NOÛS

NOÛS 49:4 (2015) 800–823
doi: 10.1111/nous.12074

Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements[1]

NEIL LEVY
Florey Institute of Neuroscience and Mental Health

Abstract

Implicit attitudes are mental states that appear sometimes to cause agents to act in ways that conflict with their considered beliefs. Implicit attitudes are usually held to be mere associations between representations. Recently, however, some philosophers have suggested that they are, or are very like, ordinary beliefs: they are apt to feature in properly inferential processing. This claim is important, in part because there is good reason to think that the vocabulary in which we make moral assessments of ourselves and of others is keyed to folk psychological concepts, like 'belief', and not to concepts that feature only in scientific psychology: if implicit attitudes are beliefs there is a prima facie case for thinking that they can serve as the basis for particular kinds of moral assessment. In this paper I argue that while implicit attitudes have propositional structure, their sensitivity and responsiveness to other mental representations is too patchy and fragmented for them to properly be considered beliefs. Instead, they are a sui generis kind of mental state, a state I dub *patchy endorsements*.

There is a great deal of evidence that many people have implicit attitudes that conflict with their considered beliefs, and that these implicit attitudes sometimes affect their explicit judgments and their actions.[2] For instance, negative implicit attitudes toward black people seem to play a role in explaining differences in political convictions and in a range of judgments—opposition to Barack Obama's healthcare reforms (Knowles, Lowery & Schaumberg 2010), the judgment that blacks bear more responsibility than whites for the riots that followed the acquittal of the police officers accused of assaulting Rodney King (Fazio et al. 1995), and a preference for white job applicants over equally qualified black applicants (Dovidio and Gaertner 2000). Negative implicit attitudes also predict a range of much more subtle behaviors that it seems appropriate to describe as racist: for instance making less eye contact with a black person than a white person (Dovido et al. 1997). Overt racism has declined in the United States and other Western countries very significantly over the past 40 years, yet dramatic racial inequalities persist; it is likely that negative implicit attitudes toward minorities play some role in explaining this fact (Pearson, Dovidio & Gaertner 2009).

Given the range of evidence suggesting that implicit attitudes play a role in explaining discriminatory behavior, understanding their nature is a pressing concern. A better understanding of what implicit attitudes *are* may enable us to avoid their

inculcation, in ourselves and in others, and to prevent their expression when they are inculcated. Further, understanding their nature is central to a proper assessment of the agents who harbor them. As Ron Mallon (forthcoming) has recently emphasized, our attributions of responsibility and other kinds of moral assessment are keyed to folk psychological concepts; to the extent to which implicit attitudes are different kinds of beasts, we may find ourselves at a loss when it comes to moral assessment. For instance, discovering that someone harbors negative *beliefs* about black people, even implicit ones, may warrant a higher degree of moral condemnation than a finding that they harbor negative *associations* alone. Perhaps the first might justify the epithet "racist", when the second does not.[3]

Several philosophers have recently suggested, on the basis of evidence that implicit attitudes have propositional structure, that they are beliefs (Smith 2005, 2012; Egan 2011; Mandelbaum 2013; Mandelbaum unpublished) In this paper, I shall argue that implicit attitudes are not beliefs, not even unconscious beliefs.[4] Though they are more than mere associations, they are not sensitive enough to evidence and to the semantic contents of other attitudes to qualify as beliefs. Rather, they are what I will call *patchy endorsements*.

In section 1 of the paper, I describe the ways in which implicit attitudes are standardly probed by psychologists, in order to lay out the background for the arguments to come. In section 2, I sketch the current debate concerning the nature of implicit attitudes. This debate has focused on how much propositional structure they exhibit, with most psychologists contending that they are essentially associations, and therefore have little structure. The associative view has difficulty accommodating a range of evidence—forcefully driven home in recent work by Eric Mandelbaum—that implicit attitudes are involved in seemingly inferential processes: explaining this evidence requires us to attribute propositional structure to implicit attitudes. But evidence of propositional structure is not sufficient to establish that implicit attitudes are beliefs. The evidence that implicit attitudes have propositional structure consists in evidence that they respond to the semantic contents of other states, and beliefs are states that respond in this kind of way, but beliefs do this *systematically.* If implicit attitudes are states that respond to semantic contents in a patchy and fragmented way, they are neither associations nor beliefs.

In section 3 I argue that implicit attitudes—specifically the implicit attitudes that cause discriminatory behavior—are neither associations nor beliefs. They are not associations for the reasons set out by Mandelbaum: they feature in content-driven transitions. But they are not beliefs because their responsiveness to content is fragmented, and often the responsiveness they exhibit is of the wrong sort to count as genuine inference. The upshot, I claim, is that implicit attitudes are not beliefs. Nor are they associations. What, exactly, are they? Something for which folk psychology lacks a word: patchy endorsements. If this conclusion is correct, we ought to be hesitant in applying our vocabulary of moral assessment to ourselves and to one another on the basis of finding that we harbor implicit attitudes: some kinds of moral assessment may need to await the development of a theory better suited to the kinds of psychological states that implicit attitudes are.[5]

1. Implicit Attitudes: The Basics

There are a number of different tasks psychologists use to measure implicit attitudes. Some tasks measure alleged neural or physiological correlates of negative implicit attitudes toward members of particular groups, such as amygdala activation and eyeblink startle response (Phelps et al. 2000); others probe language use (on the basis that subjects use more abstract language to describe behaviors consistent with their expectations; Maass et al. 1989) or utilize word completion tasks (Dovidio et al. 1997). However, two kinds of paradigms have dominated the field: those measuring response latencies and those utilizing sequential priming.

Response latencies are simply the amount of time taken to respond to stimuli. They are held to be a reliable measure of implicit attitudes because there is convergent evidence that subjects are quicker to associate two stimuli to which they have the same attitude (both positive or both negative) than two stimuli to which they have discordant attitudes. The best known (though by no means the only) task used to measure implicit attitudes via response latencies is the Implicit Association Test (IAT), versions of which have now been taken by literally millions of people (Nosek et al. 2007). In a standard IAT, subjects press buttons to associate an image with a word or a concept: say black faces with the word "good" (or "flower" or "sunshine") and white faces with the word "bad" (or "cancer" or "dirt"). The contingency changes from block to block: on some blocks, the task will be to associate black faces with positive words or concepts, and on some with negative words or concepts.[6] Typically, subjects are faster to associate positively valenced words with white faces than with black faces and faster to associate negatively valenced words with black faces than with white faces (Dasgupta 2004).

Sequential priming tasks involve two stages. First the subject is exposed to a prime: a stimulus that may dispose her toward some responses and away from others. In tasks measuring implicit attitudes toward race, the prime may be a black or white face, or a word that is associated with race (stereotypically black and white names may be used, for instance). The subject then performs a task that produces quantifiable data: for instance, rating the attractiveness of an unrelated image. If the first stimulus functions as a prime, her rating of the image will be influenced by exposure to it; rating images as less attractive after exposure to black faces than after exposure to white faces indicates negative implicit attitudes toward black people. The affect misattribution procedure developed by Keith Payne (Payne et al. 2005) uses Chinese pictograms for images and black and white faces as primes. Another sequential priming task developed by Payne is the weapon identification task (Payne 2006). In this task, subjects are asked to identify whether an ambiguous object is a gun or a harmless object, after exposure to a black face or to a white face. A significantly greater likelihood of misidentifying a harmless object as a gun after exposure to a black face than a white face is considered a measure of some kind of negative implicit attitude towards blacks.[7]

Most researchers believe that implicit attitudes, as measured by these methods (and many others) predict behavior, with implicit measures outperforming explicit at predicting some behaviors (Greenwald, Poehlman, Uhlmann and Banaji 2009).

Oswald et al. (2013) have recently claimed that the IAT is a "poor" predictor of behavior. In fact, even if we can take the results of their meta-analysis at face value (rival meta-analyses are in preparation, which apparently show a greater effect size), the claim that the IAT is a poor predictor of behavior seems to be false: even on their analysis, it remains true that the IAT reliably predicts a small amount of variance in behavior. Small effects matter: even on the most conservative estimate of the predictive power of implicit attitudes, they may cause thousands of small slights and instances of disrespect over the life of a single black person. Moreover, in many of the circumstances in which implicit attitudes make a difference to behavior, small effect sizes may make big differences: between being hired or not, for instance (or between being taken to carry a wallet or a gun).[8] Though the IAT is by far the most studied measure of implicit attitudes, and therefore the paradigm upon which researchers have concentrated to assess their predictive power, other approaches may turn out to be better predictors of some behaviors. None of the papers cited in the introduction as evidence that implicit attitudes predict behavior used the IAT; for instance, Knowles, Lowery & Schaumberg (2010) used the GNAT task, in which subjects respond as quickly as possible by pressing a button when a stimulus matches the categories of interest, and refrain from responding when it does not, to measure implicit attitudes. They found that the attitudes measured predicted opposition to "Obama care". Work on the affect misattribution procedure looks especially promising as a predictor of behavior (though it may be that its predictive power is limited to those subjects with more extreme attitudes, given the evidence that it is comparatively insensitive to non-extreme attitudes (Bar-Anan & Nosek forthcoming)). Payne et al. (2010) used it to measure the impact of implicit and explicit anti-black prejudice on voting behavior in the 2008 presidential election. They found that though negative implicit attitudes toward black people did not predict a greater likelihood of voting for McCain once explicit prejudice was controlled for, they did predict a greater likelihood of not voting for Obama—either voting for a third-party candidate or abstaining altogether. It is plausible to interpret these results as showing that implicit and explicit beliefs interact, with an explicit commitment to racial equality moderating the influence of conflicting implicit attitudes. Negative implicit attitudes toward black people do not cause prejudiced behavior on their own, but they nevertheless significantly modulate behavior by interacting with explicit attitudes. On any account of causation, these attitudes are causes of behavior.[9]

2. The Nature of Implicit Attitudes

Standard accounts of implicit attitudes model them as *associations*, which arise from agents' learning history. Frequently paired representations, or representations and evaluative responses, come to be associated, such that activation of one automatically activates or makes more accessible the other; the strength of these associations is a function of the frequency of the pairing and the reward value of the representations. Despite their differences, the most influential models of implicit attitudes share a commitment to the associative picture. According to 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), implicit attitudes are associations. Prominent philosophers have also endorsed this view (Gendler 2008).

Recently, however, rival models of implicit attitudes have been proposed. Though he stops short of endorsing it, Jan De Houwer (in press; 2014) has suggested that the field ought to seriously explore a propositional model of implicit attitudes. According to a propositional model, implicit attitudes are *psychological states with propositionally structured contents*. That is, they have constituents that are bound to one another in a determinate manner, in a manner closely akin to the way in which natural language sentences have determinately bound constituents (the close relationship between propositional structure and sentence structure is not accidental, of course: natural language sentences express propositions). If implicit attitudes are propositionally structured, then by virtue of their structure there is a highly specific relation between their constituents, whereas if they are associations there is no specific relation between their constituents. If implicit attitudes are propositionally structured, they have satisfaction conditions, which are a function of their structure: they are satisfied only if their intentional objects stand in the appropriate relations to one another. If they are mere associations, they lack satisfaction conditions (Mandelbaum 2013).

Associative models of implicit attitudes might seem to be strongly supported by the fact that much of the evidence for their existence is drawn from experimental paradigms putatively designed to measure associations, like the IAT. As proponents of propositional models point out, however, this evidence is equivocal between the two models. First, almost no one denies that agents may have associations between bona fide beliefs. Second, much of the evidence described as indicative of associative processing might equally reflect unconscious inference. Agents might *associate* black faces, and other representations that cause thoughts about black people, with violence, or they might *believe*—or at any rate token a representation with assertoric force with the content that—black people are dangerous. These different models support quite different accounts of the nature of the cognitive processes that drive behavior, and different conclusions about the people who harbor such attitudes.

Suppose my IAT performance indicates that I associate "woman" with "family" and "man" with "work". If the implicit attitudes revealed thereby are mere associations, then this fact gives us no reason to conclude that "woman" and "family" stand in a particular relation for me; that is, that I take the world to be a certain way. We can conclude only that thanks to my learning history, representing "woman" activates or makes accessible "family" (and related representations). But if my implicit attitudes are propositions, then "woman" stands in a particular relation to "family" for me. The nature of that relation will differ from attitude to attitude. Again, comparison to natural language sentences helps to illuminate the difference. The natural language sentence "Bert loves Ernie" expresses the proposition that Bert stands in the loving relation to Ernie. The inveterate watcher of *Sesame Street* probably associates "Bert" and "Ernie", but in virtue of associating them does not represent them as being in any particular relation to one another.

Proponents of propositional and associative models of implicit attitudes are committed to quite different explanations of how implicit attitudes cause behavior, for example the behavior of a man who selects a worse qualified male candidate over a better qualified female candidate as a consequence of implicit bias. If the implicit attitude at issue is a mere association, the explanation of his behavior might be something along the following lines: his association of "men" with "work" made work-related concepts more accessible to the relevant mental mechanisms and thereby affected the relative salience of, say, the man's qualifications over the woman's. But if his implicit attitude is propositionally structured, then his behavior is (in paradigmatic) cases explained by the fact that he takes there to be a determinate relation between the constituents of his attitudes. Perhaps the content of his implicit attitude is "women are not well suited to work"; because he endorses that proposition (perhaps unconsciously), he is disposed to act in ways that accord with it.

Some advocates of propositional models go further, maintaining that implicit attitudes have the propositional structure proper to *beliefs*. Assessing this claim requires us to identify the features of beliefs that distinguish them from other states. This is not the place to articulate and defend an account of the nature of belief. It seems relatively uncontroversial, however, to say that whatever beliefs *are*, they play a distinctive role in thought and—thereby—behavior. The major views on the nature of belief—dispositionalism, representationalism, functionalism and interpretationism—are all committed to the claim that beliefs respond to evidence, both by featuring in inferences and by altering when they conflict with the evidence (Schwitzgebel 2010). The basis of this commitment differs from view to view: on some, beliefs are internal states which play this role in virtue of their syntax, while for others responsiveness to evidence is criterial for belief attribution.

More specifically, beliefs may be distinguished from other states by reference to two closely linked characteristics. Beliefs are *inferentially promiscuous* and beliefs are *responsive to evidence*. Beliefs are inferentially promiscuous inasmuch as the belief that *p* can interact (appropriately) with any other propositional attitude (Stich 1978). For instance, my belief that it is raining will interact appropriately with my desire to stay dry, as well as my belief that roads can be dangerous when wet, and any other of my attitudes concerning water and wetness. Whereas inferential promiscuity is a matter of how beliefs cause behavior and update other mental states, responsiveness to evidence is a matter of how the belief itself can be expected to update, given appropriate evidence. Inferential promiscuity and responsiveness to evidence are two sides of the same coin: beliefs are inferentially promiscuous, causing the update of other beliefs, because beliefs are responsive to evidence. Any state which is inferentially promiscuous and appropriately responsive to evidence is a belief; accordingly, I will follow Brownstein and Madva (2012) in taking this kind of responsiveness to be the mark of a *bona fide* belief.

Inferential promiscuity and responsiveness to evidence are functional notions. We can infer the extent to which a state is inferentially promiscuous and responsive to evidence by careful experimentation focusing on the role of attitudes in behavior. If we can identify an attitude and control for the effects of context and other states on behavior, we may be able to assess the extent to which the attitude combines

with other states to cause behavior and how it updates given relevant evidence. This fact allows us to assess the extent to which implicit attitudes are beliefs.

Notoriously, the beliefs of actual human beings are far from perfectly responsive to evidence or apt to feature in inference. Nevertheless, most of our beliefs are sufficiently responsive to evidence to justify the attribution of (literally) innumerable beliefs to us. How much responsiveness to evidence must we exhibit to make attribution of the correlative belief appropriate? As the debate over the so-called doxastic conception of delusions reveals, this is a hotly contested issue. Delusions seem to respond to evidence relatively badly and to be relatively encapsulated from other attitudes. A sufferer from Capgras might, for instance, assert that his wife has been replaced by a robot, but fail to express any concern for the "original" or to report her disappearance to the police. This failure to draw inferences has led some philosophers to conclude that sufferers from delusions don't really believe their delusions (e.g., Currie and Ravenscroft 2002). In response, other philosophers have pointed to the fact that some degree of insensitivity to evidence and encapsulation from other attitudes is characteristic of many (apparent) beliefs, and argued that delusions are no worse on this score than other doxastic states that deserve to be counted as beliefs (Bortolotti 2009). Both sides would surely agree, however, that excessive evidence insensivity and encapsulation blocks the ascription of a correlative belief to an agent.

Of course, we might find ourselves in the same boat with regard to implicit attitudes as delusions: we might find that the available evidence is equivocal between beliefs and some other, non-doxastic, state. In that case, settling our question will require settling the debate between proponents and opponents of doxastic accounts of delusions, by motivating a more precise answer to the question how much evidence sensitivity and aptness for inference is enough, or by establishing that the matter is ineliminably vague. In advance of investigation, however, there is some reason to be optimistic: we may find the degree of sensitivity and aptness for inference to be sufficient to provide a definitive answer, without us needing to make any progress on the issues on which the debate about delusions turns. Though it may be ineliminably vague just how much responsiveness to evidence is required for a representational state to count as a belief, sufficient departure from the kind of sensitivity to evidence and aptness for normatively respectable inference we associate with a *bona fide* belief will settle the question.[10]

Note one central difference between the debate over delusions and our current concern. Because the debate over delusions concerns an explicit (putative) belief, evidence of inconsistency in the behavior of sufferers is *ipso facto* evidence against the doxastic account. Matters stand very differently with regard to the content of implicit attitudes when—as in the kinds of cases that interest us—implicit and explicit attitudes diverge. Due to this divergence, we ought to *expect* to see inconsistencies in behavior, depending on which of the attitudes is dominant in a particular context. The behavior of interest to us is therefore that subset of behavior concerning which we have good reasons to believe that implicit attitudes are causally efficacious and with regard to which we can isolate the contribution made by the

implicit attitude. Of course this makes the inference from behavior to mental state rather tricky.

Whereas *any* delusion-related behavior is grist to the mill of the person assessing the content of a delusion, valid inference about the content of implicit attitudes requires isolating the causal contribution they make. As a consequence, our primary evidence will come from laboratory-based studies, in which a variety of techniques are employed to probe the content of implicit attitudes by increasing the degree to which they are causally efficacious (requiring rapid responses, cognitive load manipulations, or sequential priming techniques, for instance). The debate between advocates of rival accounts of the structure of implicit attitudes is best understood as a debate concerning which account offers the best explanation for the responses of subjects in these experiments. Interpreting this data is challenging: *bona fide* beliefs can trigger associations, and the activation of one representation by another may mimic an inference, even if it is in fact associative (I may associate *q* with *p*, such that I appear to infer *q* from *p*). Given the amount of data available, however, we have every right to be optimistic that we can build a decisive case for one model or another. Are these responses driven by states that interact in ways that reflect their associations with other mental states—typically relations of contiguity and similarity; Payne & Gawronski 2010—or are they driven by processes that interact with other mental states propositionally (that is, in ways that mirror the semantic relations between them)? If these states do interact propositionally, how systematic is this inference—is it systematic enough to require us to hold that implicit attitudes are beliefs? If implicit attitudes interact propositionally, but do so in a patchy and fragmented manner, they are neither beliefs nor associations: they are instead (as I will argue) patchy endorsements.

The case for the view that implicit attitudes are beliefs has been advanced most systematically and powerfully by Eric Mandelbaum (2013; unpublished). Mandelbaum advances two main arguments for the belief model of implicit attitudes, one concerning the conditions under which implicit attitudes can be eliminated and one concerning how implicit attitudes interact with other attitudes. Mandelbaum claims that the evidence he cites shows not only that implicit attitudes are *not* associations, but more interestingly that they *are* beliefs.[11]

Mandelbaum's first argument aims to establish that implicit attitudes are not mere associations. Associations, he claims, can be eliminated only by extinction or by counterconditioning. A conditioned response is extinguished if it goes unrewarded a sufficient number of times: an organism that has learned that whenever a tone sounds pushing a lever produces food will eventually fail to react to the tone if the action is no longer rewarded. Counterconditioning sets up a rival and conflicting contingency—say by actively punishing lever pulling. Both of these methods are effective at breaking associations between stimuli and responses. If implicit attitudes are associations, extinction and counterconditioning ought to be effective in eliminating them. Mandelbaum (unpublished) draws on a range of studies to argue that extinction and counterconditioning are ineffective at eliminating implicit attitudes.

Mandelbaum's positive case—in favor of the view that implicit attitudes are beliefs—rests in part on the claim that while implicit attitudes are surprisingly unresponsive to extinction and counterconditioning, they are surprisingly responsive to argument. Associative attitudes ought not to respond to arguments (*qua* arguments: an argument for the conclusion that *p* might trigger representations associated with *p*). For instance, a human being who has been conditioned into learning that a tone predicts an aversive puff of air into one eye will flinch on hearing the tone, even if she has been told that the tone will not be followed by the puff of air. The response must be extinguished or counterconditioned; it cannot be *argued* away. Mandelbaum alleges that implicit attitudes *can* be argued away; indeed, that they are more effectively argued away than they are extinguished or counterconditioned. Mandelbaum claims that this responsiveness to argument is evidence that implicit attitudes interact with other attitudes in ways that reflect their semantic contents.

Here I focus on Mandelbaum's claims concerning implicit attitudes' alleged content-driven promiscuity, rather than his evidence that implicit attitudes are not responsive to counterconditioning and extinction. I ignore his negative case because it is not relevant to my concerns: as Mandelbaum recognizes, even if this argument shows that implicit attitudes are *not* associations, it does not establish that they *are* beliefs (whatever else they are, delusions do not seem to be associations; that does not settle the debate whether they are beliefs). In the next four paragraphs, I will summarize some of Mandelbaum's evidence for inferences over implicit attitudes. I do not aim to be comprehensive, but to give the flavor of the case. I will add further examples he discusses in responding to Mandelbaum's claims.

First, Mandelbaum provides evidence for what he calls the *binding* of contents. Consider the poison experiment (which Mandelbaum selects because Tamar Gendler (2008) utilizes it in arguing for the claim that "aliefs"—a broader class of attitudes which encompasses implicit attitudes—are associative). Rozin, Markwith & Ross (1990) had subjects watch while they poured ordinary table sugar from a brand-name box into two jars. They then asked subjects to affix two labels to the jars. One label had "not sodium cyanide, not poison" on it, as well as a picture of a skull and crossbones preceded by the word "not". The other label had "sucrose, table sugar" on it. Then two otherwise identical drinks were sweetened with sugar from each bottle, and subjects were asked to drink one as well as to note whether they had a preference between the glasses. Though many subjects professed indifference between them, there was a significant preference for the glass containing sugar from the jar labeled "sugar". Subjects explained this preference by noting that they were bothered by the "not poison" label.

As Mandelbaum (2013) points out, participants in the poison experiment did not experience a generalized fear response; rather, they were wary *in relation to the labeled bottle*. The affective response and motivational disposition *bound* to the bottle. Of course it is quite possible that participants did experience a generalized sense of anxiety: perhaps (for instance) we would find that their startle response was potentiated compared to controls unexposed to fear-inducing cues. Even if this is the case, however, the fact remains that their anxiety bound

preferentially to the bottle labeled "not poison". Were this not the case, their preference for the other bottle would be inexplicable (generalized anxiety should effect the bottles equally). Mandelbaum claims that this binding of attitude to referent requires that the attitude be propositionally structured. Binding is evidence for inference, presumably unconscious, on the part of the subjects: the contents of *this* bottle are dangerous; any drink that contains a substance poured from it is also dangerous.

Thus, evidence for binding is evidence of inferential processing (though whether it is ipso facto evidence that the states involved are beliefs is, as we shall, another question). Once we start looking for this kind of evidence, Mandelbaum claims, we find it everywhere, and sometimes involving inferences of great complexity.[12] Here's a second example explored by Mandelbaum: "celebrity contagion" (Newman, Diesendruck and Bloom 2011). A subset of participants are willing to pay more to purchase a sweater if it has previously been worn by a celebrity (e.g. George Clooney); moreover, the degree of contact with the sweater modulates the price they are willing to pay. Now the interesting thing here is that the effect can be reduced (though not eliminated) by a small intervention: telling people that the item of clothing has been laundered. Apparently this kind of contagion is susceptible to being washed out. This is interesting because it shows that subjects' beliefs about washing apparently interact—in ways sensitive to semantic content—with their implicit attitudes. Note that agents' learning histories will, presumably, not feature many temporal or spatial associations between 'laundry' and 'celebrity'. There seems no reason why washing ought to be associated with cleaning away some kind of ineffable property that adheres to the clothing worn by people (indeed, just because the property seems ineffable it is hard to see how any kind of associative learning *could* take place). Perhaps it might be suggested that subjects learn to associate 'washing' with 'absence of contagion', but the suggestion is on the face of it unfriendly to associationism, since contagion-avoidance mechanisms seem to be *content-specific* innate mechanisms, not the kind of thing postulated by associationists at all. Moreover, washing *lowers* the value of the Clooney-proximate sweaters, whereas on the associative story 'washing' is positively valenced and should therefore lead to positive associations. Finally, even if we grant that some kind of associationist mechanism might bind 'laundry' and 'contagion' we still need to explain why 'contagion' seems to transfer to the ineffable property that rubs off on clothes, and not something else. It seems to be the similarity *in content* and not proximal instantiation in time or space that explains the binding.

Mandelbaum (unpublished) also appeals to the enormous literature on cognitive dissonance to show that unconscious inference is pervasive. "Cognitive dissonance" refers to a conscious or unconscious state arising from the recognition (again, possibly unconscious) of a tension between two or more of an agent's mental states, or between at least one of her mental states and a state that she is disposed to self-attribute to explain her own behavior. Dissonance appears to be aversive; at any rate, it motivates behavior aimed at reducing or eliminating the conflict. The agent may change one or more of her attitudes to render them more consistent, or reinterpret her behavior, and so on. The effects may be surprising: for instance,

subjects can relatively easily be manipulated into misattributing attitudes to themselves by inducing them to perform actions that are inconsistent with their prior attitudes (Cooper 2007); subjects seem to attribute antecedently unlikely beliefs to themselves (eg, the belief that their tuition fees should rise), in order to explain their own behavior to themselves (eg, writing an essay that supports tuition fee increases). To make sense of much of the data, it seems necessary to attribute to the agents unconscious states and inferences over those states (Aronson & Carlsmith 1962). For instance, explaining some data seems to require us to postulate an unconscious state with the content I AM A GOOD PERSON *and* some kind of inference from that state (some kind of inference such as the following; I am a good person; a good person doesn't do things that are bad; what I just did isn't all that bad).

Mandelbaum's case is impressive. It certainly constitutes a persuasive argument for the existence of nonconscious inference. However, I shall argue that it falls well short of establishing that implicit attitudes are beliefs. In the next section, I will show that though he may be right that implicit attitudes feature in content-driven processes—and therefore that they have (some) propositional structure—they are not sensitive to semantic content in the manner characteristic of beliefs. They are therefore neither just associations nor are they beliefs (nor are they aliefs, insofar as aliefs are meant to be structured associatively). Rather, they are patchy endorsements. In the remainder of this section, however, I will show that some of Mandelbaum's evidence does not establish that *implicit attitudes* are involved in inferential transitions at all.

Mandelbaum's evidence seems to establish that a great deal of unconscious information processing occurs in ways that are sensitive to the semantic contents of representations. But of course that was a claim we already had good reason to accept. That claim follows from a set of other claims that are more or less uncontroversial. It follows from these claims:

(1) Beliefs are representations that feature in inferential transitions;
(2) People have beliefs;
(3) At any time, many of these beliefs are not conscious.

Together, these claims make it *very* likely that there will be a great deal of unconscious information processing involving ordinary beliefs (the claims actually entail this conclusion unless we think that beliefs have very different causal powers when they are not conscious). So the demonstration that there is unconscious information processing that is sensitive to semantics does not show that implicit attitudes are inferentially promiscuous unless we are presented with some reason to think that it is implicit attitudes that are involved. It might, for instance, be ordinary (dispositional) beliefs that feature in these inferential transitions, not implicit attitudes.

In fact, much of the data Mandelbaum cites probably does not involve any processes over implicit attitudes. Consider his claims about the inferences involved in cognitive dissonance paradigms. These inferences seem more likely to involve explicit attitudes than implicit. Indeed, direct tests of whether

cognitive dissonance paradigms change implicit attitudes seem to show that they do not (Wilson, Lindsey and Schooler 2000; Gawronski and Strack 2004). Admittedly Gawronski, Bodenhausen and Becker (2007) were able to alter implicit attitudes using a "spreading of alternatives" paradigm (in which subjects choose between two equally desirable items and subsequently come significantly to devalue the unchosen item). However, the likely explanation for this result is associative and not propositional: the chosen item came to be identified—that is associated—with the self. Evidence for this includes the fact that implicit measures of self-esteem were correlated with implicit attitudes toward the chosen item. When dissonance works by inferential processes, it seems that it works over explicit attitudes (though the inferential transitions are themselves inaccessible to introspection); otherwise it works associatively.[13]

What we learn from considering these paradigms is the need for care in ensuring that it is implicit attitudes that are involved in information processing if what we seek is evidence of the structure of these attitudes. It should be noted that with regard to most of the objects of intentional states, implicit and explicit attitudes probably correlate quite well. For this reason it is not enough to measure subjects' implicit attitudes and then demonstrate that they engaged in inferential processing over attitudes with contents that match those of their implicit attitudes to show that these very attitudes featured in inferential transitions: steps must be taken to show that the transitions were over the implicit attitudes and not the explicit. Even measuring implicit attitudes prior to and subsequent to the manipulation, and showing that they have changed, isn't sufficient to achieve this: evidence that implicit attitudes have changed is not direct evidence that the processes were over these same attitudes (equally, though, evidence that they have not changed is not direct evidence that they were not essentially involved in information processing, since beliefs often feature in inferences without altering). We need to exclude the possibility that implicit attitudes have changed due to pressure from explicit attitudes. There is evidence that discrepancies between implicit and explicit attitudes trigger more intensive information processing. The dissociation may be aversive and constitute pressure for attitude revision (Rydell & McConnell 2010). Forming new attitudes by inferential processes over explicit attitudes may often *also* cause the formation of matching implicit attitudes.[14] Given how rapidly implicit attitudes can be formed—with training lasting less than five minutes (Ranganath & Nosek 2008)—we cannot conclude that evidence of transformation in implicit attitudes is evidence of inferential processing over these attitudes.

A great deal of Mandelbaum's use of experimental data is vulnerable to the charge of overlooking the influence of explicit processing on attitude change.[15] Consider his evidence from "cognitive balance" (Mandelbaum unpublished). Cognitive balance is the phenomenon whereby subjects form impressions of other individuals or groups without knowing anything about them other than how some target individual or group, of whom they have already formed an impression, feels about or behaves toward that group. While associative processing might be able to explain how being well-disposed or being ill-disposed extends from target to new individuals or groups, via some kind of spreading of activation story, it is hard

to see how such a story can explain reversals of valence. Subjects come to be *well* disposed toward individuals to whom individuals they are *badly* disposed to behave badly, apparently on the basis of an implicit "the enemy of my enemy is my friend" principle. This is evidence of some kind of inferential processing, to be sure. But it is not evidence of inferential processing over implicit attitudes. The experiment Mandelbaum cites (Gawronski et al. 2005) *produced* the appropriate implicit attitudes but given that all stimuli were above threshold and there was plenty of time for effortful processing, this is not evidence that the inferences were *over* these attitudes.

3. What is Content-Driven Processing Over Implicit Attitudes Evidence For?

For evidence of inferential processing over implicit attitudes, we need to look to cases in which implicit and explicit attitudes diverge. Divergence of this kind makes it far more likely we will be able to dissociate the influence of implicit and explicit attitudes on information processing. Fortunately for us, interest in implicit attitudes is largely driven by cases in which there is divergence. Researchers look for divergence and concentrate on cases likely to feature it. For that reason, the kind of data we need is easily found. I now turn to this data.

Much of the evidence is equivocal: experimental findings might be driven by inferential processes or by processes that are associative (or, at any rate, non-inferential). Consider the finding by Fazio et al. (1995) that negative implicit attitudes toward black people predicted judgments of the extent to which blacks were responsible for the riots that followed the acquittal of the Los Angeles police who beat Rodney King. This finding might be due either to subjects believing (unconsciously) that blacks are untrustworthy, say, or their associating blacks with negative concepts in a manner that biases information processing by making certain representations more accessible than others. With regard to some of the data, however, something like an inferential story is very plausible.[16] I shall argue, though, that the inferential story doesn't vindicate the claim that implicit attitudes are beliefs. The reason is that the semantic sensitivity of implicit attitudes is too patchy and fragmented to justify attributing a correlative belief to subjects.

There is genuine evidence for sensitivity of implicit attitudes to semantics. De Houwer (2001) found that the characteristics of stimuli that drove behavior in an IAT were sensitive to task instructions. His British subjects performed an IAT using the categories "British" and "foreign", but including both positively and negatively valenced names in each (e.g. both "Einstein" and "Adolph Hitler" in the "foreign" category). Responses tracked "Britishness" versus "foreignness" rather than "positive" versus "negative." That seems to be evidence of implicit attitudes responding to semantic properties. Implicit attitudes are also sensitive to context, which is also sometimes sensitivity to the content of representations. Wittenbrink et al (2001) showed that when black faces were presented within the context of a church interior, implicit attitudes were less negative than when they were presented in the context of an urban environment.

However, we need to interpret evidence like this in the light of a great deal of other evidence with regard to which either no content-driven story is plausible or

with regard to which the content-driven story is of the wrong sort to count as a display of genuine inference. Evidence of the first type is weak evidence against a doxastic account of implicit attitudes. It is evidence against a doxastic interpretation of implicit attitudes because *bona fide* beliefs are supposed to be inferentially promiscuous: interacting with (pretty much) any and every other representation in a manner sensitive to its semantic content. However, honest-to-goodness beliefs may also activate contents associatively, so this kind of evidence is relatively weak. It is only insofar as the non-inferential story is more far-reaching—if implicit attitudes seem to work associatively (or at any rate non-inferentially) when we would expect a belief to (also) work semantically—that difficulty in telling an inferential story provides evidence against the doxastic conception. Evidence of processing that is not content-driven is evidence against a doxastic conception of implicit attitudes only insofar as this kind of processing crowds out inferential processing. Evidence of the second sort is more important because *bona fide* beliefs are not apt to feature in just *any* kind of transition: rather they tend to feature in normatively respectable transitions. The wrong sort of content-driven processing is strong evidence against the doxastic interpretation.

First, then, evidence that implicit attitudes often work non-inferentially. Consider the extensive evidence that these attitudes are involved in subtle and probably nonconscious social behaviors. Dovidio et al. (1997), for instance, found that implicit attitudes toward black people predicted less eye contact and more eye blinking when subjects interacted with a black interviewer. It is hard to tell an inferential story about this behavior (what belief might motivate such behavior?) Similarly, Bessenoff and Sherman (2000) found that implicit attitudes toward overweight people predicted seating distance from them and Wilson et al. 2000 found that implicit attitudes predicted the number of times subjects touched a black confederate's hand; it is hard to see what inferential process might be at work here (again, what belief might motivate these behaviors?). These kinds of findings are ubiquitous in the literature (see, further, Chen & Bargh 1997; McConnell & Leibold 2001), showing that implicit attitudes are pervasively involved in microbehaviors that are difficult to explain by citing belief/desire pairs. All by themselves, these behaviors put pressure on the doxastic conception of implicit attitudes.

To this kind of evidence we can add evidence that implicit attitudes often don't update in the kind of way beliefs are supposed to, or that they update contrary to the way beliefs should. First an example of the latter: 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: they updated when they should not. A second example, this time of failure of appropriate update: Gregg, Seibt and Banaji (2006) gave their subjects information about the members of two novel groups. Members of one did mainly positive things while members of the other 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: the information concerning the error did not interact with their implicit attitudes.[17]

Even when we can demonstrate that implicit attitudes feature in processes driven by their semantic contents, these processes may be of the wrong sort to attribute the correlative belief to the agent. Consider the finding that implicit attitudes not only predict subjects' judgments of the suitability of candidates for a job, but also cause the confabulation of the qualifications needed for the job, with subjects choosing white or male applicants over black or female, and justifying their choice by reference to the qualifications possessed by the favored applicant (Dovidio & Gaertner 2000; Uhlmann & Cohen 2005; Son Hing et al. 2008). This may be evidence of content-driven processing, but not of a kind that would justify attributing to the person an appropriate belief. Beliefs are inferentially promiscuous, recall: any inference from a proposition like "a white (male) candidate is superior" to "the kinds of qualifications possessed by the white (male) candidate are the ones relevant to the job" is an inference—if indeed it can be called that at all—that ignores too many other representations which we can justifiably attribute to the person. While it is a content-driven transition, it is encapsulated from too many of the person's other representational states to count as evidence that the attitude on which it pivots is a belief.

An alternative explanation of these experiments is that they involve genuine inference over explicit attitudes: a preference for the white, or male, candidate biases the evaluation of the qualifications, but the evaluation is explicit (prompted by the experimenters). If this explanation is correct, the relevant processes are inferential, and implicit attitudes play a role in the inferential processing, but the inference is not itself over the implicit attitude. Rather than the semantic content of the implicit attitude playing an inferential role, the attitude plays the role of disposing the person toward some options and away from others. The role played by implicit attitudes, if this hypothesis is correct, is not itself inferential; rather, they play the role in inference ascribed to Pavlovian biases by Crockett (2013), of "pruning" the decision tree. An inferential story must involve inference *over* implicit attitudes, not merely inference biased by such attitudes.[18]

Other data seem best explained by postulating content-driven processing over implicit attitudes themselves, but clearly involve transitions of the wrong sort to count as genuine inference. De Lemus et al. (2013) primed female subjects with pictures of men and women in stereotypical and counter-stereotypical settings (men and women in a kitchen versus men and women in an office setting). Subjects then performed an association task, in which they were first primed with either male or female faces and then required to categorize agency/competence related words versus warmth/community words. Previous exposure to stereotypical images actually reversed the gender stereotypes, causing subjects to respond faster toward agency/competence words when primed with a female face. The more subjects supported affirmative action, the stronger the effect.

This seems to be an example of some kind of influence of personal-level beliefs over implicit attitudes. It is unlikely that these subjects associate "female" more

strongly with agential words than with warmth/community words; the stereotypical associations are too often and too forcefully impressed upon all subjects in countries like Spain (in which the study was conducted). It may be that stereotypes can be weakened by extinction more successfully than Mandelbaum thinks (see Dasgupta 2013 for evidence) but reversal of the stereotype would require counterconditioning, and it is unlikely that the subjects had experienced sufficient counterconditioning to explain the results. Rather, it seems we must explain the results as caused by the content-driven interaction of personal-level beliefs and values with implicit attitudes. But this content-driven interaction doesn't look like inference. The subjects do not believe that women are more agentive than men. They are unlikely to assert that claim, which (other things equal) makes it inappropriate to attribute it to them as an explicit belief. And it is also extremely unlikely that they would behave as if they believed it in a variety of situations in which their implicit attitudes were dominant. Given that prior exposure to stereotypical and counter-stereotypical images was necessary to produce the effect, it seems to be the product of an interaction between implicit and explicit attitudes *none of which have the content suggested by the behavior*. Again, we have evidence that implicit attitudes have structure and interact in ways that are sensitive to semantic content, but we don't have the kind of inference proper to belief.

Indeed, one version of the very experiment that Mandelbaum (2013) cites as his central example to show that implicit attitudes have structured contents, the poison experiment (Rozin, Markwith & Ross 1990), demonstrates the patchy and fragmented nature of their content-driven processing. Rozin et al. did not label the jars "poison" and "safe"; they labeled them "not poison" and "safe". There may have been content-driven processing driving the observed behavior, but that processing was blind to the negation. That result is just one of a number of experiments that seem to demonstrate that nonconscious processes are blind to negation (Wegner 1984; Deutsch, Gawronski & Strack 2006; Hasson & Glucksberg 2006). That's a pretty big chunk of the aptness for inference associated with *bona fide* beliefs to go missing, and all by itself is strong evidence that implicit attitudes don't respond to semantic content sufficiently broadly to qualify as beliefs.

As a final set of evidence for the claim that content driven processing over implicit attitudes isn't broadly inferential, one might think of the standard tests for implicit attitudes themselves. While some of the stimuli used to measure response latencies—say associations between "female" and "warm"—might be explained via content-sensitive processing, for others it would be too much of a stretch. Think of "black" and "cancer" or "white" and "flower": the fact that the first word makes the second more accessible seems to reflect a mere association, not an inference. Sequential priming produces even stronger evidence: even if we attribute to the subjects the unconscious belief "that black people are ugly", it is hard to see how they could infer from that belief to the conclusion that a pictograph is ugly. 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. Eyeblinks, startle responses and use of abstract language, of course, are entirely resistant to any such interpretation.

In the light of all this evidence, I think it is safe to conclude that implicit attitudes are not beliefs. They don't actually look very like or update very like beliefs. Nor, however, are they mere associations. They sometimes feature in content-driven transitions, which indicates that they have propositional structure. Since aliefs are supposed to be associative representations, they're not aliefs either. The evidence seems to indicate that they are *sui generis* states for which we lack any term in our folk psychological vocabulary. I dub them "patchy endorsements". They are endorsements, because they have some propositional structure, which entails that they have satisfaction conditions, so that by tokening them agents are committed to the world being a certain way. But they are patchy: they feature in only some of the kinds of inferences and respond to only some of the kinds of evidence we expect from bona fide beliefs with the same kinds of contents (and they also are sensitive to and respond to representations in ways that beliefs do not).

It would be nice to tell some kind of systematic story about patchy endorsements: concerning which states they are responsive to and in what way. Maybe some kind of story like that can indeed be told, but I don't think we're in a position to tell it just yet: we don't yet know under what conditions they enter into inferential processes or what features of representations they are sensitive to. Quite likely there is no systematic story to tell: maybe each (token or type) implicit attitude has a distinctive propositional structure, and this structure explains which content-driven processes it features in and which it does not. We do know that implicit attitudes are not beliefs or associations, or any other state that folk psychology recognizes, but since they play a significant role in some behavior (and arguably some role in all behavior) we also know that we need to know much more about them.[19]

Conclusion

Implicit attitudes are not beliefs. They do not feature often enough and broadly enough in the kinds of normatively respectable inferential transitions that characterize beliefs. Nor, though, are they *just* associations. They do not activate contents solely associatively: they exhibit *some* of the kind of inference aptness that characterize beliefs. They do so in a patchy and fragmented manner, which indicates they have propositional structure. They are patchy endorsements.

There is good reason to think that our moral concepts, the concepts we use in assessing ourselves, others, and actions from a moral point of view, are closely linked to folk psychology. Think, for instance, of the role of belief in distinguishing between blameless and blameworthy action, and of the distinction between intention, foresight, and what the agent ought to have believed in distinguishing degrees of responsibility. Given that implicit attitudes are patchy endorsements, rather than any state that features in our folk psychology, we have good reason to think that some of our existing moral concepts apply relatively poorly to people who harbor such attitudes and to the actions that they cause. We should hesitate before we

blame, or feel shame, or guilt. Equally, though, given that they do not seem to be *just* associations, there may be room to develop analogues of our existing moral concepts that can apply to agents who harbor them. Right now, neither blame nor excuse (insofar as excuse rests on the claim that they are just associations (Levy 2014)), seem justified.[20]

Given that our moral concepts seem to be closely tied to our folk psychological vocabulary, understanding the extent to which implicit attitudes are, or closely resemble, the states that feature in that vocabulary is pivotal to coming to a justified assessment of ourselves and of others, should we or they be shown to have pernicious implicit attitudes. Should we feel shame, or guilt, and if so how much? Should we blame others for actions caused by implicit attitudes? Answering these questions requires a deeper understanding of the nature of implicit attitudes. It may also require work in moral philosophy, seeking to extend or to revise our existing moral concepts, or to invent new ones.

Notes

[1] I owe the phrase "patchy endorsements" to Susanna Siegel, whose comments on this paper made it, and my thinking, very much clearer. I am also very grateful to two anonymous reviewers for *Noûs* for extremely helpful comments on every aspect of this paper and to Eric Mandelbaum, whose advice helped me to avoid many mistakes.

[2] What makes an attitude implicit? It has widely been assumed that implicit attitudes are unconscious attitudes, but there is evidence that subjects may be aware of the content of (at least) some of their implicit attitudes (Hall & Payne 2010). Implicit attitudes might be better regarded as attitudes that are automatically activated and that cannot be intentionally inhibited, but which the person does not avow; see Fazio and Olson (2003) for discussion.

[3] Just as there are affective and behavioral, as well as doxastic, models of racism (see Faucher & Machery 2009 for these distinctions), however, there could be affective and behavioral models of what makes someone a racist. The discovery that someone harbored implicit beliefs with a kind of content we are disposed to describe as racist would not by itself settle the debate whether the person was properly described as racist. However, at minimum it would seem to bear on how we ought to assess the person morally.

[4] Denying that implicit attitudes are beliefs does not commit me to holding that they are what Gendler (2008) calls *aliefs*. An aliefs is "a mental state with associatively-linked content" (2008: 642); evidence that implicit attitudes have propositional structure is evidence that they are not aliefs.

[5] As a referee for this journal emphasized to me, it is only some kinds of moral assessment that may need to be postponed. Some axiological judgments remain appropriate: many of the implicit attitudes that have been the focus of most attention are clearly *bad*. Further, a great deal of normative theory can be done without making progress on the issues I deal with here, since there is good evidence that these attitudes cause discriminatory behavior. However, certain aretaic judgments, concerning whether individuals who harbor such attitudes should be described as vicious, as well as judgments concerning the application of some thick concepts (like 'racist'), and also judgments concerning agents' accountability for some of the actions that result from their implicit attitudes seem to me to be premature in the light of the claims of this paper.

[6] IATs may measure associations of positive words and concepts with characteristics other than race: gender, sexuality, ethnicity, and so on; IATs may even measure subjects' associations of vegetables or animals with good and bad words and concepts.

[7] Interestingly, performance on this task predicts performance on another: rating a person on a variety of measures, including likability, laziness, and ambition after reading a narrative describing the person's day. Greater bias demonstrated on the weapon identification bias predicts more prejudiced

ratings on the assessment task when the fictional person is black, compared to when they are white (Payne 2005).

[8] Many researchers believe that IAT responses are a partially controlled behavior (Huebner forthcoming). If that's correct, then we would expect to see the attitudes probed by them predicting behavior better in individuals who have lower ability to inhibit impulses. There is some evidence that this is the case: IAT responses to alcohol-related stimuli predict alcohol consumption in subjects low in executive function (Houben & Wiers 2009). If IAT responses predict responses only in subjects who (perhaps temporarily) lack the capacity for executive control, the low predictive power shown by Oswald et al. (2013) is not surprising.

[9] Some psychologists hold that the great majority of white Americans have negative implicit attitudes toward black people (Pearson, Dovidio & Gaertner 2009; Dasgupta 2013), but it may be that measurement error has exaggerated both the divergence between implicit and explicit attitudes and the prevalence of such implicit attitudes. Payne, Burkley and Stokes (2008) controlled for structural fit of explicit and implicit measures and found both that the correlation between implicit and explicit measures was higher and that the degree of anti-black prejudice was lower (among their white college student sample) than other researchers have suggested. They found that a slight majority of their sample exhibited anti-black prejudice, and that there was much more variability in attitudes than previous studies had indicated. How much comfort we should take from this finding is unclear, given that different measures of implicit prejudice do not correlate very well with one another (Fazio & Olson 2003; Bar-Anan & Nosek forthcoming), which may indicate that they tap into different representations; a lower rate of prejudice as measured by sequential priming may therefore coexist with a higher rate of prejudice as measured by, say, the IAT and it may also be true that the great majority of white Americans (and Canadians, Australians, Germans, and so on) exhibit negative implicit attitudes toward minorities on *some* measure.

[10] I thank a referee for this journal for pressing me on this question. The referee suggests that a mark of a *bona fide* belief is that it is not changeable solely by changing reinforcement conditions. I resist the temptation to think that any such condition is genuinely *necessary* for a state to be a belief. A state acquired via changing reinforcement conditions might count as a belief if it then features in a broad enough range of respectable inferences.

[11] An anonymous referee for this journal worries that Mandelbaum attacks a strawman: it is not news to social psychologists that implicit attitudes interact with semantic content. It is certainly fair to say that though social psychologists routinely say that implicit attitudes are associations, they often don't seem thereby to commit to anything very specific. However, they also routinely endorse the claims of those few psychologists who have explicitly argued that implicit attitudes work associatively, opposing associative processing to rule-based processing. This distinction was first clearly made by Smith and DeCoster (2000) and very influentially defended in Strack & Deutsch (2004). Gawronski & Strack, (2004), Gawronski & Bodenhausen (2006) and Rydell & McConnell (2006) all endorse the claim that implicit attitudes interact with other attitudes in a distinct, and distinctively non-rule-based, manner.

[12] A referee for this journal worries that the very fact that we find binding almost everywhere is a problem for Mandelbaum. Typical conditioned associations display this kind of binding, as do all phobias, and so on. The view seems to entail that some kind of inferential processing is ubiquitous, even in many of the cases that seem to be most grist to a behaviorist mill. As I read him, Mandelbaum accepts this entailment: he maintains that entirely non-propositional structures are much rarer than is commonly thought.

[13] As a referee for this journal emphasized to me, however, we need to exercise caution in maintaining that cognitive dissonance usually doesn't alter implicit attitudes. There is relatively little work on the topic, and it is always dangerous to put too much weight on null findings. Moreover, it seems no one has a really good story about why implicit attitudes change when they do in these paradigms. It is important to note, however, that even if I were to concede that cognitive dissonance does (sometimes) involve genuine inference over implicit attitudes, that's hardly a major blow to my view. I already accept that implicit attitudes feature in some inferential processes, after all. I maintain that these inferences are too patchy, fragmented, and disrespectable for implicit attitudes to count as *bona fide* beliefs. Finding that they feature in a few more inferences than I think won't threaten that (of course, finding that they feature in *many* more inferences than I think will threaten that).

[14] Most of the evidence De Houwer (2014) cites in favor of a propositional model of implicit attitudes concerns the formation of implicit attitudes via inferential processes; this evidence is vulnerable to the same charge.

[15] For similar reasons, Ron Mallon's (forthcoming) account of how stereotype threat might involve inferential processes does not offer any support to Mandelbaum: Mallon's story can be told without invoking implicit attitudes at all (in fact, given that Mallon's aim is to offer a personal-level explanation of stereotype threat, he may intend that implicit attitudes play no role; see n. 18 for an alternative explanation of his data that may be less congenial to Mallon). Mallon suggests that being reminded of some stereotypical belief about a category to which one belongs—that women aren't supposed to be good at math, say—might lower motivation to engage in a task because one doesn't expect to do well at it or because others have low expectations of one. In support of this hypothesis, he notes that subjects who endorse the stereotype are more susceptible to stereotype threat (Schmader, Johns & Barquissau 2004). That fact makes an inferential story more likely, but is consistent with inference being over explicit belief.

[16] I say "something like" an inferential story, because there is a genuine question whether the kind of responsiveness to evidence implicit attitudes display counts as inference at all. Since I don't wish to attempt to adjudicate this question, I will describe the processing in a less committal manner, as simply "content-driven processing".

[17] Gregg, Seibt and Banaji did find that implicit attitudes induced propositionally—by abstract propositions—were more responsive to evidence than implicit attitudes induced associatively (though not responsive enough to flip when participants' beliefs reversed). The evidence they present is therefore not evidence of complete insensitivity to evidence. This is hardly surprising: changing a person's beliefs is altering the way in which they think about the object of their beliefs, and all by itself that will constitute pressure to alter implicit attitudes. Why implicit attitudes acquired inferentially should be more sensitive to evidence than those acquired associatively remains unknown: perhaps in the former case some kind of 'channel' between explicit and implicit attitudes remains open for some time after attitude acquisition. I thank an anonymous referee for pushing me on the interpretation of this set of studies.

[18] Mallon's personal-level account of stereotype threat might be cashed out in similar terms; on this account, the inferential story would represent the personal-level tip of a subpersonal iceberg.

[19] Huebner (forthcoming) is more sanguine. On his proposal, implicit attitudes are not a unified phenomenon. Rather, they are the product of three independent though interacting systems: associative Pavlovian systems, associative model-free systems and computationally expensive model-based systems. At first glance, Huebner's proposal would seem neatly to explain the limited aptness for featuring in inference and responsiveness to evidence we have surveyed, with model-based processing explaining what aptness for inference there is, and the other two systems explaining why it is limited. However, while I have little doubt that these three systems interact to explain (at least a great deal of) thought, I see little evidence that model-based systems generate implicit attitudes except by generating explicit beliefs, which then—associatively, perhaps—put pressure on implicit attitude formation. While implicit attitudes may be caused by all three systems, they may only be constituted by or feature in the first two. Response latencies, for instance, might best be explained via a conflict between (avowable) model-based attitudes and associative attitudes. This leaves unexplained the limited aptness for inference implicit attitudes apparently manifest. Huebner provides a mechanism for inhibition of these attitudes—as he points out, top-down processes can dominate even the strongest conflicting responses—but not for inference aptness.

[20] The literature on responsibility for implicit attitudes, and for the actions they cause, is growing rapidly (see Holroyd 2012; Glasgow forthcoming; Washington & Kelly forthcoming; Zheng forthcoming; Brownstein unpublished; Madva unpublished for some of the highlights). Some of this work is less directly vulnerable to the claims made here than others, depending on how much in each paper turns on details of the role implicit attitudes play in cognition or behavior. However, insofar as appeal is made to our intuitions in developing the arguments that people should or should not be held responsible for implicit attitudes and for the actions they cause, caution is needed, because our intuitions may be generated by processes that respond to the agents who feature in them in ways that implicitly attribute (only) folk psychological states to them. As a consequence, our intuitions may be unreliable.

References

Aronson, E., & Carlsmith, J. M. (1962). Performance expectancy as a determinant of actual performance. *Journal of Abnormal and Social Psychology* 65: 178–182.

Bar-Anan, Y., & Nosek, B. A. (Forthcoming). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*. DOI 10.3758/s13428–013–0410–6

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition* 18: 329–353.

Bortolotti, L. (2009). *Delusions and Other Irrational Beliefs*, Oxford: Oxford University Press.

Briñol, P., Petty, R., and McCaslin, M. (2008). Changing attitudes on implicit versus explicit measures: What is the difference? *In Attitudes: Insights from the New Implicit Measures*, R. Petty, R. Fazio, and P. Brinol (Eds.). New York: Psychology Press.

Brownstein, M. (Unpublished) Attributionism and Moral Responsibility for Implicit Bias.

Brownstein, M. & Madva, A. (2012). The Normativity of Automaticity. *Mind & Language* 27: 410–434.

Chen, M. & Bargh, J. A. (1997). Nonconscious behavioural confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology* 33: 541–560.

Cooper J. (2007). *Cognitive Dissonance: Fifty Years of a Classic Theory*, Los Angeles: Sage Publications.

Crockett, M.J. (2013). Models of morality. *Trends in Cognitive Sciences* 17: 363–366.

Currie, G. and Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*, Oxford: Oxford University Press.

Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research* 17: 143–168.

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. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology* 37: 443–451.

De Houwer, J. (in press). Why a propositional single-process model of associative learning deserves to be defended. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual Processes In Social Psychology*. NY: Guilford.

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

De Lemus, S., Spears, R., Bukowski, M., Moya, M. & Lupiáñez, J. (2013). Reversing implicit gender stereotype activation as a function of exposure to traditional gender roles. *Social Psychology* 44: 109–116.

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.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology* 33: 510–540.

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. & Machery, E. (2009). Racism: Against Jorge Garcia's moral and psychological monism. *Philosophy of the Social Sciences* 39: 41–62.

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

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobstrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology* 69: 1013–1027.

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

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology* 40: 535–542.

Gawronski, B., Walther, E., and Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology* 41: 618–26.

Gawronski, B. & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin* 132: 692–731.

Gawronski, B., Bodenhausen, G. V., & Becker, A. P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. *Journal of Experimental Social Psychology* 43: 221–223.

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: 634–663.

Glasgow, J. (Forthcoming). 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.

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.

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.

Hall, Deborah L., Payne, B. Keith. (2010). Unconscious influences of attitudes and challenges to self-control. In Ran Hassin, Kevin Ochsner and Yaacov Trope (eds.) *Self Control in Society, Mind, and Brain*, New York: 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: 1015–1032.

Holroyd, J. (2012). Responsibility for Implicit Bias. *Journal of Social Philosophy* 43: 274–306.

Houben, K., & Wiers, R. W. (2009). Beer makes the heart grow fonder: Single-target implicit preferences for beer determine consumption. *Netherlands Journal of Psychology* 62: 10–21.

Huebner, B. (Forthcoming). Implicit bias, reinforcement learning, and scaffolded moral cognition. In Michael Brownstein and Jennifer Saul (Eds.) *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, Oxford: Oxford University Press

Knowles, E. D., Lowery, B. S., & Schaumberg, R. L. (2010). Racial prejudice predicts opposition to Obama and his health care reform plan. *Journal of Experimental Social Psychology* 46: 420–423.

Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. *Nous* 48: 21–40.

Maass, A., Salvi, D., Acuri, L., & Semin, G. R. (1989). Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality and Social Psychology* 57: 981–993.

Madva, A. (Unpublished). Implicit Bias, Moods, and Moral Responsibility

Mallon, R. (Forthcoming). Stereotype threat and persons. In Michael Brownstein and Jennifer Saul (eds.) *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, Oxford: Oxford University Press

Mandelbaum, E. (2013). Against alief. *Philosophical Studies* 165:197–211.

Mandelbaum, E. (Unpublished). Attitude, inference, association: On the propositional structure of implicit bias.

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.

Newman, G., Diesendruck, G., & Bloom, P. (2011). Celebrity contagion and the value of objects. *The Journal of Consumer Research* 38: 215–228.

Nosek, B., Smyth, F., Hansen, J., Devos, T., Lindner, N., Ratliff, K., Smith, C., Olson, K., Chugh, D., Greenwald, A., and Banaji, M. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18: 36–88.

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. (2005). Conceptualizing control in social cognition: How executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology* 89: 488–503.

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.

Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology* 46: 367–374.

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*, New York, NY: Guilford Press, pp. 1–17.

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.

Phelps, E. A., O'Connor, K. J., Cunningham, et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activity. *Journal of Cognitive Neuroscience* 12: 1–10.

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

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.

Rydell, R. J., & McConnell, A. R. (2010). Consistency and inconsistency in implicit social cognition: The case of implicit and explicit measures of attitudes. In B. Gawronski & B. K. Payne (Eds.), *Handbook of Implicit Social Cognition* (pp. 295–310). New York: Guilford.

Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles* 50: 835–850.

Schwitzgebel, E. (2010). Belief. *Stanford Encyclopedia of Philosophy*.< http://plato.stanford.edu/entries/belief/>

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review* 4: 108–131.

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115: 236–71.

Smith, A. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics* 122: 575–589.

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.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review* 8: 220–247.

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

Washington, N. and Kelly, D. (Forthcoming). 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.

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

Wilson, T., Lindsey, S., & Schooler, T. (2000). A model of dual attitudes. *Psychological Review* 107: 101–26.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology* 81: 815–827.

Zheng, R. (Forthcoming). 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.