Consciousness, Implicit Attitudes and Moral Responsibility

NEIL LEVY
Florey Neuroscience Institutes

What role, if any, does consciousness of our reasons for acting play in settling whether we may justifiably be held responsible for our actions? Most philosophers, and ordinary people, seem to assume that consciousness of this kind is essential for moral responsibility: if an agent fails to be conscious of their reasons for acting (and they are not responsible for that fact), they ought to be excused responsibility. Recently, however, this assumption has been rejected by researchers in a variety of disciplines, from philosophers (for example Arpaly 2002; Sher 2009; Suhler & Churchland 2009) to cognitive scientists (Wegner 2002). In this paper, I want to make a start on defending the assumption. I will not address the arguments of these researchers directly. Rather, I will set out a case for thinking that consciousness of our reasons for acting is morally significant, in a manner that at least typically entails a difference in whether agents are morally responsible for their actions.

I will frame the argument by way of responding to a recent challenge from two philosophers, Matt King and Peter Carruthers. Marshalling evidence from cognitive science, King and Carruthers (forthcoming) have recently argued that there are good reasons to think that consciousness of the propositional attitudes (beliefs, desires, and so forth) which cause our actions cannot distinguish morally responsible actions from those for which we ought to be excused, because there are good reasons to think that there are no conscious attitudes. In the light of this claim, they challenge us to justify or abandon the claim that consciousness is needed for moral responsibility. I will argue that even if there are no conscious attitudes, nevertheless coming to be aware of the content of the propositional attitudes that cause our actions makes a difference to the moral status of those actions, a difference significant enough to distinguish between actions for which we are rightly held responsible and those for which we ought to be excused. I will go on to show, however, that consciousness of the content of our attitudes plays this role only with regard to some of our attitudes: becoming conscious of the content of our implicit attitudes fails to play the role in our agency that makes consciousness of (some of) our explicit attitudes so morally significant. Getting clear on the distinction between, and the functional role of, implicit and explicit attitudes, will enable us to come to have a clearer idea of the conditions under which consciousness of our attitudes plays a role in moral agency.
Before I turn to setting out King and Carruthers’ challenge, let me say a word about how I shall use the word ‘consciousness’ in what follows. When King and Carruthers claim that there are good reasons to think that there are no conscious attitudes, I take it they do not mean that our attitudes might lack phenomenal qualities. Rather, what they have in mind is that agents may fail to have a certain kind of awareness of their attitudes. When subjects are aware of information in this kind of way, its contents are available to a variety of consuming systems. Paradigmatically, the information is available to report (to self and others), at least in normal, awake subjects, but also available directly (that is, not in a manner mediated by report) to other systems. It might be the case that awareness of this kind entails, or requires, some kind of phenomenal feel, but it does not follow that when attitudes are available in this way, agents experience some kind of associated phenomenal quality. Availability is a dispositional notion, whereas phenomenality is an occurrent property; an attitude might be available for report and for consumption by other systems without (yet) being tokened and therefore without the agent experiencing any associated phenomenal quality (if it is in fact the case that there is a phenomenal feel associated with propositional attitudes).

King and Carruthers on Moral Responsibility

Carruthers (2009) argues that our access to the content of our own attitudes is via precisely the same mindreading machinery that we use to ascribe attitudes to others. The evidence for this claim consists, in large part, of the relative ease with which subjects can be brought to misattribute attitudes to themselves. Being paid a paltry sum to tell another person that the boring and repetitive task in which we have just engaged is fun leads us to think we enjoyed it (Festinger & Carlsmith 1959); being asked to nod our heads—ostensibly to test the headphones we are wearing—makes us more likely to think we agree with an argument (Wells & Petty 1980); observing that we have chosen one item from a range that seems identical leads us to conclude that we have detected differences between them (Nisbett & Wilson 1977), and so on. This evidence—together with a great deal more cited by Carruthers—is taken by him to show that we attribute attitudes to ourselves by interpretation of our behavior. The evidence supports this view in two ways: First, the fact that we can relatively easily be misled about our attitudes at all seems to indicate that we lack direct access to their contents. Second, the fact that it is interpretive cues to attitudes that are manipulated in these experiments seems to show that it is these cues that we use to self-attribute attitudes.

In their paper, King and Carruthers aim to explore the implications of Carruthers’s view for moral responsibility. First, they argue that if our access to the content of our attitudes is indirect, in the manner suggested by Carruthers, then there are no conscious attitudes. Attitudes, as such, are not conscious. Carruthers is sympathetic to a global workspace theory of consciousness, but he denies that our attitudes are themselves globally broadcast. Rather, imagistic representations, utilizing the mechanisms of perception, are globally broadcast. Typically, these representations are in speech, inner or overt, and cannot be the attitudes themselves
because they are sometimes misrepresentations. We regularly get the contents of our attitudes right, of course, but in doing so we become aware of the contents of our attitudes, not of the attitudes themselves. We have propositional attitudes, and we often come to know their contents, but our attitudes themselves are never conscious.

If Carruthers is right, our actions are never caused by conscious attitudes. But most extant accounts of moral responsibility, King and Carruthers argue, assume that our actions are often caused by conscious attitudes, and that this fact makes a significant difference to their moral status. They illustrate their claim with the case of the alien hand, a neurological syndrome that has impressed some philosophers as a clear case in which behavior is nonresponsible due—in part—to its etiology in nonconscious states (Levy & Bayne 2004). If all our actions are caused by unconscious attitudes, the actions of the alien hand are not strikingly different from other actions, and the case for regarding them as morally different weakens considerably. Of course, there are differences between the cases. In the normal case, agents—typically—experience little resistance to self-attributing the goal of the behaviour to themselves, whereas sufferers from the pathology typically deny that the mental states in question are theirs (Della Sala et al., 1991; Goldberg & Bloom, 1990). But as we just saw, people are unreliable self-interpreters. Since we lack direct access to our attitudes, an agent’s denial that a representation which guided the movement of their hand is one that ought to be attributed to them ought not to be given much weight.

King and Carruthers concede that there may be other grounds for holding that the actions of the alien hand are not ones for which the agent in question ought to be held responsible. They caution against looking too assiduously for such grounds, however. They suggest that the intuition that sufferers from alien hand syndrome are not responsible for their actions stems from the pretheoretical idea that consciousness marks a morally significant dividing line (possibly due to an innate belief that the mind is transparent to itself (Carruthers 2008)), and that this intuition might be driving the search for an excusing factor. Just as we may confabulate the content of our attitudes, so we may misattribute our motivations for philosophical theorizing. This being the case, they suggest, we ought to accept “that explaining such intuitions should no longer be taken as one of the main goals of theorizing about responsibility”.

There is a challenge here to our practices of blaming and excusing that we ought to take very seriously. However the example King and Carruthers utilize to illustrate the challenge, the alien hand, is not well chosen. There are multiple overlapping reasons to think that sufferers from the alien hand syndrome are not (directly) responsible for the hand’s movements or their effects. First and most obviously, they have lost the ability directly to inhibit the behavior; were it not for this fact, they would not act as they do. In fact, in many cases these patients resort to indirect means of preventing unwanted behavior, for example holding it with the other hand, or even tying it down (Feinberg, Schinder et al. 1992). Second, the representations that guide the hand are not of a kind that could ground moral responsibility, on any plausible account. The movements of the alien hand are not mere reflexes; rather they are clearly responses to environmental contingencies. To
that extent, it is appropriate to describe them as guided by representations. But
the representations are likely to be very low level. They may, indeed, be nothing
more than representations of affordances. Alien hands do not engage in intelligent
behaviour, of a kind that suggests that the representations that guide their behaviour
are anything more than motor representations: representations of possibilities for
action (Feinberg, Schinder et al. (1992) describe one case as characterized by “a
prominent grasp response”). Alien hands take food from plates, raise cups of tea
to the mouth, scratch itches, tear paper or cloth, and sometimes interfere with the
movements of the other hand. These movements neither require nor suggest the
involvement of representations of any complexity. The representations that guide
this behaviour are ‘thin’: they are responsive only to a very narrow set of features.
Correlatively, they cause behaviors in ways that are inflexible, since the behavior is
not modulated by features outside this narrow band.

Because the representations that guide the alien hand are thin, there is a strong
case for refusing to identify the movements they cause with the agent. An agent’s
propositional attitudes should not be identified with thin representations; rather,
they are built up out of sets of representations. My hope that Obama wins the
next election, say, is not identical to any thin representation with an Obama-related
content; rather, it is constituted out of (perhaps inter alia) many thin representa-
tions; moreover, some of those representations may have contents that, taken alone,
would suggest that I hope he lose the next election. This is most easily seen with
regard to the attitude of believing that p. Suppose I am asked whether I believe that
p—say, whether modal realism is true, which is a question rather orthogonal to my
philosophical interests, and which I have not formerly spent any time considering.
Suppose, further, that I have never asked myself whether I believe that modal re-
alism is true. It might nevertheless be the case that there is a true answer to the
question whether I believe that modal realism is true: I might have a dispositional
belief that modal realism is true (say). My dispositional belief will be built up out
of other dispositions of mine, but identical to none of these dispositions. It cannot
be identical to any of them since some of these dispositions will have contents that
suggest I believe that modal realism is false (perhaps it seems to me metaphysically
extravagant on the face of it). In this case, my personal-level attitude is constituted
out of dispositions thinner than this attitude. And so it is with my moral attitudes
too, I claim.

A further reason to think that my moral attitudes are not identical to the thin
representations that may nevertheless sometimes cause my behavior consists in the
fact that typically at least, such representations will cause behavior only when I
lose the ability to inhibit these behaviors. This inhibition typically occurs prior
to the content of the representation coming to the agent’s awareness: it is only
in pathological cases (alien hand, imitation or utilization behavior, and the like)
that the agent comes to be aware of these representations at all, and then only
by observing their effects on his behavior. If an agent would, when functioning
properly, inhibit an action with a particular content, instead performing actions
with a conflicting content, we have good (if defeasible) grounds for attributing the
latter kind of content, and the actions it causes, to him, and not the former. But
it is only under conditions in which we are not functioning properly (when we are suffering from a lesion, most dramatically, more controversially, perhaps when we are very tired or under great stress) that our behaviors are caused by thin representations alone. Behaviors caused by thin representations are not expressions of our evaluative stance, and therefore fail to express our good or ill will. Indeed, it may be that it is because the behavior of the alien hand is not properly reflective of our agency that the person loses control over it: at least, its failure to be responsive to a broad range of representations is sufficient for a loss of control. If the behavior of the hand is not responsive (directly—as we saw, patients can and do exercise indirect control over the hand, by grasping it with the other hand, for instance) to the beliefs, values (and so on) of the agent, then the person has lost control over it. In the normal case, the fact that one abhors violence is sufficient to inhibit any impulse toward strangulation; in the pathological case, this inhibition fails.

**Responding to King and Carruthers: First Pass**

What distinguishes the attitudes that guide the alien hand from those that issue in morally responsible behavior? Though—if Carruthers is correct—both kinds of attitudes are nonconscious, nevertheless the attitudes that guide responsible behavior seem typically to be attitudes that agents self-attribute. That is, though we may never have conscious attitudes, we are conscious of the contents of the attitudes that cause responsible behavior.

The claim that we are often conscious of the content of our attitudes is a claim with which King and Carruthers seem to have no quarrel. The evolutionary hypothesis to which Carruthers is committed, according to which mindreading is adaptive, commits him to the claim that it is highly reliable: though it is adaptive to come to have true beliefs about the attitudes of conspecifics, it is likely to be costly to come to have false beliefs about their attitudes. Mindreading must therefore generate few errors with regard to other agents. It is likely to be even more reliable in our own case than in the case of others, since we have so much more behavioral evidence in our own case. Also frequently available to our mindreading machinery are the verbal expressions of our attitudes, whether in overt or in inner speech (formulation of these expressions does not require prior access to the content of our attitudes: executive systems can formulate linguistic expressions of contents using first-order representations alone). This is not direct access to the attitudes themselves: the expressions must still be interpreted by the mindreading system in order to be understood (it is because attitudes can be available for report without being—directly—available to a variety of other consuming systems that availability for report alone is not sufficient for awareness). Nonetheless this is usually a relatively trivial task, especially given that related semantic content is rendered highly accessible by the activation of the content. As a consequence, people are probably “excellent interpreters of themselves” (Carruthers 2009: 127).

Vindicating the claim that consciousness of attitudes marks a morally significant boundary can start from this fact, by identifying a functional role for
consciousness of the contents of our attitudes, or at any rate effects that such consciousness typically has. Given the global workspace account of consciousness (to which Carruthers subscribes), or any of the other available accounts belonging to what Morsella (2005) calls the “integration consensus”, it is relatively easy to identify a role that these states have and that other states lack. Whereas the representations that guide the alien hand are encapsulated from other representations, representations that enter consciousness are available (or broadcast) to a variety of the systems that drive behavior. The contents of our attitudes reliably become conscious and therefore are integrated with other personal level contents. As a consequence, they guide a broad and integrated range of behavior: availability to, or consumption by, a broad range of systems ensures that behavior reflects a correspondingly broad range of the agent’s attitudes.

This response to King and Carruthers generates a prediction: if a person self-attributes an attitude, that act of self-attribution should have an integrating effect on their behavior. That is, attribution of the state—veridically or not—should have a coordinating effect on behavior over some significant stretch of time. Unlike the representations guiding the alien hand, which do not coordinate behavior (on the contrary, these representations cause behavior that often conflicts with the personal-level goals of the agent), self-attribution of an attitude should cause the agent to pursue behaviors that are consistent with that attitude as well as with other personal-level concerns. Obviously, we should not demand too much in the way of integration: human beings are imperfectly rational. Moreover, I am here suggesting an empirical generalization, not a conceptual claim (of the kind that sometimes motivates claims like the claim that sufferers from delusions fail to believe the content of their delusion); I am suggesting that self-attributed states tend rather reliably to have this role. And the claim seems to be true. Consider the cognitive dissonance experiments that Carruthers cites as evidence for his view that our access to the contents of our attitudes is via self-interpretation and not introspection. Having attributed to themselves an attitude in a way that we have good reason to regard as confabulatory (because the probability that randomly assigned agents should differ significantly by condition in ways that don’t reflect the manipulation is low), agents go on to act—for a time—in ways that are consistent with that self-attribution. Indeed, they may retain the attitude for a very extended time, and use it as a premise in reasoning and therefore in generating further attitudes. This is exactly what happened in the most famous study of cognitive dissonance, Festinger’s When Prophecy Fails (Festinger et al. 1956). Rather than losing faith in the teachings of their leader, members of an apocalyptic group ramped up their faith in response to the failure of her prediction that the world would end on a specific date.

Carruthers himself recognizes that this prediction is borne out by evidence: the “self-attribution of mental states” is, as he notes, “likely to be self-fulfilling” (Carruthers 2009: 127). Since, however, we are pretty reliable self-interpreters, we will typically attribute to ourselves attitudes that we actually had prior to the act of self-attribution. Hence it will be attribution to ourselves of our actual attitudes that
will typically have this kind of unifying effect on our behavior. Becoming conscious of (the content of) our attitudes has an effect on our behavior. So consciousness of our attitudes has systematic effects. Why think, though, that these effects are morally significant—why think that consciousness marks a morally significant boundary?

I will have more to say about this question later in the paper. For the moment, let me simply note that autonomous agency, the kind of agency plausibly required for moral responsibility, just is a diachronic property: we are morally responsible agents insofar as we are able to pursue projects across time (Levy 2006). Arguably, being a moral agent at all requires certain capacities that are available only to beings with the capacity to formulate and guide behavior in the light of plans and projects, where this capacity requires that the agent have future-oriented desires and believes, truly, that they are able effectively to pursue these plans (Tooley 1987; Singer 1993). So insofar as consciousness underwrites the unification of behavior it underwrites moral agency, and insofar as it underwrites a kind of agency that might be uniquely human—the unification of behavior over supramodules, perhaps (Morsella 2005)—it plausibly is taken to underwrite personhood. Further, there are strong reasons for holding that these capacities, and their actual exercise, are necessary for moral responsibility; I will venture some further remarks on the relationship between moral responsibility and these capacities toward the end of this paper. For the moment, I will limit myself to pointing out that there are systematic differences between representations that guide behavior with and without consciousness, and that these differences are morally significant. The representations that guide behavior in pathologies or disorders of consciousness—in alien hand syndrome, automatism, utilization behavior and so on—are likely to be very low-level; perhaps nothing more than motor representations. Very likely, they do not reflect agents’ assessments of what they have reason to do, not even their unconscious assessments. If they are motor representations this claim is obviously true: a motor representation is automatically generated by the affordances of objects and constitutes merely a possible way of acting, not a reason for acting. Even if the representations guiding the behavior is attributable to the agent—King and Carruthers give the example of a sufferer from alien hand whose hand lashes out at his wife, presumably in response to a wish he has to injure her—we cannot identify the representation guiding the hand with the agent’s assessments of their reasons concerning what to do, for reasons we have glimpsed before: because propositional attitudes are built out of lower level representations, and not identical to any one of them. Harry’s behavior might be driven by a representation with a semantic content, but that representation does not—by itself—constitute his attitude toward his wife. His overall attitude might be loving and caring; it is because he suffers from a pathology that his behavior fails to reflect this attitude. Even when representations are complex enough to guide quite sophisticated behavior, for example the representations that guide the behavior of agents in fugue states, they represent only a small fraction of what the agent believes. These differences very plausibly make a difference: they distinguish between behaviors for which agents should be excused and those for which they are responsible.
The Inadequacy of the First Pass Response

In responding to King and Carruthers’s challenge, I have argued that consciousness of our attitudes has an integrating effect on agency. However, though it is true that coming to be conscious of the content of our attitudes can have broad and deep integrating effects on our agency, this is actually only the case with regard to some attitudes. It is perfectly possible for an agent to become aware of the content of the representations guiding some of her behavior without her thereby becoming responsible for that behavior. A sufferer from alien hand, for instance, could understand the nature of the representations to which the hand responds and yet the behavior would (or should) still be excused, I shall argue. It is only with regard to some kinds of attitudes, I claim, that coming to be aware of their contents plays the kind of integrating role in agency to underwrite both the capacity for morally responsible agency in general, and token responsible actions. With regard to other attitudes, awareness of their content fails to integrate these contents into personal level behavior; that is, fails to make the actions driven by these attitudes responsive to the agent’s personal-level attitudes as well. Only some explicit attitudes, I shall argue, are apt to play this responsibility-underwriting role.

Normal subjects often have implicit attitudes that diverge markedly from their explicit attitudes. On implicit association tests (IATs), for instance, white subjects who sincerely profess opposition to racism are typically faster and more accurate in associating positive terms with white faces than with black (Dasgupta 2004). A good many of our actions are partly caused by these implicit attitudes, and some of these actions are morally very significant. For instance, scores on an implicit association test seem to be a better predictor of certain kinds of subtle racist behavior than are explicit attitudes (McConnell & Leibold 2001). Moreover, implicit attitudes probably explain some incidents involving lethal force. Priming with black faces raises the likelihood that agents will identify ambiguous stimuli or non-gun tools as guns (Payne 2001); this fact may partially explain why police are more likely to use deadly force when confronted with black suspects. When we are under time pressure, stressed, tired or distracted, our actions often reflect our implicit attitudes, even when our explicit attitudes conflict with them. Unfortunately, many of our most significant actions occur under these kinds of pressures. We do not have the luxury of calm reflection in life-or-death situations, or in those we take (wrongly) to be life or death.

If the first-pass response to King and Carruthers’ challenge were adequate, agents would be morally responsible for the actions their implicit attitudes cause, just in case they are aware of the content of the relevant attitudes. However, as I shall show, awareness of the content of these attitudes fails to render us directly responsible for the actions they cause. This kind of awareness may be morally significant, but its significance is indirect. It may render us indirectly morally responsible for behaviors—in virtue, for instance, of failing to take steps to prevent particular implicit attitudes from causing behavior. But awareness of the content of
our implicit attitudes fails to integrate them into our personal level concerns in the manner required for direct moral responsibility.

To make progress here, we need to get clearer on the nature of the difference between implicit and explicit attitudes. This is a project that is forced on us by Carruthers’ work in any case, since his claims (together with recent work in social psychology) entail that several obvious ways of demarcating explicit and implicit attitudes fail.

It is widely held, first, that our implicit attitudes are unconscious. If Carruthers is correct, of course, this fact about implicit attitudes would fail to distinguish them from explicit attitudes. We might think, second, that the difference comes down to this: we are aware of the content of our explicit attitudes, but not our implicit. But this demarcation criterion is inadequate, empirically and conceptually. It is inadequate conceptually because awareness of the content of our implicit attitudes fails to transform them into explicit attitudes: they retain the same behavioral profiles regardless of our awareness. Subjects who participate in IATs, for instance, and are convinced that they have implicit attitudes that diverge from their explicit, will typically remain disposed to assert claims contrary to those implied by their implicit associations, and the conditions under which their behavior is driven by one set of attitudes or the other alters little or not at all. Moreover, there is a growing body of evidence that subjects often know the content of their implicit attitudes.

The correlation between implicit and explicit attitudes is typically low: one meta-analysis of the IAT data puts it at 0.24 (Hofmann, Gawronski, et al. 2005). However, this low correlation can be explained in ways consistent with subjects knowing the content of both sets of attitudes (Hall and Payne 2010). The low correlation is partially due to subjects’ motivation to present themselves in a favorable light, and therefore to assert non-prejudiced explicit attitudes. Several studies have demonstrated that the degree of correlation between implicit and explicit attitudes varies inversely with motivation to control prejudiced responses (Payne 2001; Fazio et al. 2005). The finding that subjects high in motivation to control prejudice are more likely to express attitudes that diverge from their implicit attitudes is consistent with two different hypotheses: that subjects dissimulate their attitudes, or that they identify themselves with one set of attitudes despite knowing they possess the other. Nier (2005) told subjects that attempts to misrepresent their racial attitudes would be detected by experimenters; subjects in this condition had a significant positive relation between their implicit and explicit attitudes. This evidence seems to support the first hypothesis over the second. However, there is an alternative explanation of these results: subjects may identify themselves with their non-prejudiced commitments, but fear that their implicit attitude will be discovered by experimenters who will think worse of them for denying it. When subjects are given the opportunity to express both their commitments and their implicit attitudes, they give divergent responses, bolstering this hypothesis. Ranganath, Smith & Nosek (2008) asked subjects to rate their “gut reactions” and their “actual feelings” toward gay people. People reported “gut reactions” that were more negative than their “actual feelings”; moreover, their gut reactions correlated well with their implicit attitudes.
This suggests that people know the content of their gut reactions but refuse to identify their real attitudes with these reactions. Implicit attitudes therefore do not appear to be attitudes with contents of which we are unaware. Along with the suggestion that implicit attitudes are unconscious, unawareness therefore fails as a demarcation criterion. A third possibility is that we might distinguish implicit attitudes from explicit by reference to the manner in which we come to know their contents: we introspect our explicit attitudes, but come to know the content of implicit attitudes by some indirect route. Of course, if Carruthers is right this criterion too will fail to distinguish explicit and implicit attitudes. In both cases, our awareness of the contents will be via self-interpretation. Given the failure of all three criteria for distinguishing implicit attitudes from explicit, the case for treating the behaviors they cause alike seems strengthened. However, as we saw above, there are differences in the circumstances in which explicit and implicit attitudes cause behavior. These differences seem to entail that there is a difference in the attitudes themselves. In what follows, I will argue that implicit attitudes have contents that are more loosely bound together than the contents of explicit attitudes; this looseness of binding, in turn, is the product of the ways in which the contents are activated in each case. This difference, I will claim, explains the differences in their behavioral profiles. I shall also argue that these differences entail differences in the effects that awareness of the contents of our attitudes has on our actions and on our agency.

The Structure of Attitudes

Implicit attitudes are constituted by dispositions (or by intrinsic properties that entail dispositions) to respond in various ways. These attitudes may be very thin: for instance, a phobia might be constituted by nothing more than a disposition to feel anxiety in a narrow range of circumstances (proximity to a spider, for instance). Implicit attitudes may sometimes have much broader contents, though. The implicit attitudes revealed by IATs are constituted by the dispositions associated with stereotypes: the African-American stereotype, the gay stereotype, and so on. The contents of these stereotypes are difficult to articulate and may vary from person to person to some extent, but there is a great deal of commonality across subjects. This is unsurprising, given that stereotype content is socially transmitted (Mackie et al. 1996). The content of a stereotype is, roughly, the set of traits associated specifically with representative members of a particular group. Since there are likely to be multiple such traits, implicit attitudes may be quite thick in their contents. Just how thick is difficult to say, because content activation by implicit processes is quite promiscuous. For instance, the male homosexual stereotype is associated with effeminateness (Madon 1997); as a consequence, priming the male homosexual stereotype may activate the feminine stereotype and its associated content (‘nurturing’, say). Stereotypes are multidimensional and can therefore combine negative and positive components (Carlsson & Björklund 2010). Their promiscuity makes it difficult to individuate stereotypes: the gay male stereotype is not the feminine stereotype, but the way in which activation of the first can cause activation
Consciousness, Implicit Attitudes and Moral Responsibility

31

of the second makes it difficult—and perhaps impossible—to decide where the first one ends and the second begins.

The important point is that though implicit attitudes may have quite broad contents, these contents are bound together associatively rather than propositionally (Gawronski & Bodenhausen 2011). It is because stimuli associated with the African-American stereotype automatically activate (or lower the threshold for activation of) the concept of aggressiveness that the latter belongs to the African-American stereotype, not because agents believe that there is a rational connection between the two. Explicit attitudes, at least those with a mind-to-world direction of fit, as well as others with broad contents (though not those with world-to-mind direction of fit and those with narrow content), have their contents bound together in a more rule-based way, I suggest.

Explicit attitudes with broad contents are constituted by, or at any rate entail, the set of dispositions that are normatively associated with an attitude with that content. As Schwitzgebel (2002) notes, beliefs are constituted by a set of dispositions, all or most of which subjects also explicitly accept (or would accept): the disposition to assert p, the disposition to be surprised if not-p, the disposition to rely on p, to take p as a premise in reasoning, to deny q if q is taken to be inconsistent with p, and so on. The dispositions associated with a belief are isomorphic with the dispositions that we ought, normatively, to have: the dispositions associated with the belief given acceptance of correct norms of reasoning. It is only in pathological cases (some cases of delusion, notoriously, and perhaps also extreme self-deception) that agents apparently express a belief while failing to have many of the other dispositions central to the stereotype of the belief. This constituting structure entails that these explicit attitudes—unlike implicit attitudes, which are constituted by much more loosely bound sets of dispositions—are apt to play a role in the integration of moral agency. Explicit attitudes integrate agents in two ways: when they are employed as premises in top-down reasoning, and when they activate the contents they entail.

The canonical way of distinguishing explicit attitudes from implicit is by reference to what agents are disposed to assert. Now, only attitudes with contents constituted by normatively constrained sets of dispositions have assertible contents. Implicit attitudes can be described, but they cannot be expressed in propositional form: if, for instance, I have a phobia with regard to spiders, I cannot express the content of my phobia in propositional form. I can say that I am disposed to respond as if I believe that spiders are dangerous, but it would be false for me to assert that I believe spiders are dangerous. Because there is no proposition expressing (as opposed to describing) their contents, implicit attitudes cannot be utilized as premises in reasoning. They can neither be asserted as premises, nor are there any justifiable inferences from implicit attitudes. Inferences require appropriately structured content because loose (or narrow) content is not sufficiently consistent and coherent to entail further propositions; loosely structured or narrow contents, if they can be said to entail other contents at all, entail contradictory or underspecified propositions. We might be said to make as if inferences from as if beliefs—at least if we count the conclusion of a practical syllogism as an inference—but we do not take these inferences to be justified. Rather, we typically regard these as if inferences
as mistakes. I am embarrassed by my flinching from a spider, because I do not take the response to be justified.

Explicit attitudes, when they are utilized as premises in effortful reasoning, generate contents in ways that are guided by the rules the agent (implicitly) accepts as normative. Because the contents of a belief are structured in this way, explicit beliefs can be utilized as premises in reasoning. They have contents that are assertible, and they can be used to generate further contents in ways that are justified. These contents, because they are constrained by norms of reasoning, are linked with one another in ways that are coherent, and therefore can impose coherence on the agent herself. Effortful reasoning can be used to generate plans and policies by which the agent guides her behavior. Equally, when the content of a plan is explicit it entails a consistent set of actions aimed at fulfilling it.

Because only an attitude with contents bound together in a way that is, or mimics being, rule-based can support justified inferences, and because only an attitude with a content the agent can assent to or assert can be deliberately employed in effortful reasoning, we can be confident that so far as top-down processing is concerned, the difference between explicit and implicit attitudes is stark. It is less obvious that explicit attitudes will, by virtue of their contents, integrate behavior even when they are not employed in top-down processes. The claim seems plausible: given that an explicit attitude is constituted by, or entails, a broad range of integrated dispositions, it seems likely to activate related content in a manner that clusters around these dispositions, thereby mimicking rule-based processing. There is also experimental evidence in favor of the hypothesis.

Galdi, Arcuri and Gawronski (2008) 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. But they overlook another possibility: that consciousness of the attitude influences unconscious attitudes. That is, rather than the having of a view (perhaps dispositionally) influencing our implicit attitudes, it might be the self-attribution of the view (in overt speech or to oneself) that has this effect on our attitudes. The evidence seems to better support this suggestion than their hypothesis. Given that the issue about which the subjects were asked was one of lively current interest, it seems unlikely that most of the subjects who reported an attitude at \( t \) were making their mind up at \( t \); rather, they were reporting a view they likely had already formed, perhaps dispositionally. But, given that Galdi et al. were able to detect changes in decided participants’ implicit attitudes across the relatively short span of a week, it is likely that the reporting (and thereby becoming conscious of) the attitude was affecting the rate of change: were it merely the having of the attitude that was driving change, one would expect the process of change to have been underway for some time prior to testing, and therefore one would expect the rate of any further change to be slow. Hence it seems likely that it is not the having of a view, but the reporting of it that was driving the effect.
This suggestion is also supported by Ap Dijksterhuis’s work. Dijksterhuis claims to show that unconscious processing is often superior to conscious processing, but, as he accepts, unconscious processes can produce superior results only subsequent to conscious encoding. Consider this experiment (one of many along similar lines reported in Dijksterhuis & Nordgren 2006). Subjects were asked to form an impression of a person who belonged to a minority group. They were then given information about the person, some of which was congruent with the relevant stereotype and some of which was incongruent. Later their judgments about and recall of information about the person was tested. Participants who were requested to think about the person consciously prior to testing recalled more stereotype-congruent information than stereotype-incogruent and formed a more stereotypical final judgment. Participants who were given a distractor task did not demonstrate stereotyping. Dijksterhuis and Nordgren suggest that the unconsciously activated stereotype affected how subjects consciously processed information. What matters for our purposes, though, is that conscious activation of relevant information by itself—that is, in the absence of further conscious processing—has effects in line with the normative, and not the associative, structure of representations, which supports the view argued for here. Dijksterhuis (2004; experiment 5) also showed that unconscious processing of consciously encoded information led to better organization of recalled material, again demonstrating that the content activated by consciously encoded material is normatively related to it.

The suggestion that consciously recalled information activates other contents in a manner that seems to mimic rule-based processing coheres nicely with Carruthers’ claim that it is only perceptual events, together with events (like visual and auditory imagery) that use the machinery of perception, that are globally broadcast. Simply having an attitude accessible to report will not globally broadcast that attitude, but reporting or rehearsing it—and thereby becoming conscious of it—will. This suggestion is also consistent with rival views belonging to the integration consensus: those that make conscious contents accessible to (rather than broadcast to) consuming systems, for instance.

If it is the reporting that—by activating semantically related attitudes—has an integrating effect, then we should see more integration with regard to explicit attitudes than implicit. However Galdi et al. (2008) might also serve to motivate an important objection. For those subjects who remained undecided on initial testing, Galdi, Arcuri and Gawronski found that their implicit associations at t predicted their conscious attitudes at t1. They hypothesize, plausibly, that our implicit attitudes may bias our processing of new information, making us more likely to come to a conscious view that is consistent with these attitudes. Doesn’t this demonstrate the symmetry of explicit and implicit attitudes? Doesn’t it show that whenever there is a conflict between the two, processes that will tend to bring them into line will be activated? I suggest that there are good reasons to think that there are significant differences in the ways in which this process occurs.

First and most obviously, implicit attitudes are activated, whereas explicit attitudes may be utilized as well as activated; utilized, that is, as premises in effortful reasoning. Only the latter is an exercise of agency: something we do, rather than
something that happens to us. Second, though, there is plentiful evidence that when explicit attitudes are activated, the effect they have on other attitudes is different from the effect that implicit attitudes have. Nonconscious attitudes exert their causal power in an associative manner. Part of the evidence for this claim comes from cognitive load manipulations. We have good reason to link these manipulations to implicit attitudes: when there is a conflict between explicit and implicit attitudes, cognitive load manipulations typically have the effect of bringing behavior closer into line with the latter than otherwise. There is a substantial body of evidence that cognitive load manipulations decrease the degree to which reasoning is in accordance with the norms of logic (De Neys 2006; DeWall, Baumeister and Masicampo, 2008). DeWall, Baumeister and Masicampo (2008) also found that increasing conscious motivation to be logical—by telling subjects they would need to explain their answers—increased normatively correct responses, while priming the idea of logic appears to activate the motivation to appear logical, but failed to improve performance. In fact, when there was a conflict between validity and truth (for instance when subjects were asked to assess the validity of arguments with obviously false conclusions), increasing unconscious motivation to be logical led to a decline in performance.

Part of the reason that unconscious information processing is more associative and less rule-based than conscious is that unconscious processes are blind to the structure of formal systems, but it is this structure that is essential to rule-based reasoning. Unconscious systems can process semantics but not syntax: even two word phrases cannot be used as primes as a unit (if two words are presented as a prime, each has an independent priming effect (Baumeister and Masicampo 2010)). Similarly, the unconscious is blind to negation (Deutsch, Gawronski & Strack 2006). Activating concepts unconsciously has effects on subjects’ attitudes, but these effects are associative and not logical.

Of course, we are not assuming that implicit attitudes are unconscious attitudes (that is, attitudes with contents of which we are unaware). Nevertheless, the lesson of cognitive load manipulations appears to be that implicit attitudes work in the same way as unconscious contents, activating semantically related representations in the same associative manner, whenever we lack the awareness or cognitive resources to inhibit or trump these processes. Being aware of their contents does not prevent them from exerting their causal influence associatively. It is this associative influence that we should expect, given their structure. Note that the claim that (some) explicit attitudes will activate semantic representations in a manner that tends to mimic rule-based reasoning, and therefore will have a unifying effect on behavior that is closer to what would be expected given such reasoning, is empirically tractable. Though there is experimental evidence in its favor, more evidence is needed before we can be very confident that it is true.

**Attitudes and Agency: Toward a Conclusion**

As we saw, there are good grounds for holding that the ability to pursue plans and projects across time is an essential component of moral agency. So understood,
agency depends upon unification of the person; an agent has a relatively consistent set of beliefs and desires, and is able to ensure that she acts upon those beliefs and desires. This unification depends upon explicit attitudes, I propose, because only such attitudes can cause broad and integrated behaviors. Explicit attitudes are employed by the agent to impose unity effortfully, by being taken as premises in reasoning, and through their role in coordinating plans and projects. Pursuit of plans and projects requires rule-based processing, not associative. Plans are typically complex: projects of the kind that are distinctively human typically involve the performance of actions which are not linked to the goal of the plan in a way that is simple enough for unconscious processing. Very often, our plans concatenate series of actions that seem unrelated to the ultimate goal (for instance, executing my plan to write a paper might involve my planning to get a good night’s sleep, or putting aside time to mark essays so that I can subsequently write without interruption). Our projects also often involve tower-of-Hanoi like actions in which we pursue a goal by means that seem to take us further away from it—for instance, deleting a draft of a paper and beginning again. More prosaically, successful pursuit of a project involves choosing the right time for execution of its components and the correct order in which to do it. All this is beyond the capacity of associative processes, except insofar as the agent has already acquired an overlearned script for the plan. Sometimes, we have such scripts, but the environment is rarely stable enough for their development.

Unity might also be achieved by automatic processes, but insofar as this is the case, it will depend upon the conscious encoding of explicit attitudes, in particular those with broad and coherent structures. Activation of the associatively linked structures of implicit attitudes will have no such unifying effect. Explicit attitudes will tend to activate semantically related representations in a way that matches justifiable inferences more closely than will implicit, given their more consistent structure.

The differences between the associative structure of implicit attitudes and the more rule-based (like) structure of explicit attitudes are directly relevant to our moral responsibility for actions caused by each. Moral agency—distinctively human agency more generally, including the kind of agency required for the pursuit of intellectual and artistic projects—requires the unification of behavior, and that requires that we become aware of the content of our attitudes. To that extent, the capacity for morally responsible agency depends upon awareness. There are also good reasons to think that responsibility for token actions requires that the attitudes that cause them are explicit. Here is not the place to defend a particular account of moral responsibility. In the space available, all I can do is to briefly sketch some of the reasons to think that on any of a very wide range of extant theories, agents should be excused responsibility for actions caused by implicit attitudes.

On ‘real self’ views, agents can only be held responsible for actions caused by the mental states that constitute their real selves or practical agency (Wolf 1990; Smith 2008). Real self theories need a criterion to distinguish between attitudes that belong to the agent’s real self and those that do not. On a plausible view, attitudes belong to the agent’s real self if they help to constitute a coherent perspective from
which she evaluates options and states of affairs. Call this perspective the agent’s practical identity. On a real self view so understood, agents would be responsible for actions that express their practical identity. If the preceding claims about the ways in which explicit attitudes integrate agents are true, integration of attitudes into the agent’s real selves is, as a matter of contingent fact, brought about by awareness of the content of attitudes. Only such attitudes would be genuine components of the agent’s real self because only such attitudes can help to constitute a perspective: a view on the world with relatively consistent entailments for action and for belief formation. Implicit attitudes, attitudes with thin contents more generally, cannot form a component of the agent’s practical identity because they have underspecified and often inconsistent entailments. They cannot be used in planning (though sometimes they may need to be taken account of; the agent may need to take steps to avoid triggering them or having them bias her behavior in ways she does not endorse), and when they have contents the agent rejects they are alien intrusions into her agency, not components of it.1

On other views of moral responsibility, too, agents ought to be excused moral responsibility for actions caused by attitudes with contents of which they are unaware (though some may need to be revised to accommodate the discovery that there are no conscious attitudes, in the sense urged by Carruthers). On identification theories, such as those associated with Frankfurt (1971) and Watson (1975), agents are responsible for behaviors caused by states with which they identify. If we revise these views so that agents are instead responsible for actions caused by attitudes with which they are appropriately identified (a revision that is attractive for independent reasons), such views will require that agents have a self; once again, the constitution of this self will be brought about by attitudes with contents of which the agent is aware. At least some historical accounts of moral responsibility, according to which agents must have assumed ownership of the mental states or mechanisms that cause their actions in order to be responsible for them (Mele 2006; Fischer and Ravizza 1998), will also excuse agents for actions caused by attitudes with contents of which the agent is unaware. If the historical condition is understood as requiring the integration of the attitude into the agent’s self-conception, consciousness of the content of the attitude will be required to satisfy it. It is also likely that agents will be excused responsibility for actions caused by attitudes with contents of which they are not conscious on control-based theories, for reasons we have already glimpsed. Actions caused by implicit attitudes, like the behaviors seen in the alien hand syndrome, are out of the agent’s control. Since they are not caused by the system of values constitutive of the agent, but instead only by a narrow set of representations, they are not controlled by the agent (a fact dramatically illustrated by efforts that sufferers from the alien hand make to reassert control by indirect means; say, tying the hand down). On control-based views, agents should be held responsible for actions caused by attitudes the contents of which the agent is concurrently aware. On other historical views, concurrent awareness may not be required, but awareness of the contents of the attitude expressed must have played a role in the causal history of the formation of the agent’s practical identity.
Considering an objection will allow us to get a clearer view of the essential role that explicit attitudes play in moral agency. The objection turns on the following, plausible enough, observation: people are sometimes driven by unconscious attitudes to perform a broad range of actions. When this occurs, it might be held, their lives as agents are unified by attitudes with contents of which they are unaware; thus, it is false that consciousness of the contents of our attitudes is required for the integration of our lives. As an example, consider a man whose life plan seems to be driven by a desire to rebel against his father. In this case, the man’s life may be unified by his implicit attitude toward his father, an attitude with a content opposed to his explicit attitudes. Since implicit attitudes can integrate our lives, they can underwrite moral agency as well as explicit attitudes.

Case like this one may be quite common. When they occur, it is true that implicit attitudes play a role in imposing a kind of unity on a life. But the kind of unity they impose is not the kind of unity required for moral agency, I suggest. Moral agency, I have claimed, is a diachronic property: it is only insofar as we are able to pursue plans and projects across time that we are moral agents. Such agency requires that our mental states be relatively unified, so that we are able to pursue our plans across time without later person-stages undercutting the projects of earlier person-stages. The kind of unity imposed on a life by implicit attitudes does not enable a person to pursue projects across time. Rather—to the extent to which implicit attitudes conflict with explicit—it will have the opposite effect: undermining our personal-level plans and projects. The person who is powerfully motivated by an implicit attitude of which she is unaware is likely to engage in self-defeating projects, as her implicit attitudes cause her to (for instance) interpret evidence in ways that are motivated, rather than on its merits. Even coming to be aware of the content of her implicit attitude will not suffice to allow her to pursue her projects successfully by itself; she will need to be constantly on guard against its influences on her judgments. Unity owed to implicit attitudes is not the kind of unity we require as moral agents; it is actually in conflict with such unity.

Further, for reasons we glimpsed earlier, explicit attitudes—attitudes with assertible contents, of which we are aware—are required for agents to have moral attitudes, and therefore to be capable of expressing such attitudes in their behavior. As we saw earlier, our attitudes should not be identified with any one representation, but with sets of representations. These sets are unlikely to be perfectly consistent but they must be relatively coherent to constitute a genuine personal-level attitude. Such integration is, as we seen, the domain of explicit attitudes: the having of conscious attitudes allows for the emergence of genuine moral (and nonmoral) attitudes, and therefore is required for our actions to express our moral attitudes. Integration is required for moral agency because it is required for us to be able to pursue plans and projects across time. It is required for the expression of moral attitudes because only an integrated agent has personal-level attitudes at all.

Implicit attitudes are not—necessarily—unconscious attitudes (though the relative difficulty of accessing their contents may explain why they have often been labeled unconscious). We are sometimes, and perhaps often, aware that we have them. If Carruthers is right, the route by which we discover their contents is of
the same broad kind as the route whereby we become aware of the contents of our explicit attitudes (though the route is likely to be more convoluted in the former case than the latter). Nevertheless, implicit and explicit attitudes influence behavior and dispositions in different ways. Explicit attitudes—attitudes which have assertible contents, of which we are aware—alone unify agents in the manner required for moral agency.

We can, finally, put to rest the challenge that King and Carruthers gave us. Consciousness of the content of our attitudes does matter to our moral responsibility. Even if coming to be aware of their contents does not transform them into conscious attitudes, knowing this content makes an important difference. However, it is only for those attitudes whose content we can report in direct propositional form that this is true. Awareness of the content of our implicit attitudes—though undoubtedly morally significant in all kinds of ways—cannot play an integrative role in agency because implicit attitudes have too loose a structure either to be used as premises in reasoning or to activate contents in a way that mimics rule-based, effortful, processing. These attitudes cannot therefore underwrite the capacity for morally responsible agency, and even for agents who possess this capacity, they routinely cause the capacity to be bypassed. The folk intuition that consciousness matters can be vindicated, though the picture of agency that emerges is one that departs substantially from the folk view.

Acknowledgement

The author gratefully acknowledges the support of the Australian Research Council.

Notes

1 Since King and Carruthers themselves focus on real self views, it might be worth saying a word about how the claims here avoid the criticism they make in their paper. King and Carruthers argue that the real self can only play the role of distinguishing between attitudes that genuinely belong to the agent and those that do not if it contains conscious attitudes, of the kind upon which Carruthers cast doubt. Without such attitudes available to play the role of endorsing agents’ mental states, such views would be implausible, they claim. However, we need not conceive of real self views as requiring endorsement, in the manner they suggest. Rather than attitudes being annexed to an agent’s real self by an agent’s conscious endorsement, attitudes should be considered as appropriately attributed to the agent if they are integrated into her practical identity. It is an empirical question what is required for attitudes to be so integrated. I have argued that such integration—at least typically—requires awareness of the content of the attitude integrated.

2 I owe this objection, and the example discussed, to a referee for Noûs.

References


De Neys W. 2006. Dual processing in reasoning: two systems but one reasoner. Psychological Science 17: 428–33


