

*Implicit Bias And Moral Responsibility: Probing The Data.*¹

In the laboratory, people who profess a commitment to or a belief in racial equality can be induced to behave in ways that strongly suggest that they harbor mental states that conflict with their sincere avowals. For instance:

They may be quicker to associate pictures of black people with negative words than with positive, and vice-versa for pictures of white people (Greenwald, McGhee & Schwartz 1998; Nosek, Greenwald & Banaji 2007; Greenwald et al. 2009).

Presentation of black faces as primes may cause them to judge that neutral stimuli presented immediately afterwards are less aesthetically pleasing than neutral stimuli presented immediately after white faces (Payne et al. 2005).

They are more likely to judge that an ambiguous object held by a black person is a gun than the same object in the hands of a white person (Payne 2006).

And so forth. On the basis of this kind of data, most psychologists conclude that ordinary people have two different kinds of mental representations. In addition to their *explicit attitudes*, which are the attitudes they avow and self-attribute, they have *implicit attitudes*. Implicit attitudes are automatically activated, difficult to inhibit and insensitive to agents' explicit attitudes.²

Outside the controlled conditions of the laboratory, implicit attitudes play a smaller role in behavior than within, because the experimental setting is designed to ensure that explicit attitudes are less able to exert their influence. But there is good evidence that implicit attitudes explain some of the variance in behavior, even of people with conflicting explicit attitudes. The predictive power of implicit attitudes is disputed (Greenwald et al. 2009; Oswald et al. 2013). For some kinds of subtle behaviors – say the distance from another person someone will choose to sit – implicit attitudes seem to be better predictors than explicit (McConnell & Leibold 2001). Even on the most conservative estimates, implicit attitudes will modulate or cause behaviors sufficiently often that the average black person in the United States can expect to be affected negatively thousands of times across their lifetime (Greenwald, Banaji & Nosek 2016). Most of these incidents will be trivial (less eye contact, less welcoming body language, and so on) but their cumulative effects will not be trivial. Moreover, it is overwhelmingly likely that implicit attitudes bias information processing in ways that ensure that more

momentous decisions are sometimes affected. Implicit attitudes almost certainly play a role in explaining why people often prefer white, or male, job candidates over equally qualified black, or female, candidates: why, for instance, they are more likely to offer interviews to candidates with white-sounding names (Dovidio and Gaertner 2000).

People with unprejudiced explicit attitudes and conflicting implicit attitudes will sometimes perform morally significant actions that they would have avoided had their explicit attitudes controlled their behavior. These actions may be said to have a moral character that is due to their implicit attitudes: its morally significant features (that it is racist or sexist; that it disregards the welfare of another person, and so on) are dependent on that attitude (that is, the following counterfactual is true: *had the agent's explicit attitudes controlled their behavior, the action would not have had that character*). Is it appropriate to hold these agents morally responsible for this class of actions? This question has recently become the focus of lively philosophical debate (Holroyd 2012; Levy 2014b; Faucher 2016; Glasgow 2016; Washington & Kelly 2016; Zheng 2016). We seem to have limited control over our implicit attitudes and over the ways in which they influence our behavior. We may lack insight into their existence and their content. At first glance, these facts may seem to greatly reduce or eliminate our moral responsibility for the actions they (partially) cause. But matters are more complex than they look at first glance. There are good reasons to avoid setting the bar for control over and insight into our own attitudes too high: we routinely lack insight into the nature of the processes that guide our most skillful actions and may have a limited capacity to shape our responses.

In this paper, I aim to advance the debate over whether agents are morally responsible for actions which owe their moral character to their implicit attitudes via close attention to the nature of the processes and mechanisms that underlie or realize these attitudes. This approach to the question reverses the standard method in debates over moral responsibility. The standard approach attempts to solve for two variables simultaneously: attempting to ascertain the truth regarding the cases under consideration, but also taking the intuitions generated by considering the cases as data, in the light of which one's theories may be revised. Following this approach, implicit attitudes become more grist for our theory-building mill. The approach I advocate here is more mechanical. I aim to assess the degree to which actions caused by implicit attitudes satisfy plausible necessary conditions for moral responsibility, setting aside the intuitions that might be generated by

the cases I consider. I take this approach because I suspect (and will argue) that implicit attitudes are states that differ significantly from those that feature in folk psychology. They have properties that are highly counterintuitive, by folk lights.³ Since our intuitions about moral responsibility track folk psychological states, our responses to cases involving implicit attitudes are likely to be off track. We do better to test our theories of moral responsibility in the light of more standard cases, and simply apply them to cases involving more outré entities and processes.⁴

Of course, it might turn out that our intuitions with regard to these cases are not off track. Suppose our intuitions are generated by mechanisms that treat implicit and explicit attitudes identically, and implicit and explicit attitudes are actually similar enough to vindicate these responses. The method I advocate should be able to reveal this fact. It involves close attention to the nature of the psychological mechanisms and processes that underlie or realize implicit attitudes. If they are similar enough to beliefs to justify treating them alike, then a mechanical application of our theory of moral responsibility to these cases will treat them in this way. It will not rely on intuitions, given the genuine possibility that they are off track, but it will generate the same results as the standard methodology. The approach advocated is epistemically more secure, even if it actually turns out to do no better, than the intuition-based standard methodology.⁵

Before I turn to putting this method into practice, it is worth saying a few words about the kind of responsibility I have in mind. What is at issue here is agents' *direct* responsibility for actions partially caused by implicit attitudes. Assessing agents' *indirect* responsibility requires answering different questions, and the two kinds of responsibility can dissociate. For instance, an agent may lack direct responsibility for an action caused by their implicit attitudes, because *given what her implicit attitudes were at t*, it would not be reasonable to expect her to control her behavior, or to recognize its moral significance, or what have you. But the agent might nevertheless be fully morally responsible for the behavior, because it was reasonable to expect her to try to change her implicit attitudes prior to *t*. Indirect responsibility can underwrite a great deal of praise and blame; it is not necessarily a lesser kind of responsibility.

Much of the existing literature on responsibility for actions partially caused by implicit attitudes has focused on indirect responsibility. For instance, Holroyd (2012) surveys a

range of evidence that supports her claim that agents can exercise some degree of control over the possession, or the manifestation, of implicit attitudes. There are a number of different strategies people may utilize to alter the content, or to inhibit the influence, of implicit attitudes. Counter-stereotypical priming has been shown to be effective at modulating implicit attitudes in the short term. People may formulate implementation intentions (“when a woman is speaking, I will listen attentively”) to counteract their effects. Sustained exposure to counter-stereotypical individuals may alter implicit attitudes in a way that is broad and long lasting (Dasgupta 2013). All of this evidence indicates that it is possible to affect the content and influence of implicit attitudes, and under a variety of conditions these facts may render agents indirectly responsible for actions they cause, in virtue of prior actions. If I could have reduced my gender bias but did not, I might be responsible for its expression in virtue of that earlier failure.

While the question of indirect responsibility is an important one, it cannot entirely replace the question of direct responsibility. On the most optimistic story concerning our capacities for control over implicit attitudes and over their expression, even conscientious and well-informed agents utilizing the best strategies for controlling their implicit attitudes may rarely succeed entirely in bringing them under control. There will still very likely be a gap between the explicit and implicit attitudes of such agents, and there will be circumstances in which their behavior will vary from what it would have been had there been no such gap. If that’s right, then indirect responsibility can’t do all the work we want, even under ideal circumstances, and the question concerning our direct responsibility remains an important one.

Finally, it is worth remarking on what I mean by the phrase ‘moral responsibility’. Different philosophers use the phrase to refer to different kinds of assessments of agents. I do not wish to enter into the debates concerning which (if any) of these kinds of assessment deserves to be identified with the core or central meaning of the phrase. Instead, I shall simply stipulate what I mean by it. To hold an agent morally responsible for an action is to be committed to accepting that in virtue of their having performed that action they are *pro tanto* appropriate targets of certain sorts of responses that benefit or burden them. Such responses might range from demanding that they provide a justification of their behavior through to punishment, from a grateful smile through to honors and rewards. To hold that such responses are *pro tanto* appropriate is to hold that

they are deserved (the centrality of desert to the account marks it as belonging to the family of accounts of moral responsibility centred on what Pereboom (2001) calls *basic desert*; see Pereboom (2013) and McKenna (2013) for discussion). This does not entail that the praise- or blameworthy ought, all things considered, to be responded to in any or all of these ways: perhaps there are other considerations which entail that we ought sometimes or even always to refrain from such behavior. Defeating considerations of the kind I have in mind entail that we ought not to *blame* (or praise); they do not affect whether or not the agent is *blameworthy*.

Control and Attributability

In what follows, I will focus on two conditions that, separately or together, have very widely been identified as necessary for moral responsibility. Almost every prominent theorist accepts some version of either (1) or (2):

- (1) *Control*. An agent is morally responsible for an action or for the consequences of an action only if she exercised ‘freedom-level’ control over that action or that consequence. It is this condition, of course, that is at the heart of the free will debate.⁶

- (2) *Attributability*. An agent is morally responsible for an attitude or an action only if it is appropriately attributable to the agent. Whereas the control condition is typically advanced by its proponents as a *central* condition for moral responsibility, attributability is sometimes held to be necessary but not central. That is, some accounts that maintain that attributability is a necessary condition of moral responsibility would not appropriately be described as attributability views in virtue of this fact. Nevertheless, the fact that they accept such a condition means that for some purposes we may treat them together. This disparate family of views includes those which Smith (2008: 368) calls “updated versions [...] of ‘real self views’”, on which responsibility requires that the action or attitude belongs to the agent’s real self (or is caused by states that so belong). It also includes quality of will views, like that of Arpaly (2003), on which agents are responsible for actions that express their intrinsic desire to act for moral reasons or their regard for the moral standing of other agents; on views like this,

the actions express the agent's quality of will because they are caused by states that can appropriately be attributed to the agent.

I aim to answer the following questions:

1. Do agents exercise a sufficient degree of control over actions that have a moral character due to their implicit attitudes to appropriately be held morally responsible for them?
2. Are actions with a moral character due to implicit attitudes caused by states that properly belong to the agent?

Control

An influential family of accounts of moral responsibility maintain that direct responsibility for an action or the consequences of an action requires that the agent possess a sufficient degree of control over that action or the consequence. One obvious way in which implicit attitudes seem to threaten moral responsibility is by decreasing the degree of control agents exercise over their actions and over the consequences of their actions. In what follows, I aim to demonstrate that in the class of cases which are my concern here, implicit attitudes decrease our control to such a degree that we lack what Haji (2012) calls 'responsibility-level control' over them and their consequences. Schematically, I will argue as follows (with premise 1 being assumed for the sake of the argument):

- (1) Moral responsibility requires that the agent exercises responsibility-level control over their action or the consequences of their action (depending on whether they are putatively responsible for the action or for its consequences);
- (2) In cases in which an action (or its consequences) have a moral character due to the agent's implicit attitudes, control over the action (or its consequences) is greatly diminished.

- (3) The decrease in control is significant enough to make it highly plausible that the agent lacks responsibility-level control.

If control is a necessary condition of direct moral responsibility, agents are therefore not responsible for these actions or their consequences.

As we have seen, there is ongoing debate about what proportion of behavior is predicted by implicit attitudes. Fortunately, this is not a debate into which I need to enter. I am interested only in those cases in which condition (2) is satisfied: when actions or their consequences have a moral character due to an agent's implicit attitudes. This condition is satisfied when the following counterfactual obtains: *had the agent's explicit attitudes alone controlled their behavior*, the action would have lacked its actual moral character. For instance, it would not be true, of a choice we can appropriately describe as sexist, that it is sexist. These are the relevant cases because our interest is in whether agents who are sincerely opposed to (say) sexism ought to be blamed when they act in ways that - due to the influence of their implicit attitudes - can appropriately be described as sexist. It is these conflict cases that are my concern, and - I will argue - it is in these conflict cases that agents lack responsibility-level control. The empirical debate concerning what proportion of behavior is predicted by implicit attitudes concerns how often (2) is satisfied, not the degree of control agents possess when it is satisfied. Just how frequently this occurs remains an open empirical question, but there is a great deal of - largely indirect - evidence that it sometimes occurs (this evidence, reviewed above, consists in the data that implicit and explicit attitudes diverge and data showing that implicit attitudes explain some variance in behavior; together, these two bodies of data strongly suggest that implicit attitudes sometimes cause behavior that has a moral character that conflicts with the character the action would have had were the behavior controlled by agents' explicit attitudes). The hard work in what follows will consist in showing that (3) obtains when (2) is satisfied: that agents lack responsibility-level control in these conditions.

At least typically (and perhaps always) implicit attitudes prevent what I will call *personal-level* control over the character of our actions, when that character is due to the implicit attitudes (when it is sexist, or racist, or what have you, and it is the agent's implicit attitudes which explain why it has that character). Personal-level control is deliberate and

deliberative control; it is control exercised in the service of an explicit intention (*to make it the case that such and such*) (Shepherd 2014).⁷ Personal-level control over the moral character of our actions requires that that moral character features (though not necessarily under that description) in our explicit intentions. This kind of control has some very demanding epistemic conditions, which agents routinely fail to satisfy in the cases at issue.

Personal-level control over the character of actions is typically blocked by our lack of awareness that our actions have that character. Because implicit attitudes operate below the level of conscious awareness, agents often confabulate what seem to them good (personal-level) reasons for actions partially caused by their implicit attitudes. Both the character of the action and the confabulation is caused by implicit attitudes in these cases (see Levy 2014a for discussion). Even if we know, on occasion, that we have implicit attitudes that are or might be relevant to a particular decision we have to make, for us to possess personal-level control over the character of that decision we must be able to detect how these attitudes influence our information processing, and then hit upon a method of modulating or inhibiting this influence. The causal processes whereby implicit attitudes modulate behavior and decision-making are opaque to introspection, ensuring that we lack insight into what influence they have on our perceptions and judgments, and that there are no reliable means of modulating or inhibiting this influence. Of course, there are things we can do to prevent implicit attitudes producing actions with characters we disapprove of. We can blind ourselves to the gender of applicants for jobs, we can utilize implementation intentions, and so on. But these are all things we can do *beforehand*, and if we are responsible in virtue of doing, or failing to do these things, our responsibility is indirect. Implicit attitudes always or almost always block *direct* personal-level control over the character of actions when that character is due to these attitudes.

But personal level control is a very demanding kind of control, and very plausibly agents may be morally responsible for actions if they exercise some much less exalted kind of control over them. Personal level control actually *requires* less exalted forms of control. Consider the exquisite control exercised by a skilled violinist over her performance. The musician may have an explicit intention, in the service of which she exercises personal-level control. Perhaps she intends *to articulate the 16th notes*, or *to play the largo with passion*. As an expert, she can attend to these high-level features of her performance. At the same

time, though, she exercises control over the movements of her bow and over the strings, even though the content of her intention is silent about these aspects of her performance (were she to attend to these lower-level features of her movements, her performance might significantly degrade; Beilock and Carr 2004). The improvising jazz musician might not even have an intention that refers specifically to the content of the music: that is, to the notes she selects. Her intention might be *to dialogue with the piano*. One sign of this is, notoriously, she can be surprised by her own note selection. This kind of phenomenon is an everyday occurrence even for those of us who lack the kind of expertise demonstrated by the highly skilled musician: we are nevertheless virtuosos in other spheres, and in those spheres we can exhibit the same combination of mastery and capacity to surprise ourselves. People may be surprised by their own witty remarks, for instance. Speaking fluently, playing a musical instrument, passing a football, and so on are, when done by people with the requisite skills, instances of exquisite control. But they are instances of personal-level control only at high levels of description. The conversationalist may intend *to make a witty remark*, but not intend to make *that particular pun*. Rather, her control at a finer-grained level of description consists in her sensitivity to the musical, or the conversational, or the sporting, demands and contexts of her contribution. This sensitivity is subpersonal, realized by mechanisms to which she has (at best) imperfect access.

What makes it the case that the witty remark and the movements of the bow are exercises of control? As Fischer and Ravizza (1998) have influentially argued, control consists in sensitivity to reasons: an agent or a mechanism possesses control over a state of affairs when it is capable of recognizing reasons *as* reasons to modulate that state of affairs and would so modulate it in response to some of those reasons. Fischer and Ravizza describe the former, the capacity to recognize reasons, as receptivity, and the latter as reactivity (note, though, that control over a state of affairs requires the sensitivity of that state to the actions of the mechanism or agent; reactivity therefore requires such sensitivity). The witty remark is an instance of control (if it is) because it is sensitive to conversational context and demands. Had the conversation turned personal, the remark would have been inhibited because it was now inappropriate, for instance. The musician's bowing is an instance of control because it is sensitive to the contours of strings, the pressure her other hand is exerting, the volume of the orchestra, and so on: relatively small changes in these conditions would have led to adjustments in her bowing.

As Fischer and Ravizza also emphasize, this sensitivity to reasons must be appropriately *patterned*. They say relatively little about what being patterned consists in, beyond emphasizing that responsiveness like this exhibits “a minimally comprehensible pattern” (1998: 73); their project requires no more detail than that. My project requires a little more, which I now provide.

Sensitivity is patterned, I will say, when (perhaps *inter alia*) it is *continuous*, *broad* and *systematic*. To say that it is *continuous* is to say that the relevant mechanism is sensitive to relatively fine-grained alterations in the parameters of a particular reason. The violinist exhibits continuous sensitivity to orchestral dynamics, say, when she would respond not merely to *some* alteration in the volume of the orchestra by adjusting her own volume, but when she is appropriately responsive to a (relatively) continuous dynamic range. To say that sensitivity is *broad* is to say that the relevant mechanism is responsive not just to a particular kind of reasons (however continuously) but to a range of different kinds of reasons. The violinist exhibits exquisite control inasmuch as she is sensitive not only to orchestral dynamics, but to the acoustics of the room, the mood of the conductor and the audience, and so on. Sensitivity is *systematic* when the mechanism would respond to a particular kind of reason in any context (so long as that context does not include features that neutralize the reason). The violinist exhibits systematic responsiveness to dynamics when she would respond to them in a large hall or a small, with a full orchestra or a small ensemble, and so on.

Control is not all or nothing; agents may possess more or less control over their behavior. If control is a necessary condition of moral responsibility, agents must have *sufficient* control over their behavior to be responsible for it. The more continuous, broad and systematic their control, the greater their degree of control. Contrast the controlled arm movements of the musician with the arm movements of the person reflexively jerking back from an unexpectedly hot stove. The latter movement is reasons responsive in a very minimal way: the heat of the stove constitutes a reason to rapidly withdraw the limb. But the mechanisms that cause the movement exhibit very little in the way of patterned responsiveness. The mechanism would cause the jerk in precisely the same way (or in ways that differ from case to case stochastically, rather than in response to reasons) no matter what reasons there may be for modulating the movement or refraining from making it. A person might be fully aware that there is a priceless vase just behind her and

nevertheless jerk back in the same way, with predictable consequences. This is a failure in *breadth*: the mechanism is responsive to reasons only of a narrow kind. Very plausibly, the agent lacks sufficient (direct) control over her movements to count as morally responsible for breaking the vase in virtue of this failure of breadth.

While control requires patterned sensitivity to reasons, it does not require sensitivity to *every* reason (as, again, Fischer and Ravizza recognize: in their terms, control requires *moderate*, not *strong* reasons-responsiveness). An agent (or a mechanism) exercises control when she is *sufficiently* responsive to a broad enough range of reasons. Building on Fischer and Ravizza's account, we might say that a representation, or its realizer, features in responsibility-level behavior when it is a component of a mechanism that is moderately reasons-responsive. This account generates the following result: if an action is caused by mechanism *m*, and *m* is sensitive to reasons in a suitably patterned way, then *m* realizes control over that action. I now turn to the data concerning how implicit attitudes affect reasons-responsiveness.

Reasons-Responsiveness and Implicit Attitudes.

On what has a strong claim to be described as the standard account of implicit attitudes, they are *associations* between concepts or representations. On this picture, roughly, a relevant stimulus activates or makes accessible a set of representations, or dispositions to feel and behave, which are associated with that stimulus. One representation is associated with another as a consequence of their co-occurrence in the learning history of the agent (Lee 2016). To use a hackneyed example, “pepper” is associated in this way with “salt”, because we are often exposed to the phrase “salt and pepper” and because salt and pepper are commonly placed together on restaurant and dining tables. On the associative account, a stimulus will activate or make accessible others regardless of what the agent believes (even unconsciously). “Table” may be more strongly associated with “chair” (due to features of our learning history) than with “bench”, yet we may believe that tables are more similar to benches than they are to chairs. Similarly, priming a person with a black face might activate or make accessible representations or dispositions associated with violence (due to the prevalence of cultural stereotypes) without the person believing that black people are more likely to be violent than members of other races.⁸

The associative account entails that implicit attitudes will display little reasons-responsiveness. They are sensitive to cues with which they have been associated in the agent's learning history, not to (justificatory) reasons: the implicit representation "salt" might increase the accessibility of "pepper" but won't be able to feature in inferences from propositions like "excessive salt causes high blood pressure" to conclusions like "I should cut my salt intake". If this account is correct, then the fact that a mechanism has an implicit attitude as a component will reduce its reasons-responsiveness. Of course, that mechanism will remain responsive to reasons that fall outside the domain of the attitude, but insensitivity to reasons within that domain seems to be exculpating (though we shall have occasion to revisit this question): if an action has a moral character due to an implicit attitude, the fact that the mechanism wasn't responsive to reasons that bear on that moral character seems to very greatly reduce – even eliminate – moral responsibility. On this account, it seems that we have insufficient control over the moral character of our actions, when this character is due to implicit attitudes, for us to be morally responsible for their having that character.

However, several philosophers and psychologists have recently suggested that implicit attitudes are apt to feature in bona fide inferences (Egan 2011; Mandelbaum 2013; Mandelbaum 2016; De Houwer 2014). Mandelbaum (2016) has surveyed a large body of experiments that he suggests involve implicit attitudes behaving like bona fide beliefs. Consider, for instance, the phenomenon of 'celebrity contagion' (Newman, Diesendruck and Bloom 2011). Ordinary people have positive implicit attitudes toward (some) celebrities, and these attitudes cause them to value everyday objects that have been in contact with those celebrities: people are willing to pay more for a sweater that has been worn by George Clooney. This willingness can neatly be explained associatively: the contiguity between the sweater and Clooney brings it about that his celebrity somehow 'rubs off on' the sweater. But telling subjects that the sweater has been laundered reduces the price they are willing to pay for it. It is hard to explain this fact associatively. We do not have learning histories that associate 'celebrities' with 'laundry'; further, any associations we have with washing are likely to be *positive*, not negative. Explaining the lower valuation that subjects place on the laundered sweater apparently requires us to postulate some kind of inferential interaction between their beliefs about laundry and their attitudes toward celebrity. The magic of celebrity can be washed off, like mud.

The evidence that Mandelbaum and others have provided constitutes a strong case for the claim that implicit attitudes have (some) propositional structure. It is this kind of structure that explains why they respond inferentially, for propositional structure is required for an attitude to encode a particular relation between the constituents of the attitude, and it is relational information that is required to underwrite content-driven processing. On the basis of this kind of evidence, Mandelbaum argues that implicit attitudes are *beliefs*. If he's right, then mechanisms that have implicit attitudes as components will not exhibit any reduction in reasons-responsiveness in virtue of this fact. Beliefs are *inferentially promiscuous* (Stich 1978): capable of interacting with an open-ended range of other representational states in ways that are sensitive to their content. This inferential promiscuity entails that they are capable of being responsive to an open-ended range of reasons. If implicit attitudes are beliefs, there is no reason to think that they reduce reasons-responsiveness.

Insofar as implicit attitudes are beliefs, they will exhibit patterned reasons-responsiveness: beliefs interact with other states in the kinds of normatively respectable ways that constitute such reasons-responsiveness. But a mental state might feature in content-driven processing without being a belief: if it is insufficiently patterned in its responsiveness. Implicit attitudes have some propositional structure, but there is good evidence that they do not have the right kind of structure to underwrite continuous, broad and systematic responsiveness (and, accordingly, they lack the kind of propositional structure distinctive of beliefs).

I have reviewed the extensive evidence that I take to show that implicit attitudes are not beliefs elsewhere. Implicit attitudes, I have argued, have patchy propositional structure, not the kind of continuous and broad propositional structure we rightly associate with beliefs. They are, I have suggested, not beliefs but *patchy endorsements* (Levy 2015). Implicit attitudes are endorsements because they have sufficient propositional structure to have truth conditions, and therefore may properly be asserted, but this propositional structure is too patchy for them to count as genuine beliefs. Unlike *bona fide* beliefs, which are inferentially promiscuous because their propositional structure fits them to interact in the appropriate manner with other propositionally structured representations, implicit attitudes interact appropriately with such representations only sometime, and only with

some representations. From the patchiness of inferential relations, we can infer a matching patchiness of structure. I will now outline some of the evidence of patchiness, with an eye to how the patchiness undermines patterned reasons-responsiveness.

Consider, first, the finding that implicit attitudes bias decision-making to produce both a preference for a white or a male job candidate over a minority or female candidate, and also cause the confabulation of the qualifications needed for the job, with subjects justifying their choice by reference to the qualifications possessed by the favored applicant (Dovidio & Gaertner 2000; Uhlmann & Cohen 2005; Son Hing et al. 2008). These judgments exhibit a failure of systematicity: the agent (or the mechanism) fails to respond to the qualifications of the disfavored candidate in the particular context, though he is perfectly capable of recognizing the reason as a reason in other contexts. Here the behavior has the kind of moral character it seems appropriate to attribute to it (being sexist, for example) because it is partially driven by implicit attitudes that fail to recognize a consideration as a reason; conversely, were the behavior driven by the agent's explicit attitudes alone, it would have a contrasting moral character, because these attitudes are not insensitive to the reason-giving force of the relevant considerations. There is also evidence for spectacular failures of *breadth* in responsiveness. Consider Rozin (1990). In this experiment, subjects preferred drinks sweetened with sugar from a jar labeled "sucrose, table sugar" to those sweetened with sugar from a jar labeled "not sodium cyanide, not poison," despite having seen both jars filled from the same sack of sugar. Mandelbaum (2013) argues, against Gendler (2008), that the best interpretation of the data requires us to postulate some kind of content-driven processing. That interpretation is very plausible, but it entails that the relevant mechanisms failed to process the negation in the label "not poison". There is independent evidence that implicit processes are blind to negation (Wegner 1984; Deutsch, Gawronski & Strack 2006; Hasson & Glucksberg 2006). Being blind to negation, though, entails a genuinely *dramatic* loss in systematicity of response. For a very broad class of considerations, mechanisms that have as components implicit attitudes may fail to respond to these considerations as reasons and thereby cause actions that have a moral character they otherwise would not have had.

Mechanisms with implicit attitudes as components may exhibit complete insensitivity to particular reasons. Gregg, Seibt and Banaji (2006), for example, gave their subjects information about the members of two novel groups. Members of one did mainly

positive things while members of the others did mainly negative things. In one condition, subjects were told that there had been an error: the behaviors ascribed to each group had been accidentally reversed. Subjects reversed their explicit attitudes, but not their implicit attitudes. This is a spectacular failure of breadth: complete insensitivity to a conclusive reason. Perhaps worse, these mechanisms may exhibit *perverse* responsiveness, responding when they should not. Han et al. (2006) had children learn facts about a Pokemon character and then watch a video in which other children expressed beliefs about the character that were inconsistent with what they had learned. The subjects rejected the opinions expressed by the children in the video, but the knowledge that these opinions were false did not prevent them from altering their implicit attitudes. The implicit attitudes were indeed sensitive to information, but not in the way that beliefs are supposed to be.

The evidence briefly reviewed above demonstrates the patchiness of response to semantic content exhibited by implicit attitudes; since being responsive to reasons requires appropriate sensitivity to the inferential relations entail by semantic content, this patchiness indicates a drastic curtailment of patterned reasons-responsiveness. There is further evidence that indicates that implicit attitudes are involved in informational processes in ways that are not responses to semantic content at all. It strains credibility to think of the affect misattribution procedure (or sequential priming more generally) as involving responsiveness to reasons at all. Priming with a black face leads subjects to judge that a Chinese pictogram is less aesthetically pleasing, compared to neutral primes (Payne et al. 2005). Can this perspicuously be described as responding to reasons, as opposed to something more brute? A glance at the range of primes used in the affect misattribution procedure – seals, porpoises and money, for instance, for positive primes, and guns, ruins, and snakes, among others, for negative (Payne et al. 2005) – makes an inferential interpretation of sequential priming even more strained. These primes drive judgments of relative beauty of stimuli, but clearly they do not do so in virtue of their aesthetic qualities.

Mechanisms that contain as components implicit attitudes may fail to update in response to reasons, or update perversely, taking a reason that supports *a* as a reason against *a*. They seem to be blind to negations. There is good reason to think that these failures will, in at least some cases, be relatively general: a mechanism that fails to register a woman's

superior qualifications as supporting her candidacy will fail in a similar way in a wide range of counterfactual scenarios, more or less different from the actual case. Is this sufficient reason for us to conclude that these mechanisms are not moderately reasons-responsive and that in these cases the agent therefore lacks sufficient control to make him an apt target of moral responsibility?

But we can't simply conclude from the fact that implicit attitudes limit reasons-responsiveness that when their actions are counterfactually dependent on them, agents lack responsibility-level control. There are two reasons for this. First and most obviously, the limitation in control may be irrelevant in particular cases: the fact that I am insensitive to a particular kind of reason cannot reduce my responsibility for an action, when reasons of that kind are simply irrelevant to how I acted and how I ought to have acted. Second, even when a kind of reason to which the implicit attitude limits reasons-responsiveness is relevant, that fact may not undermine moral responsibility because diminished reasons-responsiveness may be sufficient reasons-responsiveness. As Fischer and Ravizza argue, moral responsibility requires only *moderate*, not *strong*, reasons-responsiveness.

Moderate reasons-responsiveness requires only *weak* reactivity to reasons (an agent must be capable of modifying her behavior in the light of *some* reason) and *regular* receptivity to reasons, where regular receptivity gives rise to a "minimally comprehensible pattern" of receptivity (73). The receptivity characteristic of patchy endorsements may be sufficient to give rise to such a pattern; and therefore to realize moderate reasons-responsiveness. However, in some circumstances, it seems, moderate reasons-responsiveness is insufficient for moral responsibility, *contra* Fischer and Ravizza: when the mechanism is insensitive to a kind of reason, and that insensitivity explains the moral character of the resulting action. That's the case in the circumstances at issue in this context. In cases like this, recall, the counterfactual *had the agent's explicit attitudes alone controlled her behavior, the action would have lacked that moral character* is always true, and that entails that her implicit attitudes *prevented* her from responding to one or more reasons. This kind of insensitivity entails a lack of control over that moral character: while her behavior is driven by the relevant mechanisms, the agent *cannot* respond to the relevant reasons. This kind of insensitivity to the reasons that explain the moral character of an action seems to entail a lack of control over that character (Vargas 2013a: 288).

In order for an agent to be morally responsible for an action, it is not sufficient that the mechanism causing his action is sensitive to a broad range of reasons; it had better be sufficiently sensitive to the kind of reasons that give to his actions their moral character. A systematic failure to be responsive to a particular kind of reason (a failure of breadth or of systematicity) is evidence of a lack of control with regard to that kind of reason; insofar as control is required for moral responsibility, this fact ought to be exculpating. We can't blame someone for the sexist character of their action, say, if they lacked control over the fact that it is sexist; not if control is a necessary condition of moral responsibility.

Consider, again, the agent who chooses a worse qualified male over a better qualified female because his implicit attitudes bias his selection processes. If he is morally responsible for the fact *that the choice is sexist* (and moral responsibility requires control), he had better be sensitive to the considerations that make it the case that his choice has that character. If in fact insensitivity to these considerations *explains* the moral character of his action, he seems to lack the degree of control required for moral responsibility. So if moral responsibility requires control over the character of the acts for which we are morally responsible, implicit attitudes routinely block moral responsibility for actions with a moral character due to these attitudes.

At the beginning of this discussion, we noted that implicit attitudes prevent agents exercising personal-level control over the character of actions they cause, when these actions have this character due to their influence. We now see that we lack even subpersonal control over this character. We lack this subpersonal control not due (directly) to a failure to satisfy epistemic conditions on control, but due to the relevant mechanisms' lack of appropriate sensitivity to reasons.

There seem to be two ways to block this conclusion. The first is to argue that though control is required for moral responsibility, control over the character of the act is not required: a broader kind of control is sufficient. That does not seem a promising route to take. It seems unprincipled to maintain that control underwrites moral responsibility, yet hold that agents can be morally responsible for actions over the crucial features of which they lack control. The second route is more promising. It is to deny that control is

necessary for moral responsibility. One might motivate the claim that control is not necessary by considering ordinary cases of agents who are so firmly wedded to particular values or commitments that they are very insensitive to considerations that might motivate acting contrary to them. We might think of Luther's declaration that he could "do no other", so firmly was he committed to his principles (Dennett 1984). Or we might think of the racist who is so firmly wedded to his nasty views that no ordinary evidence, no matter how voluminous and authoritative, could convince him of the moral and intellectual equality of races. Most people would not be disposed to excuse the racist if he acts on these views. Cases like this, and many others, have motivated some philosophers to reject the control condition in favor of some kind of what I am here calling an attributability view.⁹ On these kinds of views, agents are responsible for attitudes that properly belong to them, and for actions caused by such attitudes.¹⁰ I therefore turn now to these views.

Attributability

Proponents of the family of views I am here describing as attributability views of moral responsibility deny that control is a necessary condition of moral responsibility. They maintain that control is relevant, when it is, because actions that are controlled by agents are typically attributable to the agent in a way sufficient to ground their responsibility for the action. Absence of control is correlated with absence of responsibility, but sometimes agents fail to exercise responsibility-level control but are responsible nevertheless. Even if they accept the claim, for which I argued in the last section, that implicit attitudes block responsibility-level control for actions that have a moral character due to these attitudes, it is open to proponents of attributability views to insist that they are responsible nevertheless, because implicit attitudes are at least sometimes deeply enough attributable to the agent to ground responsibility.

In this section, I will argue that attributability theorists should join with proponents of control-based accounts in accepting that agents are not morally responsible for actions caused by implicit attitudes in those cases in which the action has a character that is due to the implicit attitude(s). Schematically, I will argue as follows (with premise one assumed for the sake of the argument):

(1) Agents are morally responsible for actions or the consequences of their actions when they are caused (perhaps non-deviantly) by attitudes that are sufficiently deeply attributable to them.

(2) In those cases in which an action (or its consequences) have a moral character due to the agent's implicit attitudes, and would lack that character were the action controlled by their explicit attitudes, the attitude is not deeply attributable to the agent.

If attributability is a necessary condition of direct moral responsibility, agents are therefore not responsible for these actions or their consequences.

In what follows, I will argue that implicit attitudes apt to cause actions with a moral character they would have lacked were the actions caused by their explicit attitudes are outliers in agents' cognitive economy.

We can (I claim) assess the extent to which a state belongs to an agent by asking about how it is *acquired*, when (or whether) it is *eliminated* and by measuring whether it has come to be *annexed* to the self. Acquisition, elimination and annexation of attitudes are evidence with regard to attributability because (as we shall see) they are evidence for the consistency of the attitude with *other* attitudes that belong to the agent and which partially constitute her agency. To be an agent is to possess a standpoint from which one deliberates and that standpoint is constituted by a relatively coherent set of attitudes (Ismael 2015). A consideration counts *as* a reason for someone just insofar as it meshes with this web of attitudes; it is this fact that underlies standard belief-desire explanations of rational action. It is for this reason that tests of consistency with those attitudes that constitute what Ismael calls the *deliberative standpoint* is a test for the degree to which the attitude belongs to the agent. Attitudes are acquired through engagement with the deliberative standpoint only when they have appropriate relations with those that make up the standpoint; conversely, attitudes acquired in ways that bypass the deliberative standpoint will not become enmeshed in it. Of course, we can apply this test only if we can confidently identify some attitudes as properly belonging to the deliberative standpoint. We can often identify such attitudes by way of their role in the behavior of agents: insofar as they are implicated in consistent and instrumentally rational decision-making and behavior, these attitudes may be said to partially constitute the agent's deliberative standpoint.

It should be acknowledged that the claim that an attitude belongs to an agent when it partially constitutes the agent's deliberative standpoint would be rejected by some proponents of attributability accounts of moral responsibility. Shoemaker (2011) for instance denies that any kind of semantic relation between an attitude and other states of the agent is necessary for attributability. He cites agents' cares, such as the love a parent may have for a child, which may persist even after the parent comes to judge that the child is not worth her love (perhaps he has turned into a violent criminal). Like Smith (2012), I don't find these cases especially convincing: I am skeptical that the evaluative and semantic links between the parent's love and her other attitudes are severed so long as the love persists.¹¹ In any case, I don't want to argue the point. Perhaps there is a kind of attributability that is independent of such links. Nevertheless, it does not ground responsibility: not, at least, responsibility in the sense at issue here. Shoemaker, who maintains that attributability is sufficient for a kind of responsibility, explicitly severs the links between this kind of responsibility and the reactive attitudes (617). Like most theorists of moral responsibility, I am concerned with a notion of responsibility that is constitutively linked to the appropriateness of the reactive attitudes (see McKenna and Russell 2008 for discussion). I am skeptical that any conception of moral responsibility that divorces it from the reactive attitudes concerns anything that is genuinely similar enough to the kind of moral responsibility at issue here to perspicuously be referred to by the same label, but I do not wish to argue for that claim. I am content to stipulate the sense of moral responsibility at issue.

Evidence about how agents acquire implicit attitudes is evidence for the degree to which such attitudes belongs to their deliberative standpoint because under ideal conditions, agents acquire and maintain attitudes only when they are consistent with the attitudes constitutive of that standpoint; inconsistency should lead either to revision of their former attitudes or rejection of the new attitude. Evidence that agents acquire attitudes that are inconsistent with their (continuing) attitudes is therefore evidence that the attitudes acquired do not belong to the agent. Similarly, recognition of an attitude's inconsistency with other attitudes should lead to the elimination of one or other: if both persist, we have reason to think that one or the other should not be fully attributed to the agent.

Both acquisition and elimination are evidence about the extent to which an attitude belongs to an agent when the attitude conflicts with other attitudes which we have good reason to believe partially constitute the agent's deliberative standpoint. To the extent to which attitude *a* is acquired or fails to be eliminated despite inconsistency with these attitudes, *a* does not itself belong to the agent. The degree to which an attitude belongs to the agent will vary from case to case, as a function of two different factors. First, the degree to which an attitude belongs to an agent varies in proportion to its degree of inconsistency with the attitudes partially constitutive of the agent's deliberative standpoint (some attitudes are merely in tension with others, while others are logically inconsistent: for instance, a belief in small government is in tension with the belief that there ought to be a strong welfare system, while the belief that only religious people can be moral agents is logically inconsistent with the belief that a particular non-believer is a moral agent). Second, the attitudes with which a newly acquired or a maintained attitude conflicts can be more or less central to the agent's web of attitudes, where centrality is a function of mutual entailments and other semantic relations between attitudes (some attitudes are on the margins of an agent's web of attitudes, because they do not enter into semantic relations with many other attitudes). We may often be able to use conscious endorsement as a good heuristic for centrality to a web of attitudes, because conscious endorsement, as a matter of psychological fact, tends to bring it about that the attitude is tested for consistency with other representations and leads to the formation of semantic relations with other attitudes (Levy 2014a).

An important caveat should be noted: under some circumstances an attitude may conflict with others that are genuinely constitutive of the agent's deliberative standpoint and yet properly be attributed to the agent. Ordinary cases of *akrasia*, for instance, may involve attitudes that conflict *both* of which are nevertheless deeply attributable to the agent. Angela Smith (2004; 2012) has suggested that the same might be true in cases involving implicit attitudes. The unconscious racism of a person may reflect her unconscious assessment of the worth of minorities, as much as her conscious nonracist attitudes may reflect her conscious evaluations. The fact that unconscious attitudes are not endorsed by the person does not prevent it from being true that they "are rational responses that depend for their existence on the person's evaluative judgment" (2004: 345). These cases are not counterexamples to the claim that an attitude must be partially constitutive of an agent's deliberative standpoint to be attributable to her in the kind of way that grounds

moral responsibility, however. *Both* of the conflicting attitudes in these cases is embedded in deep semantic relations to many other attitudes which themselves are plainly attributable to the agent. In fact, both the conflicting attitudes in standard cases of akrasia are probably inferentially linked to many of the *same* attitudes.¹²

This fact indicates that the conflict criterion must be employed carefully. It is not sufficient to identify a conflict between attitudes constitutive of the agent's deliberative standpoint and another attitude. We must test, to the extent we can, whether the outlier is nevertheless embedded in semantic relations to other attitudes which are equally constitutive of the agent's deliberative standpoint. Evidence of conflict is *prima facie* evidence of lack of attributability and typically strong evidence, ensuring that we can employ the conflict criterion as a good heuristic for attributability; when in addition we have good reason to believe that the attitude is not supported by a network of other attitudes, evidence of conflict is extremely strong evidence of lack of attributability. I now turn to the data on the extent to which implicit attitudes may be acquired, eliminated or annexed to assess the extent to which they are embedded in appropriate relations with the attitudes constitutive of an agent's deliberative standpoint.

Let's consider acquisition first. Implicit attitudes *may* be acquired inferentially. De Houwer (2014) reports one study, for instance, in which positive pictures were presented to subjects alongside a grey square with the number 1 written on it while negative pictures were presented alongside a grey square with the number 2 written on it. Afterwards, subjects were shown two neutral pictures and told that during the previous presentations, grey squares with the number 1 always occluded one picture while squares with the number 2 on them always occluded the other. An IAT then showed that participants preferred the first neutral picture to the second. These results are difficult to explain via associationistic theories of acquisition, since the neutral pictures were never presented to the subjects along with stimuli that might be expected to give rise to positive or negative implicit attitudes. Explaining the data seems to require postulating some kind of inferential process, roughly along the following lines: because the neutral picture was occluded by a square that was presented alongside positive (or negative) pictures, it is appropriate to take the same attitude toward it as toward the positive (or negative) pictures.

Inferential acquisition of an attitude is direct evidence of consistency with the attitudes constitutive of the agent's deliberative standpoint, since inference engages this standpoint. Attitudes ought to be adopted when there are appropriate inferential relations between the attitudes constitutive of their deliberative standpoint and the new attitude, and not otherwise. There is, however, plentiful evidence that implicit attitudes can be acquired in ways that bypass and even conflict with the attitudes that constitute an agent's standpoint. Here are two examples. Ranganath & Nosek (2008) gave subjects information, positive and negative, about particular individuals from each of two groups. They were then introduced to new members of each group. Implicit attitudes to the new individuals generalized to the new individuals, but subjects did not explicitly evaluate the new individuals as relevantly similar to their fellow group members. They acquired implicit attitudes toward new individuals which diverged from the attitudes that they (rightly) took themselves to have reason to have. Moran & Bar-Anan (2013) exposed subjects to pleasant music and a highly unpleasant noise (a human scream). The presentation of a picture predicted the ending of the sounds. Subjects expressed a preference for the image that predicted the cessation of the scream over the image that ended the music, but their automatic evaluations showed the reverse pattern: co-occurrence of the images with the pleasant and the unpleasant sounds produced an automatic preference for the image paired with the pleasant noise. Again, the subjects acquired attitudes that conflicted with those they acquired inferentially. The conflict between the attitudes acquired and the attitudes constitutive of the agents' deliberative standpoints is evidence that the new attitudes did not properly belong to the agent..

We have already seen evidence that implicit attitudes can resist elimination when the person takes herself to have sufficient reason to reject them: Gregg, Seibt and Banaji (2006) gave subjects decisive reason to reverse their attitudes, but though explicit attitudes reversed, implicit attitudes remained fixed. This insensitivity of the implicit attitude to representations constitutive of the agent's deliberative standpoint is, once again, evidence that the implicit attitude does not properly belong to the agent.

Non-inferential acquisition is acquisition that fails to engage the agent's deliberative standpoint, and fails to establish inferential relations between the attitude acquired the attitudes constitutive of that standpoint. However it is acquired, though, an attitude need

not remain forever isolated from the agent's deliberative standpoint. As the phenomenon of cognitive dissonance indicates, there is pressure for psychological consistency, which would tend to lead to the attitude coming to belong to the set constitutive of the deliberative standpoint. Indeed, follow up studies reported in Ranganath & Nosek (2008) indicate that once acquired an inconsistent attitude may come to be annexed to agents' standpoint: a follow up three days later showed that explicit attitudes toward the new individuals matched implicit attitudes, because the explicit attitude had altered. However, the experimenters suggest that this occurred because the subjects had forgotten the facts that distinguished the new individuals from the old and therefore had to rely on their implicit evaluations when asked to report their judgments. This mechanism seems very unlikely to cause the integration of implicit attitudes that have a content that diverges significantly from attitudes to which the agent is committed, since it can work only when subjects forget their explicit attitudes.

Perhaps, however, there are other routes by which an attitude that was non-inferentially acquired might come to belong to the set that constitute the deliberative standpoint. Devine et al. (2002) report evidence that individuals who hold that being non-prejudiced is intrinsically important (rather than important because of social pressures not to seem prejudiced) display a smaller discrepancy between their implicit and explicit attitudes. This evidence is consistent with the hypothesis that thinking that prejudice is intrinsically wrong causes implicit attitudes to tend to fall into line with explicit judgments, though it is possible the causal arrow runs in the opposite direction, and less prejudiced implicit attitudes play a role in causing agents to believe that being non-prejudiced is intrinsically important.¹³ However, this evidence does not show that when agents perform actions that have a moral character due to their implicit attitudes, these attitudes are responsibility-level attributable to them, for two reasons.

First, while the gap between agents' implicit and explicit attitudes may be narrower in those subjects who hold that being non-prejudiced is intrinsically important, the very fact that these subjects are so strongly committed to being non-prejudiced provides grounds for denying that their implicit attitudes are not fully attributable to them. Given the agents' strongly held commitment to being non-prejudiced, one might say that these divergent implicit attitudes were *more* alien to their real selves than the negative implicit attitudes of agents who are not as committed to equality, despite the fact that the former

are less implicitly biased than the latter. We should think that these attitudes are no better integrated into their deliberative standpoint than are the more prejudiced attitudes of less well-motivated agents, precisely because their deliberative standpoint is non-prejudiced. There is no evidence to think that their implicit attitudes are embedded in rich sets of inferential relations with their standpoint; given the content of their standpoint, if they were they would alter in character.¹⁴

For all we know, however, it may sometimes (or often) happen that the pressures toward consistency bring it about that implicit attitudes come to belong to agents' deliberative standpoint.¹⁵ That brings us to the second reason to deny that when agents perform actions that have a moral character due to their implicit attitudes, these attitudes are responsibility-level attributable to them. If there are cases in which implicit attitudes come to belong to the agent's deliberative standpoint, this very fact ensures that they will no longer cause actions with a moral character due to those attitudes: the counterfactual *had the agent's explicit attitudes controlled behavior, the action would have had a different moral character* will be false. The counterfactual will be false *because* the implicit attitudes are now well integrated into the deliberative standpoint. Annexation to the deliberative standpoint, if and when it occurs, eliminates the class of actions in which we are interested here. So while implicit attitudes may come to be responsibility-level attributable to the agent, in a way appropriate to ground moral responsibility, they will not do so in the cases with which we're concerned.

It is worth noting that if implicit attitudes are patchy endorsements, they will not be entirely alien to the self. They will respond inferentially to some other attitudes which securely belong to the agent, and that suffices for *some* degree of attributability. But when the counterfactual mentioned above is true, they will be crucially at odds with the states with which we can most securely identify the agent; for that reason, we ought to think that in the set of cases in which we're interested, they will be too alien to the self to ground moral responsibility. At very least, proponents of attributability accounts owe us a story to explain why we ought to identify them with the self sufficiently strongly to ground moral responsibility in these cases.

The conditions under which an implicit attitude is annexed to an agent's deliberative standpoint are conditions under which it comes to fall into line with explicit attitudes. It

seems that on the attributability account, agents can't be morally responsible for actions caused by their implicit attitudes when those actions have a moral character that diverges from the character they would have had were their explicit attitudes controlling their behavior – for under those conditions implicit attitudes do not belong to the agent's deliberative standpoint and therefore are not sufficiently deeply attributable to them.

Conclusion

Agents who sincerely profess a commitment to, and a belief (if that's something over and above a commitment) in racial equality sometimes perform actions that have a moral character due to attitudes with conflicting contents. In this paper, I have urged that these agents are not directly morally responsible for these actions; not, at least, if control or attributability are each or both necessary conditions of moral responsibility.

Agents lack personal-level control over the moral character of actions like these, because personal control has demanding epistemic conditions. But personal-level control may not be necessary for moral responsibility. Agents may be morally responsible in virtue of sufficiently patterned sensitivity of their actions to reasons, and this kind of sensitivity does not require guidance by explicit intentions. However, mechanisms that have implicit attitudes as components do not exhibit suitably patterned sensitivity to the kinds of considerations that give rise to a moral character that diverge from the character the action would have had were their explicit attitudes controlling behavior. On a control-based account, this fact seems to excuse agents for moral responsibility with regard to this class of actions.

Attributability accounts converge on the same conclusion. Implicit attitudes do not seem properly to belong to agents' deliberative standpoints, in the way required for responsibility-level attributability. The conditions under which implicit attitudes come to be suitably annexed to the agent's deliberative standpoint are conditions under which it is no longer true that their actions have a moral character due to their implicit attitudes. So if control, or attributability, or both, are necessary conditions of moral responsibility, agents are not directly responsible for actions that have a moral character due to their implicit attitudes.

NOTES

¹ I am grateful to two anonymous referees for this journal for enormously helpful comments that enabled me to clarify the argument and better structure the entire paper. Thanks are also due to audiences at the Australasian Association of Philosophy annual conference and the University of Oxford.

² I choose racist – or apparently racist – implicit attitudes for illustration, and because their effects on behavior are among the most important from a moral point of view. There is plentiful evidence for a conflict between explicit and implicit attitudes concerning gender, sexuality, and other social categories. It is controversial what proportion of people with unprejudiced explicit attitudes harbor conflicting implicit attitudes. Some psychologists claim that the great majority of white Americans have negative implicit attitudes toward black people (Pearson, Dovidio & Gaertner 2009; Dasgupta 2013), but controlling for structural fit of implicit and explicit measures, Payne, Burkley and Stokes (2008) found that a slight majority of their sample exhibited anti-black prejudice. However, different measures of implicit prejudice do not correlate very well with one another (Fazio & Olson 2003; Bar-Anan & Nosek 2014), which leaves open the possibility that the great majority of us exhibit implicit bias on *some* measure.

³ At least this is true given plausible assumptions about what the folk believe about nonconscious states. The idea that we have unconscious states is now well entrenched in folk psychology, but the states that seem to feature in folk thought are not the *sui generis* kinds of representations that – I claim – implicit attitudes are. Instead, they seem to be unconscious beliefs and desires, which differ from explicit attitudes only inasmuch as they are unconscious. Of course, this is an empirical claim, and one that might be usefully tested by experimental philosophers.

⁴ For similar reasons, I set aside the rich philosophical literature on moral responsibility and situationism (e.g., Nelkin 2005; Brink 2013; Vargas 2013b), since this literature assumes that the psychological states that mediate behavior in the experiments at issue are those of folk psychology. Elsewhere, I have suggested that this assumption may in fact often be warranted (Levy 2014a), but it entails that the accounts developed by philosophers responding to this set of data do not transfer easily into the context of implicit attitudes.

⁵ It might be objected that the standard methodology is not all that standard in this debate: while some philosophers have emphasized intuitions, and even folk responses, to implicit attitudes in considering whether we are morally responsible for actions caused by them (see, for instance, Faucher, 2016), others have sought only to identify apparently exculpating features of such attitudes and then tested whether these features are genuinely exculpating by considering more *ordinary* cases that share this feature. Glasgow (2016) asks whether an agent's alienation from her implicit attitudes might exculpate, and concludes that it may not because there are more ordinary cases in which alienation fails to exculpate. Smith (unpublished) asks whether our lack of awareness of the content or existence of such states exculpates, and concludes that it does not because there are more ordinary cases in which such lack of awareness does not exculpate. This argumentative strategy is perfectly appropriate. However, it cannot be used by itself to show that agents are morally responsible for actions caused by implicit attitudes: it can only show that the feature that is focused on (alienation, lack of awareness) does not exculpate, insofar as that feature is genuinely shared with ordinary cases. If implicit attitudes have properties that are importantly different from the folk psychological states that feature in the more ordinary cases on which Glasgow and Smith focus, we ought to hesitate before thinking that the alienation or lack of awareness they discuss function in the same manner in the ordinary case and the cases of interest. Insofar as Glasgow and Smith rely on intuitions

about the similarities between implicit attitudes and folk psychological processes to identify apparently exculpatory features of the former, I think it is fair to regard their approaches as variants of the standard methodology.

⁶ Prominent accounts of moral responsibility which are centred around control include Kane (1996), Fischer and Ravizza (1998), Clarke (2003), Mele (2006) and Vargas (2013a).

⁷ Shepherd would not restrict control in the service of an intention to personal-level control; in addition to control in the service of an explicit intention, there may be control in the service of an unconscious or implicit intention (or motivational state).

⁸ Commitment to an associative account is common to a number of rival models of implicit attitudes: the MODE model associated with Fazio (2007); the associative-propositional model developed by Gawronski and Bodenhausen (2011) and the systems of evaluation model developed by Rydell and McConnell (2006). Prominent philosophers have also endorsed it (Gendler 2008).

⁹ Of course proponents of control-centred views have resources with which to respond to these cases. For instance, they may claim that it is appropriate to blame agents in cases like these because they exercised control over their characters: they brought themselves to be in a state in which they are now insensitive to evidence (Kane 1996 responds to the Luther case in this kind of way). Vargas (2005) and Sher (2009) each argue that this attempt to locate control in the causal history of agents will not be able to establish that agents are responsible in all the cases in which we are strongly disposed to blame or praise them.

¹⁰ It must be emphasized that there are enormous differences between the various views I am shoehorning together under this label. It is entirely fair to accuse me of neglecting their subtleties. My aim in this paper is to highlight how the properties of implicit attitudes, to which moral philosophers have previously paid relatively scant regard, affect the capacities of agents to exercise control over the moral character of certain actions, and the extent to which such attitudes may appropriately be attributed to them. Neglect of the subtle but important differences between the various accounts is justified, in my view, by the need to attend to the psychological detail. This leaves open the possibility that some view or other has resources I have overlooked whereby to reply to my arguments.

¹¹ Isn't the fierce love that many parents report for a newborn infant evidence that cares may be severed from any semantic links to other states? I doubt it: at least in typical cases the parent has already built up a rich set of hopes and imaginings around the child. It comes into the world with a place prepared for it, not only physically but also psychologically.

¹² Elsewhere, I have argued that akratic conflict involves attitudes that are each so deeply linked to the agent's evaluative perspective that either may come up for conscious endorsement (Levy 2014c).

¹³ Galdi, Arcuri and Gawronski (2008) provide evidence that – sometimes, at least – the causal arrow runs from explicit attitudes to implicit. They asked subjects for their views about a then current controversial political issue. They found that for those subjects who had made up their mind on the issue, their explicit attitudes predicted their implicit attitudes one week later significantly more strongly than they did at the time of initial testing. They conclude that the consciously reported belief tends to influence unconscious attitudes.

¹⁴ Smith (unpublished) gives several reasons for thinking that we ought to think agents' implicit attitudes are embedded in rich inferential relations to other attitudes that are constitutive of her practical identity. Some of them are empirical, concerning how these attitudes are acquired. I have dealt with these matters in the text. In addition, though, she

cites our responses to the discovery that we harbor such attitudes, as well as the responses of others when we treat them badly as a consequence of such attitudes. We may be horrified at the discovery that we harbor implicit attitudes, and others may take their manifestation to be offensive; both responses indicate that we respond to them as revelatory of our practical identity. However, these responses are the predictable consequence of the facts that our intuitive responses are keyed to the states and processes of folk psychology: when we find ourselves to harbor implicit biases, we take ourselves to unconsciously believe that (say) women are inferior to men. Since these responses are generated in this way, they should not be taken to have evidential value in this context.

¹⁵ It is an open question whether implicit attitudes *survive* such annexation: an annexed attitude possesses the appropriate set of inferential relations to other attitudes, and thereby cease to be a patchy endorsement. If implicit attitudes are always patchy endorsements, such an annexed attitude transforms into an ordinary attitude (conscious or not). It will have a content that is expressible in propositional terms. Given the right cues, it can become conscious, and the agent can affirm it (and will do so if sincere). When behavior is caused by an attitude annexed to the real self, its being implicit may be a fact about its history, not its current status.

References

AU 2014b

Arpaly, N. (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods* 46: 668-88.

Beilock, S. L., & Carr, T. H. (2004). From novice to expert performance: Attention, memory, and the control of complex sensorimotor skills. In A. M. Williams, N. J. Hodges, M. A. Scott, & M. L. J. Court (eds.), *Skill acquisition in sport: Research, theory and practice* (pp. 309-328). London: Routledge.

Brink, D.O. (2013). Situationism, Responsibility, and Fair Opportunity. *Social Philosophy and Policy* 30: 121-149.

Clarke, R. (2003). *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology* 47: 233-279.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass* 8: 342-353.

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, Mass.: MIT.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology* 91: 385–405.

Devine, P., Plant, E., Amodio, D., Harmon-Jones, E. and Vance, S. (2002). The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice. *Journal of Personality and Social Psychology* 82: 835-848.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science* 11: 319-323.

Egan, A. (2011). Comments on Gendler's 'The epistemic costs of implicit bias'. *Philosophical Studies* 156: 65–79.

Faucher, L. (2016). Revisionism and Moral Responsibility for Implicit Attitudes. In Brownstein, M. and Saul, J. (eds). *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*: Oxford, Oxford University Press (115-144).

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition* 25: 603-637.

Fazio, R. and Olson, M. (2003). Implicit measure in social cognition research: Their meaning and use. *Annual Review of Psychology* 54: 297-327.

Fischer, John Martin, and Mark Ravizza. (1998). *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science* 321: 1100–1102.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology* 44: 59-127.

Gendler, T. (2008). Alief and belief. *Journal of Philosophy* 105: 634–63.

Glasgow, J. (2016). Alienation and Responsibility. In Brownstein, M. and Saul, J. (eds). *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*. Oxford, Oxford University Press (37-61).

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464-1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97: 17–41.

Greenwald, A.G., Banaji, M.R., & Nozek, B.A. (2015). Statistically Small Effects of the Implicit Association Test can Have Societally Large Effects. *Journal of Personality and Social Psychology* 108: 553-561.

Gregg AP, Seibt B, Banaji MR. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology* 90: 1-20.

Haji, I. (2012) *Reason's Debt to Freedom*. Oxford: Oxford University Press.

- Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally-created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology* 42: 259-272.
- Hasson, U., & S. Glucksberg. (2006). Does negation entail affirmation? The case of negated metaphors. *Journal of Pragmatics* 38: 1015–32.
- Holroyd, J. (2012). Responsibility for Implicit Bias. *Journal of Social Philosophy* 43: 274-306.
- Ismael, J.T. (2015). On Being Someone. In Mele, A.R. (ed). *Surrounding Free Will: Philosophy, Psychology, Neuroscience*. Oxford: Oxford University Press (274-297)
- Kane, R. (1996. *The Significance of Free Will*, New York: Oxford University Press.
- Lee, C.J. (2016). Revisiting Current Causes of Women’s Underrepresentation in Science. In Brownstein, M. and Saul, J. (eds). *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*. Oxford, Oxford University Press (265-282).
- Levy, N. (2014a). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.
- Levy, N. (2014b). Consciousness, Implicit Attitudes and Moral Responsibility. *Nous* 48: 21-40.
- Levy, N. (2014c). Addiction as a Disorder of Belief. *Biology & Philosophy*, 29 (2014), 315-225.
- Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous* 49: 800-823.
- Mandelbaum, E. (2013). Against alief. *Philosophical Studies* 165:197-211.
- Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Nous* 50: 629-658.

- McConnell, Allen R. and Jill M. Leibold (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology* 37: 435-442.
- McKenna, M. & Russell, P. (2008). Introduction: perspectives on P.F. Strawson's "Freedom and Resentment". In McKenna, M. & Russell, P. (eds). *Free Will and Reactive Attitudes* (1-17). Farnham: Ashgate Publishing.
- McKenna, M. (2013). Directed Blame and Conversation. In Coates, D.J. and Tognazzini N. (eds.) *Blame: Its Nature and Norms* (pp. 119-140). Oxford: Oxford University Press.
- Mele, Alfred. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Moran, T. & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & emotion* 27: 743-752.
- Nelkin, D.K. (2005). Freedom, Responsibility and the Challenge of Situationism. *Midwest Studies in Philosophy* 29:181–206.
- Newman, G., Diesendruck, G., & Bloom, P. (2011). Celebrity contagion and the value of objects. *The Journal of Consumer Research* 38: 215–228.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review (pp. 265–292). In J. A. Bargh (ed.), *Automatic processes in social thinking and behavior*. New York: Psychology Press.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105: 171–192.
- Payne, B.K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology* 89: 277-293.

Payne, B. K. (2006). Weapon bias: Split second decisions and unintended stereotyping. *Current Directions in Psychological Science* 15: 287-29.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology* 94: 16-31.

Pearson, A.R., Dovidio, J.F., & Gaertner, A.L., (2009). The Nature of Contemporary Prejudice: Insights from Aversive Racism. *Social and Personality Psychology Compass* 3: 1-25.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Pereboom, D. (2013). Free Will Skepticism, Blame and Obligation. In Coates, D.J. and Tognazzini N. (eds.) *Blame: Its Nature and Norms* (pp. 189-206). Oxford: Oxford University Press.

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately, explicit attitude generalization takes time. *Psychological Science* 19: 249-254

Ranganath, K., Smith, C., & Nosek, B. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology* 44: 386–396.

Rozin, P., Markwith, M. & Ross, B. (1990). The sympathetic magical law of similarity, nominal realism, and neglect of negatives in response to negative labels. *Psychological Science* 1: 383-384.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology* 91: 995-1008.

Shepherd, J (2014). The Contours of Control. *Philosophical Studies*: 170: 395-411.

Sher, G. (2009). *Who Knew? Responsibility Without Awareness*. New York: Oxford

University Press.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics* 121: 602-632.

Smith, A. (2004). Conflicting Attitudes, Moral Agency, and Conceptions of the Self. *Philosophical Topics* 32: 331-352.

Smith, Angela M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies* 138: 367–392

Smith, A. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics* 122 :575-589.

Smith, A. (unpublished). Implicit Biases, Moral Agency, and Moral Responsibility.

Son Hing, L. S., Chung-Yan, G. A., Hamilton, L. K. & Zanna, M. P. (2008) A two-dimensional model that employs explicit and implicit attitudes to characterize prejudice. *Journal of Personality and Social Psychology* 94: 971–987.

Stich, S. (1978). Beliefs and subdoxastic states. *Philosophy of Science* 45: 499-518.

Uhlmann, E.L. & Cohen, G.L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* 16: 474-480.

Vargas, M. (2005). The Trouble with Tracing. *Midwest Studies in Philosophy* 29: 269–291.

Vargas, M. (2013a). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. (2013b). Situationism and Moral Responsibility: Free Will in Fragments. In Clark, A., Kiverstein, J. and Vierkant, T. (eds.) *Decomposing the Will* (pp. 325-349). Oxford: Oxford University Press.

Washington, N. and Kelly, D. (2016). Who's responsible for this? Implicit bias and the knowledge condition. In Brownstein, M. and Saul, J. (eds). *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*. Oxford, Oxford University Press (11-36).

Wegner, D. (1984). Innuendo and damage to reputation. *Advances in Consumer Research* 11: 694-96.

Zheng, R. (2016). Attributability, Accountability and Implicit Attitudes. In Brownstein, M. and Saul, J. (eds). *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*. Oxford, Oxford University Press (62-89).