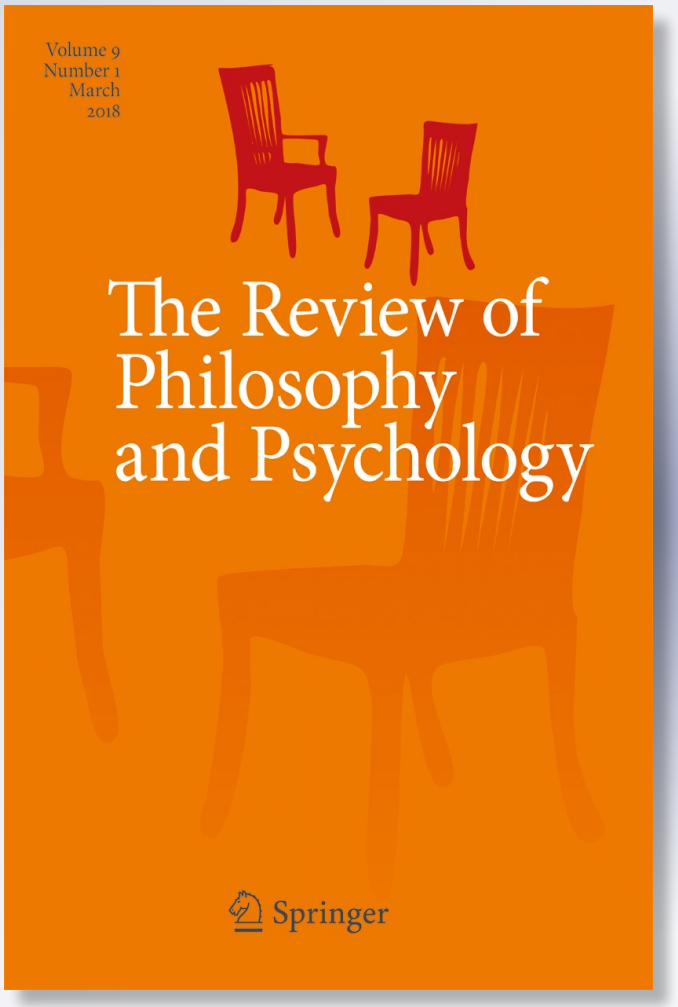
Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures?

Peter Carruthers

Review of Philosophy and Psychology

ISSN 1878-5158
Volume 9
Number 1

Rev.Phil.Psych. (2018) 9:51-72
DOI 10.1007/s13164-017-0354-3



Springer

Rev.Phil.Psych. (2018) 9:51–72
DOI 10.1007/s13164-017-0354-3

Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures?

Peter Carruthers [1]

Published online: 26 June 2017

Abstract Implicit and explicit attitudes manifest themselves as distinct and partly dissociable behavioral dispositions. It is natural to think that these differences reflect differing underlying representations. The present article argues that this may be a mistake. Although non-verbal and verbal measures of attitudes often dissociate (and frequently conflict), this may be because the two types of outcome-measure are differentially impacted by other factors, not because they are tapping into distinct kinds of representation or distinct storage systems. I arrive at this view through closer consideration than is usual of the mechanisms and processes that underlie overt behavior.

1 Introduction and Overview

A good deal of attention has been paid to the discovery of so-called "implicit attitudes." These are attitudes towards groups (often groups of people, including races and genders) that manifest themselves in various sorts of indirect test. The most famous of these is the "Implicit Attitude Test", or IAT, which uses differences in reaction-time as the outcome measure in a good/bad–target/foil categorization task (Cunningham et al. 2001). The results often cause surprise and consternation to the people who turn out to have the measured attitudes. And evidence has been building that implicit attitudes are at least moderately good at predicting real-world behavior, independent of the effects of people's explicit (verbally reported) attitudes (Banaji and Greenwald 2013).[1,2]

[1] For an argument that the real-world effects of implicit attitudes are only minor ones, see Oswald et al. (2013). For a careful and measured reply, see Greenwald et al. (2015).

[2] Throughout I shall understand *verbal* behavior broadly, to include any form of communicative response. Hence communicating one's attitudes towards black people on a "thermometer scale" counts as verbal, in this broad sense. The rationale is that all such behavior is subject to communicative and social norms, and is apt to engage reflective forms of cognition.

✉ Peter Carruthers
pcarruth@umd.edu

[1] Department of Philosophy, University of Maryland, College Park, MD 20742, USA

Philosophers and psychologists, as well as those in the education and business worlds, have struggled to understand what implicit attitudes mean for our conception of ourselves as agents, as well as their practical implications. Without a doubt, the implications are important. Individuals need to know what their implicit attitudes are, and they need to find ways to moderate their unwanted effects. Institutions need to find structures and procedures that will reduce the influence of implicit attitudes. But the *ontological* importance of their discovery may be minimal, I shall argue. Their interest lies rather in what they show us about the different ways in which representations of the same kind can get expressed in behavior, resulting from differences in their interactions with other attitudes (e.g. egalitarian ones).

In much of the social-psychology literature on (explicit) attitude formation and attitude change, an attitude is thought to be a disposition to make a certain sort of evaluative statement (Eagly and Chaiken 1993). And then on a similar dispositional reading of *implicit* attitudes, such an attitude would be a disposition to engage in certain sorts of non-verbal evaluative behavior. On this construal it follows immediately that explicit and implicit attitudes are distinct (even if they had turned out to be perfectly correlated), because they are distinct types of behavioral disposition (to engage in verbal versus non-verbal behavior respectively). However, it doesn't follow that the representations that underlie these dispositions are distinct even if (as the evidence suggests) those dispositions are not well-correlated with one another. In fact, it cannot be ruled out that the underlying representations are identical. This is what I propose to suggest. In what follows, therefore, I shall construe attitudes (whether implicit or explicit) realistically, to be identified with the core representational basis of the overt dispositions in question.

While everyone agrees that the *processes* tapped into by implicit and explicit measures are in some way distinct, most of the psychological literature on implicit attitudes is silent on the question of representation. (Some theories of attitudes make at least tacit commitments about underlying representations, as we will see in a moment.) Indeed, Greenwald and Nosek (2008) argue that the available evidence is neutral between single-representation and dual-representation accounts. They argue, moreover, that one cannot empirically discriminate between these views by looking at patterns of association and dissociation among implicit and explicit measures alone. I will suggest that when a broader range of evidence is considered, a single-representation account should nevertheless be preferred.

Our topic, then, concerns the representational structures that underlie explicit and implicit attitudes respectively. One reason for thinking that these might be distinct would emerge if one thought (as many psychologists do) that explicit attitudes can be formed "on the fly" through propositional reasoning whereas implicit attitudes need to be built up slowly through associative learning. For it is hard to see how a single type of representation could be arrived at in both of these ways, while also having dissociable effects on subsequent explicit and implicit measures. However, Mandelbaum (2016) provides an extended critique of the claim that implicit attitudes are only caused slowly through associative learning. On the contrary, they admit of one-off learning of various sorts.

Mandelbaum describes how implicit-attitude change follows a "logic" that is not merely associative. For instance, having been conditioned to have a negative implicit attitude toward person A, and on being told that person A dislikes person B, people

thereafter have a *positive* implicit attitude toward person B (Gawronski et al. 2005). This can't be explained associatively, and seems rather to suggest an inference of the form, "The enemy of my enemy is my friend." Mandelbaum also notes the findings of Briñol et al. (2009), who show that implicit attitudes can be moderated more effectively by strong arguments than by weak ones (with associative connections controlled for). And perhaps most convincingly, he describes the findings of Gregg et al. (2006), who show not only that implicit attitudes can be induced by a single act of imagining, but that the resulting attitudes are just as strong as those produced by 240 rounds of evaluative conditioning. In what follows, therefore, I shall assume that the ways in which implicit attitudes are formed and changed are not merely association-driven.

Mandelbaum (2016) fails to draw a clear distinction between implicit *affect* and implicit *belief*, however. Indeed, he treats all implicit attitudes as forms of (evaluative) belief. My own account, in contrast, will draw a sharp distinction between affect and belief (while noting that the two often interact). Moreover, Mandelbaum regards implicit and explicit attitudes as distinct *types* of belief, located in separate memory stores. My own account, in contrast, will claim that there is no distinction between implicit and explicit attitudes at the level of representation and storage. Rather, the difference lies entirely in how the attitudes in question interact with others in a given context to influence behavior. And indeed, the evidence suggests that one and the same attitude can be explicit in some conditions (e.g. speeded ones) while remaining implicit in others (where reflection is possible). For correlations between implicit and explicit measures of a given attitude increase under conditions of both cognitive load and speed (Hofmann et al. 2005).

I noted earlier that although most theories focus on the processes behind implicit and explicit responding, some carry tacit commitments about the representations that those processes operate over. For example, the well-known Associative-Propositional Evaluation (APE) model of Gawronski and Bodenhausen (2006) maintains that the processes that underlie implicit-attitude formation, activation, and change are associative ones, whereas the processes involved in explicit-attitude formation and change are propositional, involving judgments of consistency or inconsistency among propositional representations that have been activated in the circumstances. Indeed, like Gendler (2008), the model assumes that explicit attitudes are structured propositional representations whereas implicit attitudes are realized in networks of associative connections. Although this is not the main focus of the APE model, it appears to be a clear commitment of the account.

The APE model will provide a useful foil for clarifying my own proposals. While I agree that the main difference between explicit and implicit processes is that the former are *reflective* in a way that the latter are not (drawing of a wider range of information), I disagree about many of the specifics. I deny that explicit responses are entirely consistency-driven, for example. On the contrary, I will argue in Section 2 that they involve some of the same affective-evaluative processes that underlie implicit responding. I also deny that implicit evaluations are entirely associative and can only be altered in the short term by changing which components of an associative structure become activated. On the contrary, I will suggest in Section 4 that affective responses can result from sophisticated (proposition-like) appraisals, and that judgments can directly alter one's values (and affective responses) in a top–down manner. In addition, the APE model, too, draws no distinction between cognitive and affective implicit

attitudes. Where Mandelbaum (2016) treats all implicit attitudes as structured beliefs, the APE model regards all implicit attitudes as both evaluative and associative in nature. I differ from both in maintaining a clear distinction between affective implicit attitudes and cognitive (propositional) ones.

While not everyone agrees that implicit attitudes are always associatively caused, almost everyone agrees that such attitudes are *unconscious*. (Gawronski and Bodenhausen [2006] are one among a number of exceptions.) The issues here are complex (as are questions to do with consciousness generally). But they are arguably irrelevant to our topic. For even if all explicit attitudes were conscious and all implicit ones were unconscious, this wouldn't settle the question of their representational distinctness. For we know, after all, that one and the same *perceptual* representation can be either conscious or unconscious, depending on whether it attracts top–down attention. Something similar might be true in connection with explicit and implicit attitudes.[3]

What *will* prove important, however, is whether explicit attitudes are directly and reliably expressed in the communicative episodes that are used to measure them. For if they are, then explicit and implicit measures surely couldn't dissociate to the extent that they do unless the latter had a distinct representational basis. To see this, try to suppose that both sorts of task tap into the same set of underlying representations. And consider a case where someone reports explicitly that they value all races equally while displaying racial bias in an implicit task. How could a reaction-time task (the IAT) or an evaluative-priming task cause the same set of underlying (egalitarian) representations to have such different effects? If equality of value can lead to evaluative bias in an implicit test, then all of the difference would somehow have to result from extraneous factors. This might mean that the implicit task isn't a valid measure of anything. Section 2 will argue that the relationship between underlying representations and their expression in speech is complex and indirect, however, leaving plenty of room for a single-representation account to be defensible.

It is worth emphasizing that just because explicit and implicit attitudes differ in their functional roles (for example, in their responsiveness to reasons, or in their accessibility to speech), it doesn't follow that they differ in their representational basis. For there are a couple of ways in which two distinct causal profiles can come about. The first is that there are two states that are differently realized in representational structures that differ in their patterns of psycho-neural connectivity, thereby causing them to interact differently with others. Here the states in question are realized in distinct representations. But another way in which states can have different functional roles is that they *merely* differ in the ways they interact with others, either because one and the same structure can be activated in different *modes* (e.g. consciously, in working memory, versus unconsciously), or because something about the *content* of the state can lead to very different effects when interacting with others.

Within a single subject, I shall suggest, the difference between an implicit and an explicit attitude about the same topic needn't be intrinsic, but can lie rather in distinct interactions with *other* attitudes occasioned by the different measures employed. Someone who is biased against black people may be reluctant to sit next to a black

[3] In fact, I believe that all amodal beliefs (setting aside instances of seeing-as and hearing-as) are unconscious, whereas the valence and arousal components of affective states can be conscious (Carruthers 2011, 2015; see also Mandelbaum 2014). The latter point will play some role in the Section 4.

person on the bus when acting spontaneously, for example. But when asked to report how he feels about blacks, the same evaluative representation will compete with many other values to control his verbal response. Included among these might be a desire to avoid social criticism, which leads him to say, perhaps, "All people matter equally to me." In short: from the fact that different *measures* of attitudes dissociate, it doesn't follow that the underlying representations on that topic are distinct.

Across subjects, too, the difference between an implicit and an explicit attitude with the same content can reside in people's other attitudes. Someone with an implicit bias against black people may show that bias in many subtle ways when acting unreflectively. But when he reflects, he may instead express egalitarian beliefs, or be motivated to correct for his own bias through some form of affirmative action. Someone whose bias against blacks is explicitly held, in contrast, may harbor the very same underlying evaluative representation. But in this case the person has other attitudes that enable him to endorse that bias in his speech and in reflectively-guided forms of action.

It should be obvious, then, that claiming that implicit and explicit attitudes are realized in the same representations doesn't mean that one is incapable of drawing a real distinction between someone who (like a KKK Grand Wizard) is an avowed racist, and someone who harbors only implicit racial attitudes. I suggest that the evaluative structures that underlie racism can be the same in the two cases. But the Grand Wizard *lacks* beliefs about the equal status of all races, while also having beliefs concerning the rights of white people to rule (for instance). These differences make it the case that the Grand Wizard doesn't hesitate to assert the moral and intellectual inferiority of blacks. Someone whose racism is merely implicit, in contrast, has other attitudes that make such assertions seem abhorrent. It goes without saying, of course, that the overall difference between the two is a morally relevant one.

As noted earlier, I think we need to distinguish between *cognitive* attitudes (whether explicit or implicit) and *affective* ones.[4] Indeed, rather different things need to be said about the structures that underlie each, as we will see. The sorts of cognitive attitudes that concern us include stereotypes of racial and other social groups (*Blacks are criminals*; *Women are caring*; *Asians are good at math*; and so on). The relevant affective attitudes would involve responding to members of a specific social group, as such, as good or bad (liking versus disliking the group). There is evidence that stereotypes and affective-evaluative attitudes are distinct mental structures, and that they can dissociate from one another (Amodio and Devine 2006; Gilbert et al. 2012). Someone can have a stereotype (even a negative stereotype) about a social group without having a negative affective attitude toward that group, and vice versa. And this can be true using either explicit or implicit measures.

Many stereotypes have evaluative implications, of course. To the extent that one appraises criminality to be bad, the stereotype of blacks as criminals implies that they are bad, and may give rise to a negative affective response when activated. But not all stereotypes have evaluative import. Perhaps the stereotype, *Girls like reading* is mildly positive for many people; but it can also be neutral or negative, depending on how one evaluates reading. Similarly, *Nurses are female* is likely to be affectively neutral, as is,

[4] Archetypal cognitive attitudes would include memories and beliefs; affective attitudes include emotions like fear and disgust, as well as feelings of desire and repulsion.

Boys like rough-and-tumble play. So there are really three kinds of case to be considered: purely cognitive attitudes (stereotypes); purely affective attitudes toward groups; and mixed cases where affective responses to a group may be caused, in part and in context, by evaluatively-relevant stereotypes.

Sections 3 and 4 will consider cognitive and affective attitudes separately. My goal is to outline positive accounts of each, while explaining how implicit and explicit measures of each type of attitude can issue in dissociable effects while being grounded in a common type of representation (often a numerically identical representation). Section 5 will then respond to a pair of challenges to my account. One is a potential objection to my treatment of affective attitudes in particular, arising from some of the findings of Gregg et al. (2006). The other is a recent attempt by Madva and Brownstein (2017) to undermine the distinction between cognitive and affective implicit attitudes. In both cases careful attention to the underlying mechanisms involved will enable us to defuse these challenges. But first (in Section 2) something needs to be said about the relationship between explicit attitudes in general and their expression in speech.

2 Expressing Attitudes in Speech

The distinctive feature of implicit attitudes is that people don't verbally report having them. Or rather (and more strictly) indirect measures of attitudes generally dissociate significantly from people's verbal reports. As a result, part of what underlies the common belief that implicit attitudes reflect different underlying representations from explicit ones (lacking the kinds of proposition-like structures distinctive of the latter, for example) may be a naïve conception of the relationship between one's attitudes and one's verbal reports. Many philosophers, in particular, tend to think of explicit attitudes as being directly reportable in speech. In that case it would be hard to see how the reported representation could be the same (or even of the same kind) as might cause one's conflicting implicit responses, as we saw earlier. The present section is devoted to critiquing the direct-expression idea. But I propose to be brief. For it is challenged at length (on somewhat different grounds) in each of Carruthers (2011) and Carruthers (2015).

Speech is a form of action, of course, and like any other form of action it can be influenced by multiple goals and values simultaneously (as well as by multiple beliefs). When in conversation with someone, for example, one might be trying *both* to establish an affiliative relationship *and* to find out about the local political situation. Moreover, as soon as one allows that there can be unconscious attitudes that influence behavior outside of one's awareness, then it must follow that the same will be true of speech behavior. Hence reporting that one has an attitude and actually having it are two distinct things, and there will often be a significant degree of mismatch between them, depending on one's other motives that are in play in the reporting process (such as impression-management goals, for instance). I will illustrate and substantiate this claim with reference to the counter-attitudinal essay paradigm used extensively by social psychologists studying so-called "cognitive dissonance."

The basic finding in this literature is that participants induced to write an essay arguing for the opposite of what they believe will thereafter shift their expressed attitudes quite markedly, provided that their freedom of choice in writing the essay is

emphasized. (In contrast, people who write essays in support of what they believe, or who write counter-attitudinal essays in conditions where they feel they have little choice, show no changes thereafter.) For example, college students who are known to be strongly opposed to a rise in tuition costs (as measured in an unrelated questionnaire some weeks previously, perhaps) will say that they are neutral on the issue, or even moderately in favor, after writing an essay under conditions of "free choice" arguing that tuition *should* be raised.

For many years it was believed that writing a counter-attitudinal essay induced a negative feeling (called "dissonance") resulting from the perceived inconsistency between one's underlying attitude and one's freely undertaken behavior (Festinger 1957; Bem 1967). But there is good reason to think that this explanation is incorrect. While the negative emotional component of the account is well established (Elliot and Devine 1994), it turns out that similar shifts in expressed attitude can be caused by *pro*-attitude essay writing, provided that people believe their action is both freely undertaken and likely to prove harmful. This was elegantly demonstrated by Scher and Cooper (1989) who told participants of a newly discovered (but fictitious) "boomerang effect", according to which essays read early in a sequence of messages would tend to have counter-persuasive effects. Hence an essay arguing *against* a rise in tuition would be apt to induce the university committee dealing with the issue to *raise* tuition if that essay was read first or second in the series of essays consulted when considering the question. Under these conditions people who had written pro-attitudinal essays (arguing that tuition should *not* be raised) shifted their expressed attitudes quite markedly having learned that their essay would be read second, just as did those who wrote counter-attitudinal essays (arguing that tuition *should* be raised) who learned that their essay would be read second-to-last.

The best explanation of these and many similar findings is as follows (Carruthers 2011). People who have had their freedom of choice made salient to them appraise their act of essay-writing as having been bad, and this makes them feel bad. When queried later about their attitudes on the topic they consider the behavioral alternatives open to them and select one they appraise as presenting their action as *not bad*, thereby ameliorating their negative affective state. This will often involve saying something other than they believe. Indeed, people will embrace any one of a number of behavioral strategies to rid themselves of negative affect in these experiments, including not only shifting their expressed attitude on the subject matter of the essay, but also denying responsibility for the action or denigrating the importance of the issue. Moreover, they adopt the first such opportunity that is offered to them, and thereafter their responses to the remaining questions are unchanged (Simon et al. 1995; Gosling et al. 2006). As a result, it is implausible that any of their attitudes had really changed in advance of the questions being asked.[5]

What happens, then, when free-choice participants in standard (non-boomerang) counter-attitudinal essay-writing experiments are later questioned about their attitudes

[5] Do people then change their attitudes *after* the questions have been asked? They will hear themselves *as* believing that it would be okay for tuition to be raised, for example; and so they likely accept (and come to believe) that they have such a belief. This is a new belief about their beliefs, rather than a change in their first-order attitudes. But it may still have effects on behavior just as if the latter had altered. And indeed, in one of the few investigations of the persistence of dissonance-induced attitude change, the effects on people's expressed attitudes were still discernable a month after the initial experiment (Sénémeaud and Somat 2009).

is this. The question activates their standing attitude (e.g. that raising tuition would be bad) while also activating the goal of saying what one believes, or saying what is true. This goal on its own would lead them to say, "Strongly opposed." But they also have the goal of making themselves feel better (or perhaps: the goal of presenting their previous action as having been a good one). This second goal on its own would lead them to say, "Strongly in favor" (since in that case their behavior of arguing in support of a rise in tuition would be appraised positively, and not merely neutrally). But in fact participants tend to answer around the mid-point, thereby partly satisfying each goal while fully satisfying neither. Moreover, it is quite unlikely that the selection process operates consciously. Participants surely could not be aware of their attitude that raising tuition would be bad when choosing their response or they would then be aware that their answer was a dishonest one, and this would make them feel worse, not better.

As these findings illustrate, speech production in general (like speech comprehension; Hickok and Poeppel 2007) seems to proceed in parallel (or at least interactively; Nozari et al. 2011), with decisions about *what* to say being taken while one is in the process of saying it (Dennett 1991; Lind et al. 2014). This suggestion is consistent with the data reported by Novick et al. (2010), that patients with damage to Broca's area (leading to a form of production aphasia) also show much wider deficits, especially in their capacity to inhibit prepotent actions. (For example, they perform quite poorly in the Stroop test.) For the expressive difficulties experienced by some of these people emerge most clearly in cases where there are many competing things they could say. For example, when asked to generate verbs associated with a given noun, patients with damage to Broca's area may become paralyzed when prompted with "ball", since there are many related verbs to choose from ("throw", "kick", "pass", "catch", and so on). But they perform as normal when prompted with "scissors", which is associated with just a single action ("cut"). Similarly, healthy people given the same test show increased activity in Broca's area when selecting a verb out of many alternatives, as well as during conflicting-action trials of the Stroop test. At the very least these findings establish that speech production involves competition among expressive *actions*, if not competition between thoughts to be expressed.

The latter claim is supported, however, by findings from patients with Wernicke's aphasia (who often have severe speech-comprehension difficulties), as Langland-Hassan (2015) points out. For although the speech of such patients can be fluent, it is often garbled or completely unintelligible, containing misused words, non-words, and meaningless concatenations of the "Green ideas sleep furiously" variety (LaPointe 2005). Since Wernicke's aphasia is primarily a speech-comprehension deficit, we can infer that speech production normally proceeds in parallel with comprehension, evaluating the semantic contents of a range of potential speech actions while they are being constructed and selected. (See also Matsumoto et al. 2004; Aristei et al. 2011; Pickering and Garrod 2013.)

Thus it isn't the case that there is one set of propositional attitudes that is constitutively linked to expression in speech and another set that is not. Rather, all attitudes compete with one another to guide decision-making and subsequent action, and this is true of speech actions just as much as other forms of action. One difference is just that with speech there will generally be a number of things one could say in response to a request, evaluated in parallel in light of a range of different beliefs and values. Speech actions will therefore depend on a larger set of the person's beliefs and goals than will a

speeded classification of a stimulus as good or bad (for example). Hence one might disavow attitudes in one's speech behavior that one nevertheless has, and which can nevertheless guide unreflective forms of decision making.

Consider a policeman in a routine traffic-stop, for example. If the driver is black he might unthinkingly rest his hand on his gun as a precautionary measure while approaching the vehicle, whereas he doesn't do so if the driver is white.[6] His unreflective actions evince the attitude, *Black men are dangerous* or somesuch. But if asked whether this is true he might respond, "Of course it isn't true that all black men are dangerous; some are, just like some white people." The difference between spontaneous and verbal responses is just that in the latter the influence of the attitude is moderated or outcompeted by others, which have become activated in the context by the question. These might include a belief that racial profiling is wrong or a desire to appear impartial.

It may be helpful to compare the account sketched here with Gawronski and Bodenhausen's (2006) APE model. As noted earlier, the latter accounts for people's responses in explicit tasks in terms of a set of propositional representations that have been evoked into activity in the circumstances. These will generally include a proposition that encodes one's underlying evaluative attitude (such as, *I don't like blacks*), which is the main determinant of one's response in the paired implicit task. (Note that in the implicit task it isn't this proposition that does the work, but rather an associatively-caused affective response to a black face.) However, explicit questioning will generally also evoke into activity propositions that are inconsistent with this one (such as, *Everyone deserves equal respect*). These inconsistencies force respondents in explicit tasks to make a choice, often leading to an outcome inconsistent with their implicit attitudes. So in these circumstances one's implicit affective response is not among the causes of one's assertion.

It should be obvious that a mere appeal to inconsistency is insufficient, however. This is because there is nothing about an inconsistency that prescribes how it should be resolved. Plainly people need to engage in decision-making. And the best account of the latter that we have implicates affect-based prospective evaluation (Damasio 1994; Gilbert and Wilson 2007; Levy and Glimcher 2012; Seligman et al. 2013; Shenhav et al. 2013). Hence Gawronski & Bodenhausen, too, should say that people decide what to say in response to a query by appraising and evaluating a set of possible actions together with their immediate consequences. Since communication is always a social activity, the values in question will involve such things as a concern with reputation-management as well as one's moral values. But as soon as such a picture is adopted, it becomes plain that the so-called "implicit attitude" (an evaluative disposition towards black people, say) is just one among a number of affective attitudes that influence one's explicit response. The representational basis of implicit and explicit responding is (in part) the same. The latter is impacted by *more* causal factors, that is all.

Recall our discussion of the counter-attitudinal-essay paradigm: how one expresses one's attitude toward a rise in tuition will be influenced by two different affective appraisals. Saying, "Strongly opposed" feels right to the extent that it expresses one's underlying belief, yet saying, "Strongly in favor" seems good because it presents one's

[6] Note that *unthinkingly* here means *without conscious consideration of alternatives*, not *unintentionally*; on the contrary, the action results from an unconscious decision.

previous actions in the best light; with the actual outcome being a compromise between the two. Something similar may happen when someone who feels negatively towards black people is asked how he feels about blacks, except that many more affective appraisals (self-presentational, moral, and more) are likely to be in play, pushing his response further away from his true feelings. In such a case it will be one-and-the-same evaluative attitude that causes one's implicit response which is also among the causes of one's explicit response.[7]

I have argued (albeit briefly, relying on points developed more fully elsewhere) that there is unlikely to be a constitutive difference between the attitudes we express in speech and those we don't. Rather, all attitudes can operate unconsciously to influence both speech and other forms of action, with that influence being more or less direct, depending on the context. We now turn to discuss the nature of the representations underlying cognitive and affective attitudes separately.

3 Cognitive Attitudes

Most psychologists think that beliefs and other cognitive attitudes are stored in the form of structured memory representations of some sort.[8] Even those who reject the language-of-thought thesis should agree. Suppose one thinks that individual faces, locations, and other properties are represented as stored positions in multidimensional "state-spaces", for example (Churchland 2012). Nevertheless, one still has to acknowledge that an episodic memory of seeing a familiar individual in a particular spot carrying an unusual object, for instance, would be stored by linking up the region of face-state-space that represents the individual person with the regions of the state-spaces that represent the relevant location and the particular kind of object. Indeed, many accounts of hippocampal function suggest that its role is to bind such distributed representations into a spatial and temporal matrix (Eichenbaum et al. 2012). The result is a compositionally-structured representation.

As noted earlier, the implicit cognitive attitudes that interest us are mostly stereotypes like, *Blacks are criminals* and, *Girls like reading*. But implicit stereotypes can just be regular beliefs, I suggest, caused in the normal way. There has been extensive work done on so-called generic statements like, "Ducks lay eggs", "Mosquitoes bite", "Deer ticks carry Lime disease", and so on, many of which are regarded as true and used as a

[7] Note that this one-representation view provides a simple and elegant explanation for the fact that implicit and explicit attitudes, while partly dissociated, are nevertheless reliably correlated. Dual-representation accounts, in contrast, are forced to appeal to other factors to explain the correlation (such as the existence of common pathways in the acquisition process, for example).

[8] Not *all* beliefs are explicitly stored, of course. Some are dispositions to construct such representations by inference from those that *are* explicitly stored, as in Dennett's (1978) example of the belief one would manifest when asked whether or not zebras in the wild wear overcoats. Never having considered the question previously, one doesn't have a ready-stored answer. But one answers unhesitatingly nevertheless. Moreover, not all stored information should be described as a form of belief. This is because information about the statistical structure of the environment is collected and stored (perhaps associatively) for many different specialized purposes without being available to influence central decision making or verbal report. For example, it enriches the predictive models that are employed in low-level visual processing (Kok et al. 2013). There is no reason to think that stereotypes fall into this category, however. Indeed, there is good reason to think that they do not, since stereotypes are *recognized* by people who have them even if they aren't consciously endorsed.

basis for inductive inference, despite only applying to a minority of the members of the kind (Prasada 2000; Cimpian et al. 2010; Leslie 2014). (Note that only adult female ducks lay eggs, only female mosquitoes bite, and only a small proportion of deer ticks carry Lime disease.) In fact, the formation of generic beliefs seems to be the mind's default mode of generalizing, since "Some …" and "All …" statements tend to be recalled later as generics (Leslie and Gelman 2012). And generics can be based on just a few salient instances, especially where the instances in question are strikingly negative. (Consider: "Sharks attack bathers", "Muslims are terrorists" in the post-9/11 era, and so on.) Note that most if not all stereotypes can be expressed as generics. Consider: "Black men are dangerous", "Blacks are athletic", "Women are caring", "Girls like reading", "Nurses are female", "Women aren't natural leaders", "Asians are good at math", and so on. All are generics. It is therefore simplest to assume that all are stored in the same way as other generic beliefs.

In cases where stereotypes concern objectively measurable properties (such as the proportion of nurses and primary-school teachers who are women, or the proportion of murders committed by black men), many turn out to approximate the real statistics (Jussim 2012). (One notable exception concerns nationality stereotypes.) It seems that people build stereotypes, in part, by tracking those statistics, either directly or by report. Moreover, they acquire these stereotypes quite early in life. Indeed, a single instance of a property combined with a noun-phrase (e.g. "Dogs bark") is often sufficient for children to store the appropriate generic belief (Waxman 2010).

Although stereotypes are just regular beliefs, people often decline to assert them, and will likewise decline to rely on them when reasoning reflectively. For most of us are aware that they can have pernicious effects on individuals belonging to the groups in question. Moreover, there are strong normative expectations that people should be judged as individuals, not as members of the groups to which they belong. People will nevertheless acknowledge that they know *of* the stereotype. Everyone can *recognize* the stereotypes, *Black men are dangerous*, and, *Mothers-in-law are interfering harridans*, even if they decline to endorse them.[9] If such stereotypes are nevertheless stored as regular beliefs, then they are likely to find expression in intuitive, spontaneous, and unreflective behavior. And that is exactly what the evidence shows. While someone might insist, sincerely, that she rejects the stereotype, *Women aren't natural leaders*, her unreflective behavior may demonstrate that she nevertheless believes it (Uhlmann and Cohen 2005). In fact, the only thing that differentiates implicit stereotypes from other generic beliefs is that people have systematic reasons for reflectively rejecting them. But they are beliefs just like any other, all the same. And as Section 2 argued, all relevant beliefs that have become active in the context will complete with one another to influence verbal report.

Consider how this might work. When evaluating people for a leadership position and noticing that one of the candidates is a woman, the stereotype, *Women aren't leaders* may become activated, causing one to infer, *This person isn't a leader*. The result might be that the woman's dossier is placed in the "reject" pile. But now suppose that one is asked explicitly, "Do you think that women aren't leaders?", and is required to indicate one's degree of agreement on a scale of some sort. The generic stereotype

[9] This is the basis of stereotype-based humor. One could not find mother-in-law jokes funny if one was unaware of the stereotype for mothers-in-law.

will support an unqualified affirmative answer, and under speeded conditions or conditions of cognitive load that is what might result. But in normal circumstances other beliefs—likely including, *Some women are good leaders* and perhaps, *Everyone should be judged on their merits*—will also become active, and compete to control the response. This might lead one at least to moderate one's affirmative reply, or to offer some degree of negative answer.

The APE model of Gawronski and Bodenhausen (2006), in contrast, claims that explicit stereotypes are propositional structures whereas implicit stereotypes are associative ones. This is possible, but there is no good reason to believe it. Given what we already know about generic beliefs—that they are formed through either testimony or statistical learning, and can create inductive expectations outside of our awareness—it is much simpler to suppose that the same set of generic beliefs underlies both explicit and implicit responding.

I suggest, then, that implicit cognitive attitudes are just beliefs, no different in kind from one's explicit beliefs. (In this respect my views are like those of Mandelbaum 2016.) The difference is that so-called "implicit stereotypes" are generic beliefs that influence behavior in contexts that are speeded and/or non-verbal. These beliefs remain *merely* implicit when they are out-competed by other attitudes for the control of behavior in communicative contexts. Explicit stereotypes comprise the very same generic beliefs, only in the *absence* of beliefs that might prevent their verbal expression. Moreover (but now in contrast with Mandelbaum 2016), there is no reason to think that these beliefs reside in distinct memory stores. On the contrary, it can be one-and-the-same token belief that gets expressed explicitly in one context (under speeded conditions, say) while remaining implicit in another (influencing an implicit measure, but being outcompeted by others when there is more time to respond to an explicit question).

4 Affective Attitudes

Implicit affective attitudes, too, may just be affective states like any other, I suggest. Standing affective attitudes aren't stored as truth-evaluable propositional structures, in the way that episodic and semantic memories are. Rather, they are stored as structural properties of our affective / valuational mechanisms, especially in subcortical networks that include the basal ganglia (Phelps et al. 2014; Lerner et al. 2015). These structures respond to inputs on the basis of innate and acquired appraisals of relevance, and issue in both *valence* that gets directed at the current object of attention and some degree of *arousal* (including heart-rate, breathing-rate, and so on, but also related behavioral dispositions, facial expressions, and bodily postures). What happens when one acquires a new affective attitude is that a change takes place in the sensitivity of these affective mechanisms to a new class of inputs. The latter acquire the capacity to turn on affective processing by matching and activating a newly-stored evaluative structure, thereby creating an affective response.

One can, of course, have explicit beliefs about one's values or about the outcome one would prefer or think best. These are stored propositional states just like other beliefs, and they can form the basis for a verbal report. Such reports may fail to correlate with one's values as measured in other ways, of course (as revealed by one's spontaneous choices), for they may be confabulated, or adopted for reasons of self-

presentation, or whatever. But this doesn't yet demonstrate a difference in representational format between explicit and implicit affective attitudes. This is because beliefs about values aren't themselves affective states. The contrast here is between cognitive states and affective ones, not between explicit and implicit affective states.

Although beliefs about value are motivationally inert in their own right, they can influence affective processing in a top–down manner (Wager 2005; Ellingsen et al. 2013). This is the basis for placebo and nocebo effects. Believing that one has taken an analgesic will tend to decrease the pain one feels, whereas believing that a pain will be bad is apt to make it worse. Likewise, believing that a wine will taste good is apt to increase one's pleasure in its taste (Plassmann et al. 2008). Although the mechanisms underlying such effects are not yet fully understood, we know (from the papers just cited and many more) that these are not just effects of value-beliefs on behavior. Rather, the activity of subcortical reward-systems is directly modulated. Moreover, value-beliefs can have lasting effects on affective processing, continuing long after the beliefs themselves are likely to have been forgotten (Sharot et al. 2012).[10]

These findings explain, I think, how the acquisition of new value-beliefs can have an immediate effect on one's implicit affective attitudes, as we will see in Section 5. For those beliefs can modulate one's underlying evaluative mechanisms in a placebo-like manner. And in addition, of course, although stereotypes themselves are just generic beliefs, they can become affectively laden by modulating the appraisal-conditions for affective responding. Stereotypes like, *Black men are dangerous* are intrinsically motivationally inert. But they can nevertheless exert a top–down influence on values and evaluative processing. When activated, the properties encoded in an evaluatively-laden stereotype will be received as input by subcortical evaluation systems, altering one's appraisal of the situation, and contributing to an affective response of the appropriate valence.

As already noted, standing affective attitudes aren't propositions, but are, rather, dispositional properties of subcortical evaluative mechanisms to respond to certain input representations with affective output (including valence and arousal). But affective systems can nevertheless respond to propositional inputs. In addition to the top–down placebo-like effects already mentioned, the appraisal processes that issue in affect can be both sophisticated and highly context sensitive. It makes a great deal of difference to how one responds affectively to the sight of an enraged grizzly whether or not the bear is safely contained behind bars in a zoo, for example. If it is, then the sight may merely be exhilarating rather than fear-inducing. Moreover, there will be cases of this sort that cannot be explained in terms of an acquired association between bars-in-a-zoo and safety. For one might respond as described even on one's first visit to a zoo, influenced by one's background belief that zoos (in order to function) must arrange things to render their visitors safe.

Here, in outline then, is how affect is produced. Whenever one confronts an object or situation, or entertains the thought of a potential action, one's evaluative systems set to work processing value-relevant properties of the stimulus. (The "stimulus" here can

[10] Although initially puzzling, in may be that this is one of the main mechanisms of evaluative learning. In highly social creatures such as ourselves it is surely adaptive to have a mechanism that can convert socially-acquired evaluative beliefs into felt values (affective dispositions). This is especially likely given that values tend to be one of the main markers that distinguish in-groups from out-groups, and given that the desires and preferences shared by most members of one's community are likely to be adaptive ones.

either be a product of perception or be internally generated from memory or reflective thought.) The result is some degree of affect directed toward the object in question. The valence component of affect can become an aspect of one's experience, leading one to see the thing as to some degree good or bad, thereby issuing in intrinsic approach or avoidance motivation, and biasing one's decision-making accordingly (Carruthers 2017). But one has no direct access to which properties of the thing resulted in one's affective reaction toward it. Indeed, as numerous studies in social psychology have demonstrated, some component of one's affective reaction may result from previously experienced or concurrently experienced objects or events that are actually irrelevant, without one being aware that this is so (Schwarz and Clore 1983; Forgas 1995; Higgins 1997; Li et al. 2007; Halberstadt et al. 2013).

When one is choosing whether to approach a black person in the street for directions, then, or when considering a black candidate for a job, one may have a valenced reaction that one is aware of. The person in the street strikes one as a bit shifty or ominous (leading one to walk on by) and the job candidate strikes one as not quite right for the position (leading one to put the dossier in the "reject" pile). But one has no introspective access to the properties of the people in question that elicit these reactions. There will, of course, be multiple properties of the person in the street that might potentially contribute to one's negative impression—body language, facial expression, clothing, and so on, as well as features of the context in which the encounter takes place. And likewise there will be many aspects of the person's dossier that might contribute to the feeling that he is not a good fit for the job. Although experimental evidence (for example, results from a racial IAT) can show that these responses are actually caused (at least in part) by race, this is not something to which people themselves have introspective access. One has introspective access to the feelings produced by one's value systems, not to which properties produce those feelings. But this is true of value in general. There is nothing specific to implicit evaluative attitudes here.

If the valence component of an implicit attitude is conscious, then one might be puzzled that people can be so surprised (and dismayed!) by their results in implicit tests like the IAT. But the explanation is straightforward, given the points just made. Although affective valence is a component of people's conscious experience, they have no direct access to which properties of their current or recent environment (or of their own thoughts) caused the valenced response. Nor do they have access to the decision-making processes that underlie the production of a verbal answer. So when someone says when asked (and says sincerely, or as sincerely as any assertion ever is, given that there will always be a multiplicity of factors involved) that she feels just as warmly toward blacks as toward whites, she has little or no access to the process that issues in this statement. (She might be aware of fleeting memory images that occur to her while answering, for example, but without any awareness of their causal relevance.) She thinks she is merely "speaking her mind", and saying what she really feels. But in reality her assertion results from the sorts of competitive processes described in Section 2, in which her negative affective attitude toward black people participates (albeit outcompeted by others in this instance).

The IAT, in contrast, taps much more narrowly into the value that the person's affective system places on blacks. But this, too, is something that she has no access to. For although she can be aware of her affective response when confronted with any particular black person, she has no introspective access to the fact that the property that

produces this response is group-membership defined by blackness of skin. In any particular case the negative response can be attributed to other causes (to the person's rudeness, or accent, or lack of friendliness, or even to aspects of one's background mood). There is no feasible way to learn of one's affective attitudes through introspection. Put differently: one can only learn what it is about a stimulus that produces an affective response by inference from previous experience, or by forming a hypothesis about the most likely cause in the circumstances.

The evidence provided by Hahn et al. (2014) is consistent with these claims, despite initial appearances to the contrary. The authors show that people can predict their own IAT results for racial and other categories with moderate-to-strong levels of accuracy (with an overlap between predicted and actual scores of around .55, even after controlling for explicit-attitude results). And they can do this even if they have no previous experience with the IAT, but know only that it is a form of psychological test that can differentiate between people's implicit and explicit attitudes.

There is no reason to think, however, that these findings reflect *introspective* access to implicit racial attitudes. Participants in these studies were presented with sample photos of the racial categories that would be used in the subsequent IAT, and they were asked to predict what a test of their implicit attitudes would show about how positive or negative they feel toward the races in question. So they were, in effect, told that race was the relevant dimension of evaluation, and they could consult their affective response when looking at the photos to formulate their predictions. They were not *introspecting* that they feel negatively toward black people in general (say), but rather inferring this from their negative response towards a particular black face taken together with their knowledge of the conditions of the experiment.

As Hahn et al. (2014) note, there are a number of reasons why people's IAT predictions should deviate from the explicit thermometer-scale ratings they gave in the same experiment. Most importantly, I would say, explicitly rating how one feels about a racial group is a social act for which one can be held responsible. If one says (or otherwise indicates), "I don't like black people very much" one can be challenged to justify one's feelings, others will judge one's character on the basis of one's feelings, and so on. A raft of different motivations thus come into play as one appraises where best to place one's mark on the thermometer scale. *Predicting* how one might score in a test of one's implicit attitudes is a very different thing, and people would likely have felt they wouldn't be held responsible, nor have their characters judged, on the basis of the results. These findings are thus fully consistent with the account I have outlined: implicit measures of affect tap fairly directly into the relevant evaluative dispositions, whereas explicit measures are subject to multiple affective influences.[11]

As previously, it may be helpful to compare the account outlined here with the APE model of Gawronski and Bodenhausen (2006). According to the latter, implicit affective attitudes are realized in associative structures. I have suggested, in contrast, that affective attitudes in general can comprise sophisticated

[11] Hahn et al. (2014) also found that when people had their explicit attitudes re-tested following completion of the IAT, those attitudes had shifted significantly towards their implicit ones. Given the framework I have outlined this should not be surprising. For one of the motives in play in explicit-attitude reporting (albeit competing with others) is to say what one believes one's attitude to be. Since people had recently received evidence of their racial biases through learning of their IAT results, this would exert a pressure to alter their explicit reports accordingly.

proposition-like appraisal-conditions for subcortical affective mechanisms. Moreover, according to the APE account, the only means of top–down influence on these attitudes is by modulating which aspects of the associative network become active. I don't deny that top–down influences can manipulate the inputs to evaluative mechanisms (thus influencing the output), of course. Indeed, I suggest that this happens when the context activates one group-stereotype rather than another (such as, *Blacks are rhythmical* rather than, *Blacks are criminals*), thus altering one's appraisal of the affective significance of the situation. But I have also suggested that top–down influences can have a direct effect on the values encoded in the underlying affective mechanisms themselves. In addition, it is these same stored values that underlie both implicit and explicit responding (with the latter being subject to a range of additional influences).

According to the APE model, explicit tasks will lead one to encode one's attitudes into truth-evaluable propositional form, such as, *I don't like black people* or, *Black people are bad*. But other, inconsistent, propositions might also be formulated, such as, *Everyone is of equal worth*. What one says aloud is some unspecified outcome of these conflicts. While I agree that an explicit response will often be the outcome of competing pressures, I think these will include evaluative / affective pressures, rather than always being formulated propositionally. Along the lines sketched in Section 2 for speech production, I suggest that someone considering how to indicate his attitude toward blacks on a thermometer scale will be appraising the appropriateness (that is, seeming goodness) of the different possible responses along multiple dimensions, producing an overall intuition of the best way to respond. One of the options will "feel right" in the circumstances (given both evoked beliefs and evoked affect), and will be the one selected. Included among these dimensions will be the affective response to black people that underlies implicit responding. Hence an explicit judgment will incorporate among its causes the same affective appraisal that underlies implicit performance.

5 Two Objections

The present section will consider two objections to the shared-representational-basis account of implicit and explicit attitudes offered here. The first focuses on the account of affective attitudes presented in Section 4. The second challenges the reality of the distinction between cognitive and affective implicit attitudes that I have assumed throughout.

5.1 Asymmetric Change

As noted in Section 1, Gregg et al. (2006) found that implicit and explicit evaluative attitudes could each be induced by both evaluative conditioning and by acts of imagining. In contrast, they found that while the explicit attitudes thus induced could easily be reversed, this was not true of the implicit ones, which persisted in the face of updated information. While this study might be claimed to demonstrate an important difference in the representational basis for

implicit and explicit attitudes, I will argue that, properly understood, it does no such thing.[12]

The first two experiments conducted by Gregg et al. (2006) found that both evaluative conditioning and mere supposition of the properties of two imaginary social groups produced equally significant shifts in both explicit and implicit (IAT-performance) evaluations of those groups. It is no surprise, of course, that being given an explicit narrative about the qualities of the two groups should induce strong explicit evaluations of them. For participants would surely have generated very different evaluatively-laden stereotypes for the two. Nor is it any surprise that evaluative conditioning would give rise to very different implicit evaluations (as measured by the IAT).

Why would evaluative conditioning give rise to differences in explicit attitudes, however? In light of the framework set out in previous sections, this should not be puzzling. For following training, participants would experience positive valence whenever one group was named and negative valence whenever the other was. This would directly influence their choice of descriptor when asked to say what they feel about the two groups. Note, moreover, that in these experiments there would have been none of the social pressures that influence people when evaluating real-life groups—such as races—thereby moderating what they say.[13]

What about the finding that mere supposition can induce novel implicit evaluations, however? How can a mere supposition that the Niffites are just like the Jebbians (who are already positively valued) lead the former to be *implicitly* valued just as strongly (with the equivalent happening in reverse when learning that the Luupites are just like the evil Haasians)? Although remarkable, this finding is fully consistent with the framework outlined in Section 4. It is a testament to the power of top–down influences on affective processing. The mere belief that one group is as good, or as bad, as another can be enough for members of that group thereafter to be intuitively appraised, in swift online tasks, as good or bad. This is consistent with many other findings in the literature. For example, the mere belief that one wine is better than another is enough to ramp up the response of subcortical pleasure networks to its taste (Plassmann et al. 2008).

The real challenge to the framework provided in Section 4 comes from the findings of Experiments 3 and 4 of Gregg et al. (2006). Each initially induced both reported and implicit preferences between two social groups before attempting to undo them again through different supposition-type manipulations. In both experiments the finding was that supposition could reverse the valence of explicit evaluations but not implicit ones.

Participants were first caused to have strong preferences for one group over another, either by reading a graphic narrative, or by undergoing evaluative conditioning, or both. In Experiment 3 (which will be our focus here), one group of participants was then tested on both explicit and implicit measures, before the experimenter pretended that the computer program that had assigned the materials to each individual had made an error, and had got the names of the two social groups the wrong way round. (The other

[12] Hu et al. (2017) report findings comparable to those of Gregg et al. (2006). These findings admit of a similar explanation to the one I advance here for the latter, however.

[13] In addition, it seems likely that participants would *also* have formed stereotype representations of the two groups through the conditioning process. For they were told at the outset that they would be learning about the characteristics of the two social groups, and they were asked to keep clear in their minds which group possessed which characteristics. A name drawn from one group paired with the word "vicious", for example, would then lead participants to believe that members of that group are vicious.

group of participants were a control, and were given some other pretense for re-taking the two tests.) In an attempt to salvage some data from the error, the participants were asked whether they would mind taking the two tests again, but this time imagining that the two groups were reversed. That is, having previously learned that the Niffites were good and the Luupites were bad (say), they were now to believe the reverse, and were to provide both implicit and explicit evaluations of members of the two groups (who had distinctive and easily recognizable names). The finding was that explicit evaluations of the two groups reversed at this second round of testing, whereas implicit (IAT) evaluations did not. It seems that while implicit evaluations can easily be *induced* by mere story-telling or supposition, they cannot likewise be undone.

How is this finding to be explained if there is really no difference in the representations underlying implicit and explicit affective attitudes? That implicit attitudes resist reversal whereas explicit ones do not might seem to suggest that there *is* a representational difference between the two. In fact, however, this finding may be readily explicable. Suppose that appraisals of novel objects can be *created* swiftly through top–down expectations, but that once these appraisal-conditions for sub-cortical evaluative mechanisms have been set, they can only be reversed more slowly and incrementally. This would make sense, for it is comparatively rare in nature for types of object to reverse their values overnight. And indeed, top–down-induced changes in preferences caused by merely imagined choices, as well as the responsiveness of subcortical evaluation systems, can last for three years or more (Sharot et al. 2012). (See also Mandelbaum 2014, for a similar point in relation to belief.) That would explain why the IAT measures failed to reverse when people were asked to suppose that the characters of the two social groups had flipped. But at the same time it is possible that the change in *explicit* measures didn't really reflect any change in *affective* evaluation of the two groups. Rather, participants completed the explicit-evaluation component of the task by using working memory and following task instructions. That is, knowing, and feeling, that the Niffites are good (say), but knowing that they were supposed to be evaluating the two groups in reverse, they simply scored the value of each individual accordingly.

The findings of Gregg et al. (2006) are thus consistent with the claim that the representations underlying our affective attitudes are the same, whether those attitudes are measured implicitly or explicitly. What differs is just that implicit measures tap more directly into the properties of one's affective system, whereas explicit measures can be influenced by many different goals and values.

5.2 Cognitive versus Affective

I now turn to the second of the two challenges. My argument that implicit and explicit attitudes have the same representational basis has assumed a distinction between cognitive attitudes (in this context, stereotypes) and affective ones (in this context, often referred to as "prejudices"). Madva and Brownstein (2017)) argue against a dissociation between cognitive and affective attitudes, however. They claim that all implicit attitudes are clusters of semantic-affective associations, thus defending a position quite close to that of the APE model of Gawronski and Bodenhausen (2006). This means that they inherit the disadvantages of the latter. More importantly, however, the ontological status of these postulated clusters is opaque. How does a semantic representation "associate" with an affective state, or vice versa? One semantic

representation can associate with another, since both are stored representations that can become linked in such a way that activating the one will activate the other. But how does a concept associate with a feeling? Affective states are the outputs of evaluative processes, not stored representations of any kind. My own account, in contrast, can tell a principled story to accommodate the same data, guided by what is known about the distinction between semantic and affective networks in the brain.

Recall how I noted in Section 1 that cognitive and affective attitudes may richly interact. In particular, stereotypes that are evoked in a context may provide appraisals for evaluative mechanisms that modulate the latter's affective response. This is why people primed with concepts related to jazz, for example, may show lower implicit prejudice against blacks in the IAT than those who have been primed with concepts related to gun violence. The concepts evoked in the two conditions may lead one to appraise a given black face as musical in the one case and dangerous in the other, with consequent effects on affective processing. Likewise, people primed with negative affect may display greater influence of negative stereotypes (whether these are measured implicitly or explicitly). While I agree that cognitive and affective networks richly interact, this is consistent with an ontological distinction between them. Wrapping up these points about interaction into the claim that there is a single type of semantic-affective disposition it neither helpful nor illuminating.

Semantic networks are largely distinct from affective ones, and are realized for the most part in regions of temporal, parietal, and prefrontal cortexes.[14] They have been designed to store action-guiding information (albeit especially information that is relevant to one's values). Affective networks, in contrast, are realized in the basal ganglia together with ventromedial prefrontal cortex, among other regions. They have been designed to appraise the evaluative relevance of stimuli, producing some degree of valence and arousal, and motivating action accordingly (albeit relying on semantic-cognitive appraisals of those stimuli). Both sets of networks employ computations over structured representations, but one set is designed to represent (and predict) the environment, whereas the other set is designed to compute the subjective value of what is represented. Moreover, although Madva and Brownstein (2017)) note, correctly, that cognition and affect interact at all levels of the mind, this does nothing to challenge the distinction between cognitive and affective attitudes.

By varying which semantic representations are activated, and which cognitive stereotypes are evoked, one can influence the inputs to affective processes. This will impact both implicit and explicit affective responses. And conversely, by priming positive or negative affect (or more specific affective states like fear) one can influence cognitive processes in a variety of ways, again at both implicit and explicit levels. We can at least begin to understand how these interactions take place. But if implicit attitudes are characterized indiscriminately, as clusters of semantic-affective associations, then we have little hope of understanding how they work, in my view.

[14] Indeed, Gilbert et al. (2012) show using fMRI and multivariate pattern analysis that the information whether black or white faces are seen can be decoded from orbitofrontal cortex when participants are making likely-friend judgments, but can be decoded from anterior medial prefrontal cortex when participants are making athleticism judgments. The former region is a classic part of the affective network, whereas the latter is often implicated in social-cognitive and stereotype judgments.

6 Conclusion

I have argued that it is the same types of underlying representation that give rise to both explicit (communicative) and implicit (non-communicative) behavior. Because the behavior-types dissociate, this can create the impression that there are two kinds of representation in play: explicit and implicit. But this may be an illusion, resulting from our lack of familiarity with the different causal processes involved in each case. Once these are detailed it becomes plausible that the underlying representations don't really differ in type. Indeed, in many cases it can be the very same token representation that is involved in the causation of both kinds of behavior under different conditions.

Note that there are both commonalities and differences in the story of how these behavioral dissociations arise across cognitive and affective attitudes. What is common is that verbal tasks of both sorts always tap into causal factors that are absent in implicit tasks. This is because verbal tasks are always to some degree social and/or reflective in character, and a range of different attitudes can be evoked and compete for control of one's verbal response. What differs is that cognitive and affective attitudes differ in kind. Cognitive attitudes are stored propositional structures. Affective attitudes, in contrast, are stored appraisal-conditions for one's value-processing mechanisms.

In the end, however, the causal structures underlying both explicit and implicit attitudes are the same, I suggest. (Of course nothing has been demonstrated conclusively here. All my arguments have been inferences to the best explanation of the data, and are defeasible.) What differs is just that there are factors that influence explicit (verbal) responses that don't influence implicit (non-verbal) ones, and vice versa. Hence the discovery of implicit attitudes, although of great practical and methodological importance, adds nothing to our mental ontology. The theoretical interest of their discovery lies rather in what they reveal about the different ways in which representations of the very same type can be manifested in thought and behavior.

Acknowledgements I am grateful to Luca Barlassina, Greg Currie, Dan Kelly, Dan Moller, Jeremy Pober, Brent Strickland, and anonymous reviewers for comments on earlier drafts of this paper.

References

Amodio, D., and P. Devine. 2006. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91: 652–661.

Aristei, S., A. Melinger, and R.A. Rahman. 2011. Electrophysiological chronometry of semantic context effects in language production. *Journal of Cognitive Neuroscience* 23: 1567–1586.

Banaji, M., and A. Greenwald. 2013. *Blindspot: Hidden biases of good people*. New York: Delcorte Press.

Bem, D. 1967. Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74: 183–200.

Briñol, P., R. Petty, and M. McCaslin. 2009. Changing attitudes on implicit versus explicit measures: What is the difference? In *Attitudes: Insights from the new implicit measures*, ed. R. Petty, R. Fazio, and P. Briñol. New York: Psychology Press.

Carruthers, P. (2011). The Opacity of Mind. Oxford: Oxford University Press.

Carruthers, P. 2015. *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: Oxford University Press.

Carruthers, P. (2017). Valence and value. Philosophy and Phenomenological Research. doi: 10.1111/phpr.12395.

Churchland, P.M. 2012. *Plato's camera: How the physical brain captures a landscape of abstract universals*. Cambridge: MIT Press.

Cimpian, A., A. Brandone, and S. Gelman. 2010. Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science* 34: 1452–1482.

Cunningham, W., K. Preacher, and M. Banaji. 2001. Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science* 12: 163–170.

Damasio, A. 1994. *Descartes' Error*. London: Papermac.

Dennett, D. 1978. *Brainstorms*. Brighton, Sussex: Harvester Press.

Dennett, D. 1991. *Consciousness explained*. London: Penguin Press.

Eagly, A., and S. Chaiken. 1993. *The psychology of attitudes*. Boston: Wadsworth Publishing.

Eichenbaum, H., M. Sauvage, N. Fortin, R. Komorowski, and P. Lipton. 2012. Towards a functional organization of episodic memory in the medial temporal lobe. *Neuroscience and Biobehavioral Reviews* 36: 1597–1608.

Ellingsen, D.-M., J. Wessberg, M. Eikemo, J. Liljencrantz, T. Endestad, H. Olausson, and S. Leknes. 2013. Placebo improves pleasure and pain through opposite modulation of sensory processing. *Proceedings of the National Academy of Sciences* 110: 17993–17998.

Elliot, A., and P. Devine. 1994. On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* 67: 382–394.

Festinger, L. 1957. *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Forgas, J. 1995. Mood and judgment. *Psychological Bulletin* 117: 39–66.

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

Gawronski, B., E. Walther, and H. Blank. 2005. Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology* 41: 618–626.

Gendler, T. 2008. Alief in action (and reaction). *Mind & Language* 23: 552–585.

Gilbert, D., and T. Wilson. 2007. Prospection: Experiencing the future. *Science* 317: 1351–1354.

Gilbert, S., J. Swencionis, and D. Amodio. 2012. Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia* 50: 3600–3611.

Gosling, P., M. Denizeau, and D. Oberlé. 2006. Denial of responsibility: A new mode of dissonance reduction. *Journal of Personality and Social Psychology* 90: 722–733.

Greenwald, A., and B. Nosek. 2008. Attitudinal dissociation: What does it mean? In *Attitudes: Insights from the new implicit measures*, ed. R. Petty, R. Fazio, and P. Briñol. Hillsdale: Erlbaum.

Greenwald, A., M. Banaji, and B. Nosek. 2015. Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553–561.

Gregg, A., B. Seibt, and M. Banaji. 2006. Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology* 90: 1–20.

Hahn, A., C. Judd, H. Hirsh, and I. Blair. 2014. Awareness of implicit attitudes. *Journal of Experimental Psychology: General* 143: 1369–1392.

Halberstadt, J., D. Pecher, R. Zeelenberg, L. Wai, and P. Winkielman. 2013. Two faces of attractiveness: Making beauty in averageness appear and reverse. *Psychological Science* 24: 2343–2346.

Hickok, G., and D. Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience* 8: 393–402.

Higgins, E. 1997. Beyond pleasure and pain. *American Psychologist* 52: 1280–1300.

Hofmann, W., B. Gawronski, T. Gschwendner, H. Le, and M. Schmitt. 2005. A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin* 31: 1369–1385.

Hu, X., B. Gawronski, and R. Balas. 2017. Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychology and Personality Science*. doi:10.1177/1948550617691094.

Jussim, L. 2012. *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. Oxford: Oxford University Press.

Kok, P., G. Brouwer, M. van Gerven, and F. de Lange. 2013. Prior expectations bias sensory representations in visual cortex. *Journal of Neuroscience* 33: 16275–16284.

Langland-Hassan, P. 2015. Hearing a voice as one's own: Two views of inner-speech self-monitoring deficits in schizophrenia. *Review of Philosophy and Psychology*. doi:10.1007/s13164-015-0250-7.

LaPointe, L. 2005. *Aphasia and related neurogenic language disorders*. 3rd ed. New York: Theime Medical Publishers.

Lerner, J., Y. Li, P. Valdesolo, and K. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology* 66: 799–823.

Leslie, S.J. 2014. Carving up the social world with generics. *Oxford Studies in Experimental Philosophy* 1: 208–232.

Leslie, S.J., and S. Gelman. 2012. Quantified statements are recalled as generics. *Cognitive Psychology* 64: 186–214.

Levy, D., and P. Glimcher. 2012. The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology* 22: 1027–1038.

Li, W., I. Moallem, K. Paller, and J. Gottfried. 2007. Subliminal smells can guide social preferences. *Psychological Science* 18: 1044–1049.

Lind, A., L. Hall, B. Breidegard, C. Balkenius, and P. Johansson. 2014. Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science* 25: 1198–1205.

Madva, A., and M. Brownstein. 2017. Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs*. doi:10.1111/nous.12182.

Mandelbaum, E. 2014. Thinking is believing. *Inquiry: An interdisciplinary journal of philosophy* 57: 55–96.

Mandelbaum, E. 2016. Attitude, inference, association: On the propositional structure of implicit bias. *Noûs* 50: 629–658.

Matsumoto, R., D. Nair, E. LaPresto, I. Jajm, W. Bingaman, H. Shibasaki, and H. Lüders. 2004. Functional connectivity in the human language system: A cortico-cortical evoked potential study. *Brain* 127: 2316–2330.

Novick, J., J. Trueswell, and S. Thompson-Schill. 2010. Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass* 4: 906–924.

Nozari, N., G. Dell, and M. Schwartz. 2011. Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology* 63: 1–33.

Oswald, F., G. Mitchell, H. Blanton, J. Jaccard, and P. Tetlock. 2013. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105: 171–192.

Phelps, E., K. Lempert, and P. Sokol-Hessner. 2014. Emotion and decision making: Multiple modulatory neural circuits. *Annual Review of Neuroscience* 37: 263–287.

Pickering, M., and S. Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36: 329–347.

Plassmann, H., J. O'Doherty, B. Shiv, and A. Rangel. 2008. Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences* 105: 1050–1054.

Prasada, S. 2000. Acquiring generic knowledge. *Trends in Cognitive Science* 4: 66–72.

Scher, S., and J. Cooper. 1989. Motivational basis of dissonance: The singular role of behavioral consequences. *Journal of Personality and Social Psychology* 56: 899–906.

Schwarz, N., and G. Clore. 1983. Mood, misattribution, and judgments of well-being: Informative affective states. *Journal of Personality and Social Psychology* 45: 513–523.

Seligman, M., P. Railton, R. Baumeister, and C. Sripada. 2013. Navigating into the future or driven by the past. *Perspectives on Psychological Science* 8: 119–141.

Sénémeaud, C., and A. Somat. 2009. Dissonance arousal and persistence in attitude change. *Swiss Journal of Psychology* 68: 25–31.

Sharot, T., S. Fleming, X. Yu, R. Koster, and R. Dolan. 2012. Is choice-induced preference change long lasting? *Psychological Science* 23: 1123–1129.

Shenhav, A., M. Botvinick, and J. Cohen. 2013. The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron* 79: 217–240.

Simon, L., J. Greenberg, and J. Brehm. 1995. Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology* 68: 247–260.

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

Wager, T. 2005. The neural bases of placebo effects in pain. *Current Directions in Psychological Science* 14: 175–179.

Waxman, S. 2010. Names will never hurt me? Naming and the development of racial and gender categories in preschool-aged children. *European Journal of Social Psychology* 40: 593–610.