in whether subjects would notice that some of the statements had been negated, or whether they would offer reasons in favour of the opposite view to the one they had originally expressed. If subjects were expressing stable moral beliefs, we would expect them not to then justify the

opposite of the view they expressed. But if they do go ahead and justify the opposing view, then it seems that they are not expressing a stable moral belief at all.

The switch of statements was very well designed. The survey was on two pages of paper attached to a clipboard. The statements on the first page were actually printed on a separate piece of paper invisibly glued to the page. On the back of the clipboard, in exactly the right place, was a patch of stronger glue, so that when the subject turned to the second page the statements from the first page stayed on the back of the clipboard, revealing a different set of statements underneath. Some of the statements in this set were the same as the ones that had been glued over them. Some were the negations of those statements. The experimenters recorded the switch as having been detected if the subject spontaneously corrected for it by changing their response rather than defending it, or if the subject expressed any suspicions about the statements in the post-experimental discussion, or if the subject could correctly identify which statements had been reversed once they had been told how the experiment worked. Even with this generous range of forms of detection, the majority of subjects did not detect any change to the statements.[1]

Two conditions are necessary for a subject to fail to detect the change in statement and blithely justify the opposite of the view they originally expressed. One is that the subject does not remember the original statement. The other is that the subject's judgments on the topic are not consistent across situations. For if the subject did hold a stable view on the issue that the statement concerned, then the subject should express that same stable view when they first complete the survey and when they are asked to justify the responses on the page in front of them. The two situations in which the subject is asked to express a view on this topic differ only in one respect. In the second situation only, they are provided with false evidence concerning the view they expressed moments earlier. For the majority of the subjects, those who did not detect the switch, this difference in situation is enough to negate the judgment that they express. In

[1] A film of the survey being used can be seen at: http://www.lucs.lu.se/cbq/

terms of the cognitive-affective personality system, the two situations caused different sequences of activity through the network of mental states leading to different judgments being expressed.

What of those who did detect the switch? Perhaps detection was due to the subjects holding stable views on the matters that the switched statement concerned. This would explain the consistency in judgment across the two situations. In the second situation, the subject would recognise that the view expressed on the page is not their own view, so would assume that something had gone wrong, perhaps that they had misunderstood the question first time around. In terms of the cognitive-affective personality system, there are two ways in which such a stability of judgment might occur. One is that the judgement simply expresses a particular mental state in the system, a moral belief that remains constant irrespective of its position in the network. The other is that the judgments in both cases were generated by the personality system as a whole, with the difference between the two situations being insufficient to cause a different outcome from this cognition.

However, stability of judgment across situations is not the only possible explanation of detection in this experiment. For it was a short survey and subjects were asked to justify their responses as soon as they had finished it. So it remains possible that those who detected the switch did so simply because they remembered the original statement. This leaves us with two candidate explanations of the experiment overall. One is that some of the subjects had stable moral beliefs where others did not. The other is that none of the subjects expressed stable moral judgments, but some subjects were better than others at remembering the statements. Either way, the experiment presents evidence against the idea that moral judgments generally express stable moral beliefs.

4. Strength and Influence in Attitude Psychology

Which of these two explanations of the moral choice blindness experiment is correct? Empirical research into the nature and influence of evaluative attitudes suggests that some subjects had sufficiently stable attitudes concerning the topic of the switched statements to detect the switch, but other subjects did not. An attitude's stability over time is a matter of the 'strength' or firmness with which that attitude is held. This is distinct from the attitude's content. For example, you might hold a positive attitude towards democracy as a political system. The overall content of this attitude is the degree to which you approve of democracy, though the content can also be characterised in more detail to include what you think of various aspects of democracy and various different democratic systems. Attitude psychologists reserve the term 'strength' for a different dimension of the attitude. This is the degree to which the attitude is embedded in your cognitive system. An attitude that is strong in this sense is not easily changed by persuasion or reconsideration.

One classic experiment concerning the effects of attitude strength measured the relation between subjects' attitudes towards Greenpeace and their response to an opportunity to donate to Greenpeace (Holland et al 2002). At the first stage of the experiment, subjects completed a lengthy questionnaire, which included questions about their attitude towards Greenpeace. They were asked how much they approved or disapproved of Greenpeace, how certain they were of their attitude towards Greenpeace, how important this attitude was to them personally, whether this attitude is central to their self-image, and whether this attitude reflects values they hold to be important. The first of these questions measured the attitude content, the other four measured its strength or firmness. Subjects returned a week later for a different experiment. After that experiment was over, they were paid for their participation. The payment consisted of a set of coins. They were then offered the opportunity to donate some of the money to Greenpeace and were asked to complete a short questionnaire about Greenpeace. One of these questions asked the subjects to evaluate the work of Greenpeace on a scale of 1 to 10.

Subjects whose attitudes towards Greenpeace were strong, or firmly held, when measured at the start of the experiment acted in line with those attitudes when offered the chance to donate to Greenpeace a week later. Those with strong attitudes in favour of Greenpeace donated, whereas those with strong attitudes against did not. By contrast, there was no significant relation between the attitude reported at the start of the experiment and the response to the opportunity to donate to Greenpeace among those subjects whose attitudes towards Greenpeace did not score highly on the strength measures at the start of the experiment. Moreover, the results of the second attitude measure, taken immediately after the opportunity to donate, show that those whose attitudes had scored highly on the strength measures a week earlier tended to report the same attitude at this point. Subjects who had originally reported weak attitudes, on the other hand, did not tend to report the same attitude at this point. Indeed, the attitude they reported at this point reflected whether or not they had just donated to Greenpeace, which in turn was unrelated to their original attitude report.

This experiment illustrates a finding that attitude psychology has gradually converged upon, that strongly held attitudes consistently manifest in judgments about their objects and in behaviour, whereas weakly held attitudes do neither.[2] The experimenters explain this in terms of the structures of attitudes. A strong attitude, they argue, is a persisting state, but a weak attitude is constructed at the time at which it is needed (Holland et al 2002). The mental states that a weak attitude is constructed from will vary with the occasion, since their relative levels of accessibility vary according to their recent employment in cognition and since the amount of time and cognitive resources used to construct the attitude will vary. In terms of the cognitive-affective personality system, we can say that attitude strength is determined by the strengths of the

[2] This relation between attitude strength and behaviour, but not the point about consistency of explicit judgments across situations, was rather dramatically demonstrated some years earlier by Danny Axsom and Joel Cooper (1985). For discussion of this experiment in relation to Aristotle's theory of trait habituation, see Webber 2013: § 4.

associative connections between the mental states that compose the attitude. When these are sufficiently strong, the set of mental states will continually influence cognition together as a whole. But mental states linked by connections that are not stronger than most connections in the system will influence the flow of cognition individually, each according to its own range and strength of connections.

5. Attitude Strength and Consistent Judgment

We can understand the moral choice blindness experiments in terms of this relation between attitude strength and consistency of evaluative judgment across situations. If your attitude towards democracy, for example, is firmly held, then you are unlikely to be tricked into thinking that you have just expressed the negation of that attitude. In a moral choice blindness experiment where a switched statement concerned democracy, you would be likely to detect the switch. If the statement concerned some topic on which you do not hold a strong attitude, however, you would construct your response at the time on the basis of the available relevant beliefs and desires. When presented with your purported response to such a statement and asked to justify the response, if you did not remember that this was not in fact your response, then you will again construct and explain an attitude, but this time one of the most salient mental states drawn on in constructing the attitude would be the false belief about your response to the statement moments earlier. So the mental states drawn upon in the attitude construction will feature those most closely and strongly associated with the content of that purported response. These will be the reasons you then give.

A detail of the experiment supports this interpretation. The survey asked subjects whether they held strong moral opinions in general and whether they were politically active. Responses to the first of these did not correlate with whether the subject detected the statement negation. This is consonant with the explanation in terms of attitude strength, for one's answer to this general question seems unlikely to correlate with strength of attitude on the specific topics that the

negated statements concerned. There was, however, some correlation between detection of statement negation and the second question. More specifically, this correlation held for one group of subjects but not the other. For there were two versions of the survey. One presented highly general moral statements, such as 'even if an action might harm the innocent, it can still be morally permissible to perform it'. The other presented more specific statements, such as 'the violence Israel used in the conflict with Hamas is morally defensible despite the civilian casualties suffered by the Palestinians'. Subjects who considered themselves politically active and who were given the more specific moral questions were significantly more likely than any other subjects to detect the statement negation. This is unsurprising. Politically active people are more likely to hold strong attitudes on the moral aspects of specific political issues, but it does not follow that they are likely to hold strong general moral attitudes. Indeed, such a person might express agreement with a general moral statement on the basis of strong attitudes concerning specific applications of it, but when presented with evidence that they disagree with the general statement might justify this in terms of strong attitudes concerning other specific applications.

We can understand the Milgram experiment and the implicit bias experiment in the same way. Before one has encountered the Milgram experiment, one is extremely unlikely to have a strong attitude concerning how one ought to behave in precisely that situation. When asked to predict how one would behave in the experiment, therefore, one constructs an attitude from relevant mental states. When actually in the experiment, one also judges and acts on the basis of attitudes constructed at the time. Many subjects in the experiment do construct the attitude that people predict they would construct. This is manifested when the subjects argue with the experimenter, seek confirmation that their actions are causing no harm, and even briefly refuse to continue. But each prompt from the experimenter requires the subjects to construct their attitudes anew. As it does so, each prompt also changes the relative levels of accessibility of the mental states drawn on to construct the attitude. For this reason, a subject is likely to vacillate between judging that the experiment should stop and judging that it can continue. Neither judgment, moreover, is

wholehearted. Each attitude constructed incorporates considerations in favour of continuing and considerations against.

Do the explicit measures that are compared with the results of the implicit association test record persistent strong attitudes or weak attitudes? Although the original experiment found that the implicit test results were often contradicted by the same subject's explicitly reported attitudes, a subsequent replication with a minor alteration eliminated this divergence. In the original experiment, the explicit questions followed the implicit test (Greenwald et al 1998). In the replication, this order was reversed (McConnell and Leibold 2001). This suggests that the explicit measures generally recorded weak attitudes that were dependent on the situation. Since the reaction time differences measured by the implicit association test are imperceptible to the subject, the test often leaves the subject with the false impression that they have treated black and white faces equally in what is clearly a scientific measure of their attitudes. Since weak attitudes tend to confirm recent behaviour, as we saw in the Greenpeace experiment, these subjects are likely to report attitudes consonant with their false belief that they have just treated these two ethnic groups equally. When the question is asked before the implicit association test is taken, on the other hand, the weak attitude constructed is likely to manifest the same associations and accessibility levels as are then manifested in the test itself.[3]

If this is right, then neither of the two construals of situationist and implicit bias experiments that we began with is correct. One of these construals held that our behavioural dispositions are not always in line with our evaluative beliefs, which assumed our explicit moral judgments to be consistent across situations. The other construal denied this assumption, portraying moral

[3] The authors of the replication study hypothesise that having taken the implicit association test might increase the role of self-presentation effects in answering the explicit questions (McConnell and Leibold 2001: 440-1). Since the explicit measure was thoroughly anonymised, the authors must have in mind the presentation of one's self to oneself. But they do not explain why the explicit measures themselves would not have this effect to a sufficient degree to shape one's responses to them even without having taken the implicit association test.

judgments as varying with some details of the situation. What the attitude strength and moral choice blindness experiments suggest, however, is that those of our explicit moral judgments that express firmly held attitudes are thereby consistent across situations, whereas others are constructed when needed from resources that vary across situations. Moreover, the strong attitudes that manifest in consistent judgments also manifest in consistent behaviour. So the moral task these experiments pose is neither one of bringing behavioural dispositions into line with evaluative beliefs nor one of undertaking strategies to ensure careful deliberation in morally important situations. It is to ensure that one holds the right moral attitudes sufficiently strongly that one's judgments and actions will express them consistently.[4]

6. The Negative Programme of Moral Improvement

How should one aim to ensure that one's moral attitudes are not only correct but also sufficiently firmly held to manifest in consistent judgment and behaviour? What practical ethical guidance is the best response to the cognitive structures that explain situational manipulation and implicit bias? One kind of response would be to prescribe a negative programme of moral improvement. The aim would be to identify the features of situations that lead one to judge and act in morally problematic ways, then undertake strategies to prevent these features from having this malign influence. This is a negative programme because it aims to negate the morally negative influence of particular aspects of situations. Such recommendations are not uncommon as responses to both situationist experimental results and the implicit bias experiments.

One form of this response recommends that we simply avoid situations that might lead us to behave badly (Doris 2002: 147-8). This might be sage advice for some kinds of situation, but the scope for this kind of control is clearly very limited. A more promising form of this response is the converse recommendation to preserve and promote the features of situations that support

[4] This is the idea of virtue ethics and situational variation ascribed to Plato, Aristotle, and the Stoics by Rachana Kamtekar (2004: 277-286).

morally desirable behaviour (Kamtekar 2004: 490-1). For example, if one's behaviour towards a particular ethnic group is biased by the presence of strong negative associations in one's cognitive system, then one can alter one's environment in ways that are likely to weaken these negative associations or strengthen more positive ones to counteract them. The efficacy of such a strategy is strikingly illustrated by 'the Obama effect': implicit measures of white people's cognition in response to images of black people found a significant decrease in evidence of negative associations as a result of the widespread media coverage of Barack Obama's first presidential campaign (Plant et al 2009).

Other forms of the negative programme dispense with the idea of managing one's situation and instead focus on shaping one's cognitive system more directly. One such strategy is to aim to alter one's pattern of cognitive associations with some particular situational feature through regularly forming the desired associations in action or in conscious imagination (Mischel and Shoda 1995: 261; Snow 2006: 556, 560). Alternatively, one can formulate and rehearse intentions to behave in a particular way in response to particular situations (Mischel and Shoda 1995: 261; Kamtekar 2004: 487-8; Besser-Jones 2008: 328-9). In terms of the cognitive-affective personality system, these strategies aim to forge and strengthen particular pathways through the architecture that generates judgments and behaviour, so that when the relevant situations arise one's cognitive system tends towards producing the outcomes that one has trained it to produce.

However, this negative programme might seem inordinately demanding. A very wide variety of subtle situational cues can influence our judgments and actions. Although the experimental literature on implicit bias tends to focus on responses to women and to black men, for example, there is no reason to assume that such biases are limited to these categories. The negative programme of moral improvement should include all the biases relating to the full range of ethnic and religious identities we encounter. Moreover, there is evidence of widespread bias concerning an individual's height and their body shape. It seems plausible that there are further biases concerning aspects of social background indicated by a speaker's accent. Once all of these

are taken into account, the negative programme seems rather daunting. But it may be even more so, since there remains the question of how these biases interact. Must a white person's bias concerning Oriental women, for example, simply be a function of distinct biases concerning Oriental people and concerning women? Or is it a specific bias that would require its own strategy for overcoming? The same question arises about the interplay of these biases and the classic situational manipulations. Do ethnicity and gender feature in our perceptions of authority figures or passive bystanders in the same way that they feature in our perceptions of students in a seminar room?

Even if these issues were to be resolved satisfactorily, so that a full set of strategies could be formulated for counteracting one's most morally problematic implicit biases and situational weaknesses, and even if it were accepted that this set of strategies was not overly demanding, then there would still remain the further question of how one can ensure that this programme of strategies will actually be carried through. This is most obvious in the case of the strategy of avoiding certain situations. The problem here is not that one might find oneself in such a situation for reasons beyond one's control. It is rather that when one is making a decision that determines whether one enters that situation, one might at that point be subject to situational influences or implicit biases. On a larger scale, a full programme of strategies for counteracting unwanted influences is itself an intended sustained pattern of behaviour that seems vulnerable to the kinds of influence it is intended to counter. The negative programme, that is to say, seems to preserve a vestige of the idea that our behaviour manifests our reflectively endorsed beliefs. We need to take more seriously the finding that simply deciding to adopt a certain mode of behaviour, or to pursue a range of strategies, is not enough to ensure that we actually do so. Our strategies can be derailed by the influence on our cognition of the subtle features of our situations, of our biases concerning the people we deal with, and of the demands of our everyday lives.

7. The Positive Programme of Moral Improvement

How can we deal with the problems of situational manipulation and implicit bias without our strategies for doing so being undermined by the same aspects of cognition that produce these problems? The findings of attitude psychology afford an answer to this question. To pursue some behavioural strategy, it is important that one habituate in oneself a sufficiently strong attitude in favour of that strategy. One's judgments about whether to expend effort in pursuit of this strategy will remain consistent across situations only if they manifest a firmly held attitude. If they are merely constructed out of the most accessible relevant beliefs and desires each time, they will vary with the relative accessibility of one's mental states, which in turn will vary with the situation one is in, the situations one has recently experienced, and one's recent cognitive activity. Indeed, these factors will determine whether one even consciously thinks of the strategy on a given occasion when it could be pursued.

Once we see this problem in terms of attitude strength, however, it becomes clear that the negative programme is not the only way to address the practical problems of situational manipulation and implicit bias. For the influence that a strong attitude has on cognition and behaviour is not restricted to situations in which some feature of the environment is directly related to the attitude. Attitude strength determines the attitude's general degree of accessibility to cognition. The stronger it is, the more accessible it is, the greater its influence over cognition generally. For this reason, a strong attitude can shape cognition and behaviour in ways that are not responses to the manifest features of the situation (Webber 2013: § 4). In order to reduce one's susceptibility to situational manipulations and implicit biases, therefore, one can aim to instil in oneself a few firmly held moral attitudes, such as attitudes in favour of fairness or against discrimination. With a sufficiently high degree of accessibility, these general attitudes should serve to counteract the influence of situational manipulations and implicit biases. Unlike the negative programme of moral improvement, the aim of this programme would be to identify and

embed in one's cognition the attitudes that tend one towards the right behaviour. This is a positive programme of moral improvement, which relies on the holistic nature of the cognitive-affective personality system.

This proposed programme of moral improvement echoes Aristotle's account of habituation. Character traits develop, according to Aristotle, through critically reflective practice. This habituation serves two ethical purposes. One is that it embeds one's values into one's behavioural cognition sufficiently firmly that one will judge and act in ways that manifest those values even in the face of temptations to do otherwise (NE: 1152a25-33). The other is that it refines one's understanding of the nature and demands of those evaluative commitments. One learns what justice really is and what it requires through repeatedly thinking about what justice requires in particular situations (NE: 1104a5-10). In terms of attitude psychology, the first of these purposes is served by strengthening the attitude and the second by refining the set of mental states that compose that attitude. In the context of the cognitive-affective personality system, these are processes of strengthening the associative connections between the relevant mental states, connections that strengthen each time they are used. The positive programme of moral improvement, therefore, would carry out the Aristotelian recommendation of habituating the ethical virtues.[5]

At present, there is no direct empirical confirmation of this theoretical prediction that one can reduce susceptibility to situational manipulation and implicit bias by habituating the requisite attitudes. Research into situational effects has developed independently of attitude psychology. Research into prejudice and stereotyping began in the context of attitude psychology, but soon

[5] Peggy DesAutels (2012) also suggests a positive programme of counteracting implicit biases. Her suggestion is grounded in a different area of cognitive science. Whether her account of the generation of implicit bias is compatible with the one detailed in this paper is too complicated a matter to be properly addressed here.

developed independently of the main path of development of attitude psychology.[6] However, there are experimental findings that support confidence that such empirical confirmation could be found. Research into the efficacy of goals is particularly suggestive in this regard, because goals share with attitudes the feature of having a dimension of strength or firmness as well as a dimension of content.

In one such experiment, subjects were shown a series of words, each preceded by a picture of a woman or a man, and asked to pronounce the word. If a picture of a woman had activated concepts stereotypically associated with women, then the subject should be able to pronounce any stereotypical words more quickly. Subjects had been tested for whether they strongly held the goal of treating women equally with men. The experiment found that a strong goal of egalitarianism towards women prevented the activation of stereotypical associations in response to pictures of women. Moreover, the speed at which the subjects responded indicates that the egalitarian goal had this effect without the subject's conscious intent (Moskowitz et al 1999).

This goal of egalitarianism towards women is more general than the goal of treating women equally with men if they are of a particular ethnicity, or a particular range of body shapes, or in particular social roles or situations, but less general than the positive programme recommends. Perhaps at least some of the subjects held this goal because they were committed generally to egalitarianism, but the experiment did not test for this. A more recent experiment tested the effect on stereotyping of a broader egalitarian goal, defined as 'treating people equally regardless of their ethnicity, gender, race, physical appearance'. This experiment did not measure how firmly each subject was already committed to this value, but manipulated the subjects so that this value became temporarily highly accessible for some, much less accessible for others. It found that subjects for whom this general egalitarianism was highly accessible did not make stereotypical associations in response to images of black people, whereas those subjects for whom this value

[6] For the history of this divergence and a detailed argument in favour of an reintegrating attitude psychology and the psychology of prejudice and stereotyping, see Maio et al 2010.

was not highly accessible did. Again, the speed of the experiment makes it impossible for this stereotype inhibition to be the result of conscious thought (Moskowitz and Li 2011).

Precisely how the idea of a goal employed in these experiments is related to the conception of an attitude emerging from attitude psychology is a matter that cannot be properly addressed here. Nevertheless, these results do suggest that direct empirical support could be found for the theoretical prediction that strengthening the right attitudes reduces the automatic activation of problematic associations. The design and analysis of such experiments would need to be grounded in an integrative conceptual approach to the findings of these divergent research areas. Debates over the ethical implications of situational manipulation and of implicit bias, and indeed over the prospects for virtue ethics more generally, would benefit greatly if such experiments concerning the positive programme of character development were to be undertaken.[7]

[7] This paper was developed through presentations at the 'Ethics and the Architecture of Personal Dispositions' conference at the Sorbonne and at a workshop of the Leverhulme 'Implicit Bias and Philosophy' project at University of Sheffield, both in July 2012. I am grateful to the organisers and participants of those conferences for discussion. I am also grateful to Alberto Masala for comments on an earlier draft.

References

Aristotle. NE. *Nicomachean Ethics*. Translated by Christopher Rowe. Introduction by Sarah Broadie. Oxford: Oxford University Press, 2002.

Axsom, Danny and Joel Cooper. 1985. Cognitive Dissonance and Psychotherapy: The Role of Effort Justification in Inducing Weight Loss. *Journal of Experimental Social Psychology* 21: 149-160.

Besser-Jones, Lorraine. 2008. Social Psychology, Moral Character, and Moral Fallibility. *Philosophy and Phenomenological Research* 76: 310-332.

DesAutels, Peggy. 2012. Moral Perception and Responsiveness. *Journal of Social Philosophy* 43: 334-346.

Doris, John. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Greenwald, Anthony, Debbie McGhee, and Jordan Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74: 1464-1480.

Hall, Lars, Petter Johansson, and Thomas Strandberg 2012. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *Plos One* 7: e45457.

Holland, Rob W., Bas Verplanken, and Ad van Knippenberg. 2002. On the Nature of Attitude-Behavior Relations: The Strong Guide, The Weak Follow. *European Journal of Social Psychology* 32: 869-876.

Kamtekar, Rachana. 2004. Situationism and Virtue Ethics on the Content of Our Character. *Ethics* 114: 458-491.

Maio, Gregory, Geoffrey Haddock, Russell Spears, and Antony Manstead. 2010. Attitudes and Intergroup Relations. In *The Sage Handbook of Prejudice, Stereotyping and Discrimination*, edited by John Dovidio, Mies Hewstone, Peter Glick, and Victoria Esses. London: Sage.

McConnell, Allen, and Jill Leibold. 2001. Relations Among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *Journal of Experimental Social Psychology* 37: 435-442.

Merritt, Maria. 2009. Aristotelian Virtue and the Interpersonal Aspect of Ethical Character. *Journal of Moral Philosophy* 6: 23-49.

Merritt, Maria, John Doris, and Gilbert Harman. 2010. Character. In *The Moral Psychology Handbook*, edited by John Doris and the Moral Psychology Research Group. Oxford University Press.

Milgram, Stanley. 1974. *Obedience to Authority: An Experimental View*. New York, Harper and Row.

Moskowitz, Gordon, Peter Gollwitzer, Wolfgang Wasel and Berndt Schaal. 1999. Pre-conscious Control of Stereotype Activation Through Chronic Egalitarian Goals. *Journal of Personality and Social Psychology* 77: 167–184.

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: 246-268.

Moskowitz, Gordon, Peter Gollwitzer, Wolfgang Wasel, and Bernd Schaal. 1999. Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals. *Journal of Personality and Social Psychology* 77: 167-184.

Moskowitz, Gordon, and Peizhong Li. 2011. Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control. *Journal of Experimental Social Psychology* 47: 103-116.

Plant, E. Ashby, Patricia Devine, William Cox, Coey Columb, Saul Miller, Joanna Gople, and Michele Peruche. 2009. The Obama Effect: Decreasing Implicit Prejudice and Stereotyping. *Journal of Experimental Social Psychology* 45: 961-964.

Russell, Nestar. 2011. Milgram's Obedience to Authority Experiments: Origins and Early Evolution. *British Journal of Social Psychology* 50: 140-162.

Snow, Nancy. 2006. Habitual Virtuous Actions and Automaticity. *Ethical Theory and Moral Practice* 9: 545-561.

Shoda, Yuichi and Walter Mischel. 1996. Toward a Unified, Intra-Individual Dynamic Conception of Personality. *Journal of Research in Personality* 30: 414-428.

Webber, Jonathan. 2013. Character, Attitude and Disposition. *European Journal of Philosophy*. DOI: 10.1111/ejop.12028