Implicit Attitudes and Awareness

Synthese 197(3): 1291-1312.

Jacob Berger
Department of English and Philosophy, Idaho State University

(Please quote the published version.)

Abstract
I offer here a new hypothesis about the nature of implicit attitudes. Psychologists and philosophers alike often distinguish implicit from explicit attitudes by maintaining that we are aware of the latter, but not aware of the former. Recent experimental evidence, however, seems to challenge this account. It would seem, for example, that participants are frequently quite adept at predicting their own performances on measures of implicit attitudes. I propose here that most theorists in this area have nonetheless overlooked a commonsense distinction regarding how we can be aware of attitudes, a difference that fundamentally distinguishes implicit and explicit attitudes. Along the way, I discuss the implications that this distinction may hold for future debates about and experimental investigations into the nature of implicit attitudes.

Keywords: implicit attitudes; awareness; consciousness; belief

1. Introduction

There is much evidence that people harbor implicit attitudes (for discussion within the psychological and philosophical literatures, see respectively, e.g., Gawronski & Payne 2010; Brownstein & Saul 2016). So-called implicit-association tests ("IATs") reveal that many white participants who explicitly claim not to have racist beliefs about black people are nonetheless slower to associate positive terms with black faces than with white faces (e.g., Greenwald et al 1998). Similarly, white participants primed with black faces are more apt to classify harmless objects as guns than if they are primed with white faces (e.g., Payne 2001). Such findings are typically explained by positing that these white individuals have implicit attitudes regarding black people that influence their behavior.

Despite the near consensus that there are implicit attitudes,[1] there remains little unanimity regarding how they differ from their explicit counterparts. Perhaps the most widely accepted

[1] I acknowledge that there is growing controversy about the existence of implicit attitudes (for popular summary, see, e.g., Goldhill 2017). Recent meta-analyses, for example, seem to indicate that the predictive validity of measures of implicit bias and stereotyping behavior is rather small (for an overview, see, e.g., Greenwald et al 2015). But implicit

account of the difference—a difference often simply assumed by both psychologists and philosophers alike—is that we are (or readily can be made) *aware* of our explicit attitudes, which are *conscious*, but not aware of our unconscious implicit attitudes (e.g., Greenwald et al 1998; Bosson et al 2000; Saul 2013; Holroyd 2015). Such an account would seem to explain many findings regarding implicit attitudes: for example, the experimental evidence that measures of implicit bias, such as IATs, are typically weakly correlated with standard measures of explicit attitudes, such as verbal report (e.g., Hofmann et al 2005; Nosek 2007). In other words, participants readily articulate their explicit attitudes, but rarely if ever report—and often even deny—their implicit ones. A natural explanation of these results is that individuals are aware of the former but not aware of the latter.

This explanation of the difference between implicit and explicit attitudes has, however, recently been challenged (e.g., Hofmann et al 2005; Gawronski et al 2006; Hahn et al 2014; Cooley et al 2015). It has been found, for example, that the correlation between measures of implicit and explicit attitudes is inversely related to the strength of participants' motivation to avoid appearing biased (e.g., Nier 2005). Perhaps more strikingly, Hahn, Judd, Hirsh, and Blair (2014) revealed that participants are quite adept at predicting the results of their IATs. Such research has led some to conclude that people are, despite appearances, aware of their implicit attitudes. As Cooley, Payne, Loersch, and Lei (2015) summarize:

> Our results also have implications for theoretical debates about the phenomenology of implicit attitudes. Many researchers have used the term implicit as synonymous with "unconscious" when referring to attitudes, leading to the assumption that people are unaware of the content of their implicit attitudes. However, careful studies have recently shown that people can be quite accurate at reporting the content of their implicitly measured attitudes… [which] suggests that people have introspective access to their implicit attitudes (p. 114).

---

measures such as IATs do seem to reveal *some* variety of psychological kind (but, see, e.g., Machery 2016); I thus proceed on the assumption that there are implicit attitudes, whether or not they are comparatively weak attitudes.

Indeed, such evidence might suggest that the expression 'implicit attitudes' does not pick out a genuine psychological category distinct from explicit attitudes; one might conclude that such attitudes are simply a class of beliefs intrinsically the same as their explicit counterparts, except insofar as people are reluctant to report on them—perhaps because these beliefs are socially unacceptable or because they contradict the individual's more considered explicit beliefs (cf. Goldhill 2017).

I believe that such conclusions would be too hasty. My goal in this paper is to propose a new hypothesis regarding the difference between implicit and explicit attitudes. I argue that, though we can be and often are aware of our implicit attitudes, we are not aware of them *in the same way* as we are aware of our explicit ones. While some theorists have distinguished various ways that we can be aware of attitudes (e.g., Gawronski et al 2006; Holroyd 2015), most researchers in this literature have overlooked another commonsense distinction between how we can be aware of mental states, a difference that arguably distinguishes implicit and explicit attitudes.

I begin in section 2 by surveying the current debate over the nature of implicit attitudes, issuing some caveats about the approach pursued here. Then, in section 3, I describe this distinction between modes of awareness of attitudes, proposing what I call the 'Awareness View' of implicit attitudes. I offer several arguments for the Awareness View in section 4. Section 5 addresses how this account is consistent with the evidence that we can be and often are aware of our implicit attitudes. Although my primary goal here is to explicate and motivate the Awareness View, I discuss in section 6 some implications of this view for the debate over the *intrinsic nature* of implicit attitudes—that is, the debate regarding whether or not implicit attitudes are simply ordinary beliefs or belief-like states (e.g., Schwitzgebel 2010; Egan 2011; De Houwer 2014; Mandelbaum 2016), or *sui-generis* states quite unlike beliefs or other ordinary propositional attitudes recognized by folk psychology (e.g., Gendler 2008; Gawronski & Bodenhausen 2011; Levy 2015; Madva 2016).

2. The current debate over implicit attitudes

Although many theorists maintain that at least one pertinent difference between explicit and implicit attitudes concerns our awareness of them, it is also often assumed that another difference, if not the central difference, is that explicit but not implicit attitudes are *conscious*. As Cooley and colleagues' (2015, p. 115) comments in the Introduction reveal, 'implicit' and 'unconscious' are often used interchangeably (see also, e.g., Hahn et al 2014, p. 1370). As their remarks also indicate, however, many theorists assume that any kind of awareness of an attitude renders it conscious; 'awareness' and 'consciousness' are often used interchangeably as well. Based on the evidence that we can be aware of our implicit attitudes, some theorists might therefore conclude that both implicit and explicit attitudes can be and often are conscious (e.g., Gawronski et al 2006). Other theorists, by contrast, propose that *all* attitudes are unconscious (e.g., King & Carruthers 2012). On both of these approaches to consciousness, it would seem that consciousness cannot mark the difference between explicit and implicit attitudes. Both can be conscious or neither can be.

Thus arises a first caveat about the approach that I pursue here: because the nature of consciousness remains vexed (see, e.g., Block 2009), I stay neutral regarding whether or not implicit attitudes can be or ever are conscious. But whether or not implicit attitudes are conscious, there is an independent question of whether and in what ways we are aware of such attitudes. There is, after all, a distinction between a *state's being conscious* and a creature's *being conscious or aware of a state*—what Rosenthal (2005, p. 4) calls the difference between a state's being 'state conscious' and one's being 'transitively conscious' of a state. To preserve this distinction, I use 'being aware' and related expressions to refer to the latter condition.

Recent debate regarding implicit attitudes has largely focused not on our awareness (or lack of awareness) of implicit attitudes—an arguably *extrinsic* feature of those states—but rather on the

intrinsic nature of implicit attitudes. This focus may have arisen in the face of evidence that a difference in awareness seemingly cannot distinguish implicit from explicit attitudes. But this focus may also reflect the fact that there appears to be much indication that implicit and explicit attitudes differ in their psychological or functional roles. Experimental evidence seems to suggest, for example, that implicit biases operate in an *automatic* manner, unlike their *controlled* explicit counterparts (e.g., Payne 2001; Ranganath et al 2008).

Many thus advocate versions of what I call 'Intrinsic Views', on which implicit attitudes function unlike beliefs because they are intrinsically unlike them (e.g., Gendler 2008; Gawronski & Bodenhausen 2011; Madva 2016). A common model in psychology, for example, conceives of explicit attitudes as propositionally structured conceptual states and implicit attitudes as nonpropositional (socially conditioned) *associational states* (for an overview, see, e.g., Gawronski & Bodenhausen 2011). On this associational version of the Intrinsic View, whereas an explicit belief might have the propositional content *that black men are dangerous*, an implicit attitude may only associate the concepts BLACK MAN and DANGEROUS. An associational variety of Intrinsic View thus explains the purported functional differences between implicit and explicit attitudes—such as the fact that explicit attitudes may function as premises in reasoning, whereas implicit attitudes seemingly cannot—because of this difference in propositional structure (see, e.g., Mandelbaum 2016, pp. 636-637). In a similar vein, Levy (2014a; 2015; 2017) has recently proposed that implicit attitudes are what he calls 'patchy endorsements'; while he grants that implicit attitudes have propositional content, Levy proposes that such attitudes differ functionally from explicit attitudes (e.g., they putatively cannot be controlled) because their contents are not sufficiently integrated with the rest of one's other person-level attitudes in the way that beliefs' contents can be.

Other theorists instead endorse what I call 'Doxastic Views', on which implicit attitudes are simply ordinary thoughts—most likely weak, recalcitrant, or unendorsed beliefs or belief-like states

that may contradict one's more considered explicit beliefs (e.g., Schwitzgebel 2010; Egan 2011; De Houwer 2014; Mandelbaum 2016).[2] On the Doxastic View, the purported differences in the psychological roles of explicit and implicit attitudes are illusory. As Mandelbaum (2016, p. 648) argues, for example, even if it is the case that implicit attitudes cannot be controlled, many of our ordinary explicit beliefs are formed quickly and cannot be controlled—such as beliefs formed on the basis of perception (see, e.g., Quilty-Dunn 2015). Likewise, Mandelbaum cites evidence that explicit suppositions both inculcate and modulate implicit attitudes (Gregg et al 2006; cited on 2016, p. 644)—facts that would be hard to explain if implicit attitudes did not have at least some propositional structure.

I am myself drawn to a Doxastic View on which there are in fact no intrinsic differences between implicit and explicit attitudes. That is, on my view, the only (functional) differences between explicit and implicit attitudes are to be explained by our different modes of awareness of those attitudes. But since the literature on this topic is vast, I cannot address all of the psychological evidence in favor of Intrinsic Views. I therefore offer the following second caveat: I remain neutral regarding the intrinsic nature of implicit attitudes. But I will argue in section 6 that the considerations in favor of the Awareness View undercut some of the arguments for Intrinsic Views. I now turn to presenting the Awareness View.

3. The Awareness View introduced

It is crucial first to clarify what it is to be *aware of* an attitude, as several theorists distinguish different uses of 'awareness'. Gawronski, Hofmann, and Wilbur (2006, p. 486), call the awareness of an attitude itself 'content awareness,' which they distinguish both from *source awareness*—the awareness

[2] It is not unreasonable to think that we can and often do have contradictory attitudes—even contradictory explicit beliefs (e.g., Lewis 1982; Egan 2008; for a review of some relevant experimental evidence, see Mandelbaum 2014, p. 79, fn. 58).

of the cause or origin of an attitude—and from *impact awareness*—the awareness of the influence the attitude has on other states or behavior. Most theorists who assume that we lack awareness of our own implicit explicit attitudes seem to have in mind that we lack *content* awareness of them. What the evidence mentioned above suggests, however, is that we often if not always are content aware of both our explicit and implicit attitudes.

The other sorts of awareness do not seem to be good candidates for distinguishing implicit from explicit attitudes either. After reviewing evidence that we can be content aware of implicit attitudes, Gawronski and colleagues (2006, p. 496) propose that the difference between implicit and explicit attitudes consists in impact awareness, citing evidence that people are often unaware of the lamentable consequences of their implicit biases. But impact awareness cannot mark the difference: as Holroyd (2012, p. 294) observes, we are often unaware of the impact of many of our explicit attitudes. I do not know the full range of psychological and behavioral effects of my explicit belief that 2+2=4. Moreover, it even seems that we often *do* have awareness of the effects of our implicit attitudes. Holroyd (2012, p. 292) discusses a study by Montheith, Voils, and Asburne-Nardo (2001), which found that that over half of participants while taking IATs were able to detect a discrepancy between their responses and how they thought they should have responded.

We likewise often have *source* awareness of our implicit attitudes. After taking an IAT, a subject might infer that he or she has implicit biases—thereby becoming content aware of them—and surmise that they must be a product of one's culture. Moreover, we often lack source awareness of our explicit attitudes. We have many explicit beliefs, such as the belief that 2+2=4, without being able to pinpoint the origin of these beliefs.

In a different context, however, Rosenthal has distinguished between *two commonsense kinds of content awareness* (e.g., 2005, p. 185 fn. 24).[3] It is this distinction, I propose, that plausibly distinguishes explicit and implicit attitudes. Going forward uses of 'awareness' and related expressions refer to content awareness, unless otherwise specified.

3.1. Two kinds of content awareness

Consider the way in which I can be aware of my explicit attitudes, such as my explicit belief that my father lives on Main Street. Suppose I try to recall where my father lives. My attitude simply seems to pop into awareness, as it were. While I may also have source or impact awareness regarding that belief, I need not. Even if I had no idea why I believe that my father lives on Main Street, or what that belief causes me to do, I would still be aware of myself as believing it. That is to say, I am aware of the belief in a way that does not seem to be the result of inference or observation. It is natural to say that my awareness *seems direct* or is, as Rosenthal puts it, *subjectively unmediated*:

> **Subjectively unmediated awareness ("SU-awareness")**: One is SU-aware of an attitude A just in case one is aware of A in a way that does not seem to be the result of inference and/or observation.

Rosenthal's notion of awareness here is grounded in folk psychology, but a long tradition in the history of philosophy has assumed that we are often if not always aware of our mental states in this kind of unmediated way (e.g., Descartes 1988; Strawson 1994). Many assume that each time that we think, we are or can be aware that we think and that we need not infer or observe anything about ourselves to come by this awareness.

[3] Rosenthal offers this distinction in the context of defending his *higher-order theory* of consciousness, on which a mental state is conscious just in case one is aware of being in it in a *suitable* way. And as we shall see, in regards to the hypothesis proposed here, a higher-order theorist might consider explicit but not implicit attitudes as conscious; indeed, I am drawn to this view. But since my goal here is not to become embroiled in debates over consciousness, I will not defend this position. Any theory of consciousness must be consistent with Rosenthal's distinction in modes of awareness.

But as Rosenthal (e.g., 2005, pp. 183-184) observes, not all awareness of mental states is like that. If I am aware of another person's attitudes or if I become aware of being in a state on the basis of testimony, my awareness of those states seems to be due to inference and/or observation. Suppose that after talking with my psychiatrist, I come to realize that I have the distorted belief that my father is responsible for all of my problems—an *Oedipal* distortion. I must have had this belief prior to talking to my psychiatrist because the psychiatrist hypothesizes that this belief explains my past behavior. But I was not aware of having that belief—and I only come to be aware of it on the basis of my psychiatrist's testimony. Crucially, I come to be aware of this belief in a different way than the way that I come to be aware of my belief that my father lives on Main Street. While I am unsurprised to learn of the latter belief, I may initially balk at the insinuation of the former though I may acquiesce after self-reflection.

In other words, in the case of my distorted belief, I become aware of it in a manner that *seems indirect*—that is, my awareness is what Rosenthal calls 'subjectively mediated':

> **Subjectively mediated awareness ("SM-awareness")**: One is SM-aware of an attitude A just in case one is aware of A in a way that seems to be the result of inference and/or observation.[4]

In this case, I become content aware of the distorted belief by making inferences (of which I am aware) on the basis of my impact awareness—that is, by reflecting upon my psychiatrist's report and my past behavior. Though I may trust my psychiatrist—and so even form the disposition to become aware of my problematic belief about my father in relevant situations—my awareness remains subjectively mediated.

I become aware of this mental state in a way akin to the manner in which I might become aware of one of my nonmental bodily states. While I may become aware of *cirrhosis* of my liver—for

[4] SU- and SM- modes of awareness are not mutually exclusive. One might be aware of an attitude in both ways if, for example, I have SU-awareness of the attitude but also receive testimony that I have it.

example, on the basis of observing my jaundiced skin or the testimony from my hepatologist—there is no straightforward sense in which it seems to me that I am or can be SU-aware of this state. It is not that I do not believe that my liver is in that state; it is simply that it seems to me that I am aware of that state only on the basis of inference or observation. Though many philosophers have assumed that we always have direct access to all of our mental states, this assumption is questionable. Assuming there are invariably unconscious mental states, we can be only SM-aware of them (for arguments to this effect, see, e.g., Rosenthal 2005; Mandelbaum 2014).[5]

3.2. The Awareness View

My proposal, the Awareness View, is that this commonsense distinction between modes of awareness explains the way that explicit and implicit attitudes differ:

> **Awareness View**: An attitude A is explicit only if one is (or disposed to be) SU-aware of A; an attitude A is implicit only if one is not aware of A or only SM-aware of A.[6]

It makes sense that we can be and often are SU-aware of our explicit attitudes. Whatever else may be true of explicit attitudes, one of their distinguishing marks is that we can spontaneously report them. This is why the gold-standard measure of explicit attitudes is verbal report. Indeed, explicit attitudes arguably are constitutively linked to (spontaneous) verbal report, insofar as they are often

[5] What explains the difference in these kinds of awareness? I do not defend a specific proposal here, but Rosenthal (e.g., 2005, chapter 7) offers the sensible hypothesis that one's awareness of an attitude consists in having a suitable occurrent higher-order thought ("HOT") about that state. To be aware of my belief that *p* is to have the occurrent HOT that I believe that *p*. The difference in modes of awareness thus consists in whether or not one is aware of mental processes that cause the relevant HOTs about one's states. If I am SU-aware of my belief that *p*, then I have the HOT that I believe that *p* and I need not be aware of inferences or observations that may have caused that thought. By contrast, if I am SM-aware of my belief, I have the HOT that I have that thought and I am aware of inferences or observations that caused me to have that thought. But other explanations may be available.

[6] I focus here only on cases wherein attitudes are *occurrent*. But we often talk of attitudes in their dispositional forms: one might believe that 2+2=4, even if one is not currently having that thought, insofar as one is disposed to have the occurrent thought that 2+2=4. Thus one need not be aware of that attitude at all. But we may regard such beliefs as explicit insofar as one is disposed to be SU-aware of it. Some theorists seem to assume that some types of attitude, such as beliefs, can only occur unconsciously or implicitly (e.g., Schwitzgebel 2010; Mandelbaum 2014). But one arguably can have the very same occurrent attitude both implicitly and explicitly. Since some do use 'belief' only in the dispositional way, one might replace instances of 'belief' throughout with 'occurrent assertoric propositional thought'.

operationalized as the attitudes that can be expressed via self-report measures (e.g., Hahn et al 2014, pp. 1369-1370). Perhaps the most reasonable explanation of the fact that we can readily report our explicit attitudes is that we are or are at least disposed to be SU-aware of them. As I argue shortly, we are not even disposed to be SU-aware of our implicit attitudes.[7]

One might object that experimental evidence seems to show that we are never directly aware of *any* attitudes. Consider King and Carruthers' (2012) view, according to which individuals do not have special introspective access to their attitudes; rather, they are aware of their own attitudes via the same mindreading process that is dedicated to determining other people's attitudes. Such a view is in part motivated by the kind of evidence canonically provided by Nisbett and Wilson (1977), who described participants reporting having selected as desirable a particular consumer good on account of its intrinsic qualities, when the item was in fact selected for its arbitrary location within an array of goods. These findings seem to suggest that people lack direct access to the nature of their preferences.

In the case of SU-awareness, however, I am not claiming that we are *in fact* aware of any attitudes in unmediated ways. I am quite moved by the considerations in favor of King and Carruthers' account on which we at least often arrive at SU-awareness of our own mental states by processes of inference of which we are unaware. What is crucial for SU-awareness is that one is not aware of any inferential processes that may lead one to be aware of an attitude: it does not *appear* to one that one is aware of a state via a mediating process. The distinction between SU- and SM-awareness concerns our *mental appearances*—our impressions of how our mental lives seem to us

[7] Does it follow that we should regard *Oedipal* beliefs of which we are only SM-aware as implicit? It need not follow, as my account only holds that it is a necessary condition on an attitude's being implicit that one not be SU-aware of it. It may be that we use 'implicit attitude' to refer to socially biasing attitudes, rather than any attitude of which are not disposed to be SU-aware. That said, I see nothing wrong with calling an *Oedipal* belief 'implicit'. What about bodily states, such as my liver's state of cirrhosis? Plainly we do not regard such states as implicit because they are not even mental states, let alone attitudes.

from the first-person perspective. Thus Nisbett-and-Wilson-type evidence is no trouble for Awareness View. The view is neutral with regard to whether or not we ever have direct access to any attitudes, but it claims that we can and often do *seem* to have such direct access. That it often seems that way is illustrated by the fact that so many in the history of philosophy have assumed that we *in fact* have direct access to all of our mental states. The present account in that way splits the difference between these extremes: it grants to the traditional view that it often seems to us that we have direct awareness of our mental lives while simultaneously yielding to the skeptic that we rarely or even never have such direct access.[8]

The distinction between SU- and SM-awareness thus also cuts across Holroyd's (2015, p. 514) taxonomy of *observational*, *inferential*, and *introspective* modes of (content) awareness. While SM-awareness is characterized by inference and/or observation, SU-awareness may involve inference and/or observation too. Moreover, SU-awareness of states need not be introspective. As some so-called 'higher-order theorists' such as Rosenthal (e.g., 2005) himself have urged, SU-awareness is arguably a feature built into ordinary nonintrospective consciousness.[9]

In the next section I argue that there are good reasons to accept the Awareness View.

---

[8] This compromise may still seem implausible. It might seem that participants in the Nisbett-and-Wilson experiments do not at any time prefer the consumer goods for their intrinsic qualities. But such a view is puzzling: which of the participants' attitudes regarding the goods at the time of verbally reporting their reasons for their judgments are then explicit? I favor the explanation that participants have had contradictory preferences: at the time of selecting the good, they implicitly preferred it for its location; and at the time of verbal report, they explicitly preferred the good for its intrinsic qualities. That is, the latter preferences drive their verbal reports and the former drive their selection behaviors. The participants are, of course, incorrect about which attitude drove their selection behavior. Moreover, it is plausible that at the time of verbal report such an implicit attitude was extinguished when the explicit attitude was formed. But what reason would we have for thinking the confabulated attitudes reported are explicit, if not for the fact that participants are SU-aware of and can thus spontaneously report them? I address the possibility that the attitudes that one fails to report are actually explicit but quickly forgotten or rejected in section 4.2.

[9] The present account has some affinities with the view recently proposed by Levy (2017, p. 535; cf. 2014a, p. 30), which holds that a crucial feature of implicit attitudes is that we can be only inferentially, and not introspectively, aware of them. Moreover, Levy similarly seems to hold that we can be introspectively aware of explicit attitudes, while acknowledging that we may not in fact have introspective access to any attitudes. However, the present view differs from Levy's suggestive remarks insofar as he does not draw a distinction between one's genuinely direct awareness of an attitude and SU-awareness; nor does he hold that the characteristic feature of explicit attitudes is that we are SU-aware of them. Rather, he endorses an Intrinsic View, which locates the central difference between explicit and implicit attitudes in their roles in inference and other psychological processes.

4. Arguments for the Awareness View

All of the arguments that follow share the same abductive form. I argue that if one rejects the Awareness View and holds that we are aware of explicit and implicit attitudes in the same way, it would seem that we must attribute to people considerable insincerity or artifice. Most people who exhibit implicit biases are quite reluctant to report them and often deny them outright. The Awareness View plainly explains this phenomenon: people are simply not suitably aware of their implicit attitudes—and so do not report them. But if we maintain that people are SU-aware of these biased attitudes, then it seems that these individuals are at best unforthcoming and at worst dishonest.

Yet it is a core practice in cognitive science (not to mention daily life) that we take people at their word, unless there is excellent reason to do otherwise. In the meantime, it is reasonable to assume, as many do (e.g., Sullivan-Bissett 2015, pp. 552-553), that people who exhibit racial biases on measures such as IATs may nonetheless report honestly that they do not have racist beliefs.

One might, however, think that people's awareness of their implicit attitudes does not require us to attribute insincerity to them because we can appeal to an Intrinsic View. One might argue, for example, that the fact that implicit attitudes are mere conceptual associations—and not full-blown beliefs—entails that the white participants who exhibit bias on IATs are not insincere when they report that they do not believe that black men are dangerous. The problem for such view is that if people are aware of such conceptual associations in the same way as they are aware of their explicit attitudes, then we should at least expect them to sincerely report *some kind* of (perhaps ambivalent) bias—and they do not do even this, except under rather controlled experimental conditions.

Levy offers a different explanation of why people would not readily report their implicit attitudes—even if they are (SU-)aware of them—which need not involve attributing to them any deception. Levy observes that:

> subjects may identify themselves with their non-prejudiced commitments, but fear that their implicit attitude will be discovered by experimenters who will think worse of them for denying it. When subjects are given the opportunity to express both their commitments and their implicit attitudes, they give divergent responses, bolstering this hypothesis. Ranganath, Smith & Nosek (2008) asked subjects to rate their "gut reactions" and their "actual feelings" toward gay people. People reported "gut reactions" that were more negative than their "actual feelings"; moreover, their gut reactions correlated well with their implicit attitudes. This suggests that people know the content of their gut reactions but refuse to identify their real attitudes with these reactions (2014a, p. 29).

That is, because people's implicit attitudes contradict their more considered explicit ones, they are not being insincere when they report that they are not biased—and have good reason not to report their implicit attitudes, which would only complicate matters.

Levy's explanation is a possibility. But note that even this explanation involves at least holding that these individuals are, in most ordinary circumstances, not being completely forthright.[10] What I propose is that, by adopting the Awareness View, we can more reasonably explain why people do not readily report their implicit attitudes, without countenancing even this weak kind of pretense.[11] I do not believe that the existing evidence demands that we do otherwise—and I return

[10] I do not deny that individuals may insincerely report that they do not have biased attitudes of which they are SU-aware. Because such attitudes are socially unacceptable, we should expect that some might lie about such attitudes. But I do suspect that there are cases of honest people who, despite biased performances on implicit measures, genuinely do not believe that they have such attitudes at all.

[11] Levy (e.g., 2014a; 2014b; 2017) offers what might seem like another reason to deny that people are in any way deceptive when they fail to report their implicit attitudes—namely, that since on his Intrinsic View implicit attitudes are not sufficiently integrated into our mental lives, one cannot be *morally responsible* for actions driven by them. But even if one cannot be blamed for failing to report one's implicit attitudes, if it were the case that one is SU-aware of them, then it would seem at least that one is not being totally forthcoming—and again would be reasonably expected to report such states. So even if an Intrinsic View were true, we are left with the unsavory conclusion that we must attribute to people a kind of (perhaps blameless) caginess. In any case, the idea that we cannot be morally responsible for actions driven by implicit attitudes is questionable (see, e.g., Brownstein forthcoming; but see Levy 2017, p. 547, fn. 11). Moreover, as I

to this issue, addressing *inter alia* Ranganath and colleagues' (2008) evidence, in section 5. I now turn to the more specific arguments for the Awareness View.

4.1. The Argument from Surprise

Consider first the fact that many people are often surprised and dismayed to find that they harbor implicit biases. Here, for example, is how Banaji, a pioneer in the use of the IAT, describes how it felt for her to take the test for the first time:

> So when I took the test… it was stunning for me to discover that my hands were literally frozen when I had to associate black with good… the first thought that I had was: 'Something's wrong with this test.' Three seconds later, it sunk in that this test was telling me something so important that it would require a re-evaluation of my mind, not of the test (Johnson 2013).

Banaji's remarks suggest that she was initially unaware of having implicit biases; only as she took the test did she come to recognize that she might have them. And Banaji's reaction is not uncommon. Recall Montheith and colleagues' (2001) study, which found that that many participants taking IATs were able to detect a discrepancy between their responses and how they thought they should have responded. The experimenters also found that a significant number of participants who were able to detect these discrepancies were stunned by their performances.[12] As Montheith and colleagues report, "on several occasions, our participants expressed surprise and genuine concern about their biases during debriefing sessions" (2001, p. 413).[13]

---

argue shortly, the Awareness View undercuts many reasons to endorse Intrinsic Views such as Levy's. Much light would be shed on the issue of our moral responsibility vis-à-vis implicit attitudes if the Awareness View proves true and implicit attitudes are mere beliefs of a kind. But that is a topic for future exploration.

[12] Such results need not come as a shock or disappointment to everyone—some people may, of course, have biased explicit attitudes that are not in tension with their implicit attitudes.

[13] As far as I can tell, there is not much direct experimental evidence that people are often surprised to find that they harbor implicit attitudes, which is why I draw here only on Banaji's anecdotal remarks and comments from Montheith's debriefing sessions. This is plainly an issue that could be explored by future experimental work.

The fact that people who take IATs are often surprised by their results is difficult to square with the idea that people are always (SU-)aware of their implicit attitudes. If we were (disposed to be) SU-aware of our implicit attitudes, it seems unavoidable that we must attribute a kind of insincerity or artifice to people who report such surprise.

The Awareness View provides the best explanation of these results: these individuals were at first not even disposed to be SU-aware of their implicit attitudes. Taking the IAT, however, rendered them impact aware of their attitudes, which enabled them to infer that they harbor such biases, thereby becoming SM-aware of them. And this SM-awareness proved shocking and disappointing.

Moreover, the fact that Banaji became SM-aware of her implicit biases explains why she did not at first experience them as her own. It is plausible that one *ipso facto* experiences one's explicit attitudes as one's own. When one is SU-aware of an attitude, one need not be aware of any data that could explain why one is aware of the attitude—and thus the attitude seems in an immediate way to be an aspect of one's mental life. If one becomes SM-aware of an attitude, by contrast, then one may doubt the adequacy of the explicit inference or observation that lead to such awareness and thereby remain suspicious that one really has the attitude. Just because my hepatologist informs me that my liver is in a state of cirrhosis, I need not believe her—and I may remain doubtful that my liver is in that state, even if it in fact is. This explains why many are often reluctant to report such states, even if they come to be SM-aware of them. Of course, just as I may trust my hepatologist and so believe that my liver is in a state of cirrhosis, I may regard an implicit attitude as my own too because I come to endorse the relevant inferences. But because I am aware that my awareness of that attitude is based on inference, there remains room for me to doubt it.

4.2. The Argument from Confabulation

Similarly, consider the well-established result that people can and often do confabulate the reasons for actions or judgments driven by implicit attitudes (for an overview, see, e.g., Sullivan-Bissett 2015). Uhlmann and Cohen (2005), for example, demonstrated that male participants' implicit biases against women not only often predict that they will select male job candidates over female ones, but also that they will confabulate the qualifications for the job so that they are consistent with the selected candidate.

Again, assuming that participants in such situations report honestly, the best explanation is that they were neither SU-aware nor disposed to be SU-aware of the attitudes that drove their behavior. Indeed, participants were not aware that they were selecting candidates for biased reasons: they self-rated their levels of objectivity and, as Uhlmann and Cohen observe, "perceiving one's judgments as objective and free of bias predicted greater gender bias. Participants were, apparently, under an illusion of objectivity" (2005, p. 477). Though participants can be made SM-aware of their implicit attitudes by reflecting on their behaviors (of which they explicitly disapprove), the best explanation of the fact that they regard their judgments as objective is they are made only SM-aware of their implicit attitudes via such impact awareness.

One would not expect such confabulation if people were aware of their implicit and explicit attitudes in the same way, even if an Intrinsic View were true. One would expect that, even if participants did not select candidates because of their *beliefs* about their qualifications, participants would report honestly about the non-belief attitudes that drove their selections. That is to say, if the Awareness View were false, we would again have to posit a kind of widespread artifice or irrationality on the part of participants to explain such confabulatory explanations.

There is, however, experimental evidence that people can and do confabulate attitudes even when the confabulated attitudes are explicit. In a classic study, for example, Bem and McConnell (1970) found that people can be led to forget certain explicit attitudes and generate new conflicting

explicit attitudes, without any impression that their attitudes had changed. One might thus think that the fact that people often confabulate the explanations for their biased actions is no evidence that the attitudes that drove them are implicit.[14]

But confabulation of explicit attitudes seems to be comparatively rare—and to occur only in controlled experimental conditions—whereas confabulation regarding implicit biases seems to be the norm. Moreover, the experimental situations in which people confabulate explicit and implicit attitudes are not suitably parallel. In their study, Bem and McConnell induced changes in explicit attitudes by requiring participants to write passages defending positions that conflicted with their reported pre-manipulation attitudes. This manipulation is the only factor in the study that explains why participants changed their explicit attitudes and why they mistakenly report that their pre- and post-manipulation attitudes are the same. But in studies of implicit bias, there seems to be no such analogous explanation. Participants rarely report their biased attitudes at any time and they undergo no learning or other process that would explain why they would change, and then forget, those attitudes in order to report (mistakenly) that their behaviors were unbiased. It is not simply that people seem to have been aware of, but forgotten, their biased attitudes; it is that people seem never to have been aware of them in the first place.[15]

To be clear, I am not arguing that since people sometimes lack impact awareness regarding their implicit attitudes, they can have only SM-content-awareness of them. I agree with Holroyd that we often lack impact awareness regarding our explicit attitudes too. Nor am I claiming that the fact that we can only have SM-content-awareness of implicit attitudes entails that we always or even

[14] I thank an anonymous reviewer for this journal for this interesting objection.

[15] Indeed, in a partial replication of Bem and McConnell's work, Chris and Woodyard (1973) found that participants who rated their pre-manipulation attitudes as subjectively important were less likely to forget them, even if the manipulation did result in a change in those attitudes. That is to say, participants were more likely to confabulate their explicit attitudes if these attitudes were unimportant to them. On the reasonable assumption that one's attitudes about social (in)equality would be rather important, we would thus expect that people would be unlikely to forget their biased attitudes, even if such attitudes did change during the experimental conditions. But, again, this is not what we find.

often lack of impact awareness of their effects. But even if we often are impact aware of our implicit attitudes, it is far from obvious that we are ever SU-aware of the attitudes themselves. The best explanation of the confusion regarding what caused an action is lack of awareness of its (actual) mental cause. In cases where one is not aware of the effects of an explicit attitude, we would expect one to be surprised or confused about the action, but not about what attitude(s) caused it. But this is not what we find in the case of implicit attitudes: people are often not only unaware of actions driven by their implicit biases, but also confabulate what mental states caused those actions once they are made aware of them.

4.3. The Argument from Dissonance

Here is a final reason to doubt that we are ever SU-aware of implicit attitudes. Though there is much debate regarding the intrinsic nature of implicit attitudes, it is reasonable to think that, whatever their nature, their contents often somehow contradict the contents of our explicit attitudes. If a Doxastic Account is correct, then these are simple cases of ordinary contradictory beliefs (e.g., Lewis 1982; Egan 2008; Mandelbaum 2014). But even if an Intrinsic View were true and implicit attitudes' contents literally could not contradict the contents of our explicit unbiased attitudes, implicit attitudes' biased contents would at least in some sense *in tension* with those unbiased explicit contents. If we were SU-aware of both our explicit and implicit attitudes, however, we would expect that our minds would do something to resolve these tensions. Minds abhor contradictions—whether they be between explicit beliefs or beliefs and other sorts of attitudes. This insight is the basis of *dissonance theory*, which holds that the mind often attempts to resolve conflicts between attitudes by jettisoning or reframing certain attitudes (for classic discussion, see Festinger 1957).

Prior to becoming SM-aware of implicit attitudes, however, people do not typically exhibit dissonance with regard to the tension between them and their explicit attitudes. This is evidenced by

the facts that many do not usually attempt to compensate for the behavioral effects of their biases until they learn of how their implicit attitudes contradict their explicit ones. Again, one could attribute a kind of insincerity to people—holding that that they do exhibit dissonance that they fail to report or even deny—but such a proposal is avoidable.

The Awareness View explains why we *genuinely* lack motivation to reduce the tension between our implicit and explicit attitudes.[16] Though dissonance of some sort is likely to arise in cases wherein one both explicitly believes *that p* and explicitly believes *that not p*, it is not hard to see how one might have an explicit belief and a contradictory implicit attitude, without resultant dissonance, if what makes an attitude implicit is that one is not SU-aware of it. That we are not SU-aware of implicit attitudes explains how they might outright contradict our explicit ones without dissonance. As Freud (1949) famously hypothesized, it may be that early in life we are SU-aware of all of our attitudes, but some attitudes later come into conflict with other preferred explicit attitudes, such that that we later repress some attitudes out of SU-awareness. In the case of implicit biases, this explanation seems to be borne out by recent experimental evidence. Dunham, Baron, and Banaji (2008) review data suggesting that children as young as six years of age are often willing to report their social biases, but such willingness typically decreases with age. A Freudian reading of these results holds that we push these attitudes out of SU-awareness to avoid experiencing the conflict and contradictions that they pose.[17]

One might object that there are many examples of contradictory explicit attitudes that do not generate dissonance. One might point to cases of known illusions wherein, for example, a stick half-submerged in water looks bent, even though one knows that it is not. Some, such as Quilty-

[16] This line of argumentation was suggested to me by Jake Quilty-Dunn.

[17] One might object that such evidence demonstrates only that people as they mature in age are unlikely to report their social biases, not that they have ceased to be SU-aware of them. But, again, such an explanation thus faces the objection that people are not being sincere or at least forthright—and so a more natural explanation of the fact they do not report such attitudes is that they are not (SU-)aware of them.

Dunn (2015), have argued that such cases involve contradictory beliefs. In the stick example, it seems that the stick is bent insofar as one has the perceptual thought—that is, a thought formed quickly or automatically on the basis of visual sensation—that the stick is bent, though one more reflectively believes that it is not bent. In such cases, one may be SU-aware of both attitudes, but one does not seem to exhibit dissonance. However, as Quilty-Dunn also urges, it is arguable that such beliefs are separated by cognitive-architectural boundaries—that perceptual thoughts are architecturally separated from central beliefs in a way that prevents the latter from altering the former. And this may also explain why their contradictions do not issue in dissonance. Importantly, Quilty-Dunn offers independent considerations for this view in the case of perceptual thoughts—after all, they are closely tied to perception. But even if an Intrinsic View were correct, there are no independent reasons to think that implicit attitudes would not be a kind of central cognition.

5. Can we be SU-aware of implicit attitudes?

What about the elephant in the room: the growing body of research that suggests that we are often if not always aware of our implicit attitudes—for example, Hahn and colleagues' (2014) remarkable evidence that people are quite good at predicting the results of their IATs? Similarly, we recall Ranganath and colleagues' (2008) evidence that participants are likely to report biases consistent with their implicit attitudes when asked to report their "gut feelings" rather than their "actual feelings" about socially marginalized groups. Although most theorists do not draw the central distinction between SM- and SU-awareness sketched here, one might think to think that such evidence reveals that can be and often are *SU-aware* of our implicit attitudes.

The Awareness View is compatible with these findings. First, this sort of experimental evidence does not demonstrate conclusively that we are always aware of them, nor that we are necessarily (content) aware of them in the same way as we are aware of our explicit attitudes. As

Sullivan-Bissett observes while discussing Gawronski and colleagues' (2006) work, "though [such] evidence suggests that source and content awareness of implicit attitudes is *possible*, that is not to say that subjects always have such awareness in ordinary settings" (2015, p. 549, emphasis hers). That people can be compelled to predict or report their implicit biases does not show that prior to the experimental situation they were aware at all of those attitudes. It is quite plausible that prior to the experiment, participants were completely unaware of their implicit biases and the experimental situations *prompt* participants to engage in explicit inferences regarding their implicit attitudes, a process which explains how they can predict or report on them. After being asked to predict their results of their IATs, for example, participants may explicitly reason that most people harbor negative implicit biases and infer that they probably harbor them themselves. Of course, given the widespread coverage of implicit bias in the media and elsewhere, ordinary people might engage in such inferences as well for nonexperimental reasons—for example, simply in the interests of self-exploration. But whatever prompts people to come to have SM-awareness of their implicit attitudes, they never have SU-awareness of them.

Hahn and colleagues (2014, study 3) attempted to control for the possibility that people were making inferences about biases in general by showing that individuals can accurately predict the variance between their own IAT scores and the "average score"—suggesting that people have particular access to their own implicit attitudes. But even this result does not undermine the Awareness View. Just as a jaundiced patient might infer that she has cirrhosis of the liver from observing certain of her own symptoms, so too might participants be capable of inferring their own implicit biases through observing their previous biased behavior. And even if one predicts that one has cirrhosis of the liver, it is still may come as a shock to have that prediction confirmed by a doctor. Again, since one becomes SM-aware of the bodily state, one can remain skeptical about one's prediction—and a doctor's testimony can strengthen one's belief in the state. Likewise, even if

people are able to predict their particular implicit biases with some accuracy, it may still feel like a surprise to find those predictions confirmed by IATs.[18]

Interestingly, Hahn and colleagues did not ask participants how they came to make such predictions about their IAT performances. But arguably one kind of relevant data regarding these predictions includes how people *justify* how they are aware of their attitudes.[19] In future work, then, experimenters might ask participants *why* they have certain attitudes. If participants were to focus on the inferences that led to their having those attitudes—for example, they must have the attitudes because they have noticed that they often act badly in the presence of certain groups—then it is arguable that they have only SM-awareness of those attitudes. Such evidence would show, at least, that participants had awareness of the inferences that led to their awareness of those attitudes. Indeed, it is arguable that in such cases participants would be making predictions about how they will perform on the IAT simply on the basis of their impact awareness of their biases without content awareness of the attitudes at all.

If, by contrast, participants are dumbfounded as to why they have attitudes or if they give justifications for the attitudes but not for *why* they are aware of them (e.g., that a certain group is simply disgusting), then it is arguable that they are SU-aware of their attitudes.[20] I predict that people's awareness of their implicit biases will fall into the former category.

Another likely possibility is that participants predicted their IAT scores on the basis of SU-awareness of what Ranganath and colleagues (2008) dub 'gut feelings'. But such gut feelings are plausibly explicit emotional reactions, such as having a sinking feeling in the pit of one's stomach when thinking about or confronting members of a certain group. Just as one might infer one's implicit attitudes from impact awareness of one's biased actions, one might do so from seemingly

[18] For a similar discussion, see Carruthers (forthcoming).
[19] I thank James Skidmore for this suggestion.
[20] For a related account of dumbfounding regarding moral beliefs, see, e.g., Haidt & Hersh (2001).

unmediated impact awareness of one's biased emotional reactions. Some theorists do identify various kinds or features of implicit attitudes, which may include such reactions. Amodio and Devine (2006), for example, distinguish the *cognitive or semantic content* of implicit attitudes from the *emotional or affective aspects* of such attitudes and suggest that these features may belong to distinct kinds of attitude. Since I am neutral here regarding the underling nature of implicit attitudes, I will not take a stand on whether or not they involve, or merely often accompany, such affective states. Either way, these affective states are arguably not (complete) implicit attitudes.[21]

6. Comparison to Intrinsic Views

Although many theorists today are drawn to Intrinsic Views, it is an important implication of the Awareness View that an attitude's being implicit/explicit is arguably an extrinsic property of that attitude—a property that the attitude may gain or lose depending on one's mode of awareness of it. If one is not aware or disposed to be aware in an SU-way of an attitude, then it is implicit—and yet one may perhaps come to be SU-aware of that attitude, thereby rendering it explicit. The Freudian perspective on Dunham and colleagues' (2008) data supports this idea. Or consider Cooley and colleagues' (2015) findings that participants who are led to classify their implicit biases as their own are more likely to have negative explicit attitudes consistent with their implicit biases than participants who were not so led. One reading of these findings is that reflecting on one's reasons (however erroneous) for having an implicit attitude tends to engender SU-awareness of it, thereby rendering it explicit.[22]

---

[21] Some theories of emotions regard them as mere "feelings" with no content, whereas others hold that emotional states represent, among other things, bodily states or relations between the environment and the subject (for an overview, see, e.g., Prinz 2004). But it is at best unclear that emotional states have the suitable sort of content to qualify as (complete) implicit attitudes.

[22] One might worry that since the Awareness View holds that an attitude is explicit just in case one is aware or *disposed to be* SU-aware of it, the fact that we can become SU-aware of an attitude that was putatively implicit shows that it was explicit all along. That is, this evidence might instead seem to suggest, like Hahn and colleagues' work, that we can be

Some theorists would seem to disagree. Levy, for example, maintains that "awareness of the content of our implicit attitudes fails to transform them into explicit attitudes: they retain the same behavioral profiles regardless of our awareness" (2014a, p. 29). But since Levy does not distinguish SM- and SU-awareness, it is unclear whether he would amend his view. Moreover, the Awareness View is consistent with the claim that one's newfound SU-awareness of a previously implicit attitude need not affect its behavioral profile insofar as it is unclear that extrinsic awareness of attitudes would affect them.[23]

There are three possibilities regarding the (putative) functional differences between implicit and explicit attitudes, among which I remain neutral for present purposes. There is the possibility, to which I am most sympathetic, that the apparent functional differences between implicit and explicit attitudes are simply illusory. Again, it is likely, for example, that some explicit attitudes are, like many implicit attitudes, uncontrollable. Alternatively, it may be that there are genuine functional differences between explicit and implicit attitudes due to the fact that such states are intrinsically distinct (e.g., propositional beliefs vs. mere conceptual associations). The Awareness View is, after all, compatible with Intrinsic Views. Naturally, whether or not attitudes that are implicit ever become explicit depends upon the facts that explain why some attitudes are implicit in the first place; perhaps the intrinsic nature of certain attitudes makes it difficult or even impossible for us to be SU-aware of them. Lastly, it could be that the relevant type of awareness engenders some kind of functional difference in the attitudes. Settling between these options is a task for future exploration.

---

and easily are aware of our implicit attitudes. But notice that participants in this study would not report attitudes in line with their implicit biases if they were not so led to reflect on their attitudes—and my proposal is that such reflection changes how one can be aware of those attitudes. Though we can become SU-aware of a once-implicit attitude, this does not show that, prior to becoming aware of it, one was disposed to become SU-aware of it. The fact that a bachelor can get married does not entail that he is disposed to be married. That is, I argue that certain psychological changes, such as reflecting on the reasons for one's having an attitude, might alter one's dispositions regarding it.

[23] On Rosenthal's account (see fn. 5), the HOTs in virtue of which we are SU-aware of our extrinsic attitudes do not alter their functional profiles (see, e.g., Rosenthal 2005, p. 185).

But whatever kind of attitudes figure in implicit bias, as long as such awareness is extrinsic, it remains at least *conceptually* possible that such attitudes might occur explicitly as well.[24]

Even if there were such intrinsic differences between implicit and explicit attitudes, they would not be what *make* attitudes implicit. Whether or not a Doxastic or Intrinsic View is correct, the Awareness View is what *fundamentally* distinguishes implicit from explicit attitudes insofar as Intrinsic Views cannot explain the kinds of evidence offered in support of the Awareness View. Even if some version of the Intrinsic View were true, if people were SU-aware of their implicit attitudes, then one would not expect the kinds of surprise reactions that figure in the Argument from Surprise, nor would we expect people to often confabulate the explanations for actions driven by their implicit attitudes.

6.1. Undercutting motivations for Intrinsic Views

Though I remain neutral here regarding the intrinsic nature of the attitudes that occur implicitly, I argue in this section is that the plausibility of the Awareness View challenges many of the reasons for adopting an Intrinsic View; the Awareness View thus paves the way for a Doxastic View on which the alleged intrinsic differences between implicit and explicit attitudes are illusory.

Consider Levy's discussion of the study by Uhlmann and Cohen (2005), wherein participants confabulate the qualifications of a job to select a candidate of a preferred group. Levy (2015, p. 814), cites such results as evidence that implicit attitudes do not behave like beliefs. We should not, Levy claims, attribute to participants the belief that "the kinds of qualifications possessed by the white

[24] Some theorists do maintain that the relevant kind of awareness of states is an intrinsic feature of those states (e.g., Kriegel 2009), in which case the Awareness View trivially amounts to a kind of Intrinsic View. But that's an optional commitment of the account. If the relevant awareness is extrinsic, then the present proposal amounts to a version of a *single-process model* (e.g., Fazio 1990). And other theorists recently seem to be converging on similar views. Carruthers (forthcoming), for example, defends the view that implicit and explicit attitudes have the same representational structures, though their different behavioral manifestations can be explained by the kinds of other states with which they are tokened. The present view is in many ways compatible with Carruthers' account, which differs from it insofar as Carruthers does not emphasize a difference in *awareness* as the fundamental maker.

(male) candidate are the ones relevant to the job" (2015, p. 814). Levy maintains that beliefs have certain characteristic features: for example, beliefs must be *inferentially promiscuous* insofar they sufficiently update other beliefs and drive behavior. Implicit attitudes, however, would seem to be too inferentially isolated from other attitudes to count as beliefs. Participants do not seem to act in ways consistent with the kind of belief that Levy describes (cf. Goldhill 2017). Moreover, attributing this kind of belief to participants seems especially implausible, as it would involve attributing to them wildly contradictory beliefs: the explicit belief that the particular qualifications used to select candidates were the legitimate criteria and the implicit belief that *whatever* qualifications the preferred candidate possessed were the relevant criteria.

But, again, the present analysis explains why the capacity to harbor such contradictions without dissonance is not implausible. Likewise, the fact that it may seem bizarre to attribute such outwardly irrational and contradictory beliefs to participants may be explained by the fact that we, as theorists, are rarely if ever SU-aware of having such beliefs. Indeed, the fact that we are never SU-aware of implicit attitudes explains why it may *seem* in general that such attitudes are functionally unlike explicit attitudes. It may be, for example, that we can and often do control our implicit attitudes; it may simply be that we are never SU-aware of such instances of control.

Perhaps most importantly, even if we are SM-aware of ourselves as having implicit biases, our SM-awareness may be *partial*, distorting how we theorize about those biases. If one maintained that we were ever in fact directly aware of attitudes (and not merely seemingly so), one might urge that we can be aware of their complete natures—or at least their complete contents. But since we are only SM-aware of implicit attitudes, if aware at all, it is plausible that we are often not fully aware of their contents. Suppose that one has an implicit attitude with the propositional content *that black men are dangerous*. After reflecting on one's past behavior regarding black men, one might become partially SM-aware of this attitude as involving a vague relation between the concepts BLACK MAN

and DANGEROUS. One might be aware of having *some* negative attitude about black men. This does not entail that the implicit attitude is not itself a belief with a propositional content—it may be—but it need not seem that way. Consequently, people classify these seemingly thin attitudes as mere associations, when in fact they are beliefs.[25]

7. Conclusions

I have hypothesized here that what makes an attitude explicit or implicit are one's modes of (un)awareness of it. Though there is much evidence that we can be aware of implicit attitudes, it is plausible that we are not aware of them in the same subjectively unmediated way that we are aware of our explicit attitudes.

Moreover, I have argued that the forgoing analysis has implications for the debate over the intrinsic nature of such attitudes. This account renders questionable the proposal that implicit attitudes are not, or do not involve, beliefs. But whether or not implicit attitudes are beliefs, I have argued that their intrinsic nature is not what *makes* such states implicit. An attitude is implicit just in case one is not aware of it in a subjectively unmediated way. Theorists and experimentalists alike should seek to explore other ways to substantiate (or refute) this hypothesis—and explore its various consequences.

Acknowledgements I thank Ralph Baergen, Zac Gershberg, Rocco Gennaro, Neil Levy, Eric Mandelbaum, Myrto Mylopoulos, Bence Nanay, Jake Quilty-Dunn, Evan Rodriguez, James Skidmore, Brent Strickland, two referees for this journal, as well as the audiences at the Conference on Interdisciplinary Perspectives on Moral Responsibility at

[25] There is, however, evidence for Intrinsic Views that the Awareness View does not as clearly bear upon. For example, Levy (2015, p. 815) discusses fascinating work suggesting that implicit attitudes seem to be, unlike beliefs, unresponsive to negation—that they are equally affected by presentations of expressions and negations of those expressions (e.g., Deutsch et al 2006; cf. Madva 2016). While there are reasons to think that implicit attitudes can be responsive to negation (e.g., Mandelbaum 2016, p. 640), I cannot settle this issue here.

Utah Valley University in March 2015, the Anthropology Department Colloquium at Idaho State University in April 2015, and the 2017 Meeting of the Southern Society for Philosophy and Psychology for their helpful discussions of these issues or comments on previous drafts of this material.

References

Amodio, D. & Devine, P. 2006. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91(4): 652-661.

Bem, D. J. & McConnell, H. K. 1970. Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation atttitudes. *Jounal of Personality and Social Psychology* 14(1): 23-31.

Block, N. 2009. Comparing the major theories of consciousness. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences IV* (pp. 1111–1121). Cambridge, MA: MIT Press.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. 2000. Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology* 79: 631-643.

Brownstein, M. forthcoming. Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology* DOI: 10.1007/s13164-015-0287-7.

Brownstein, M. & Saul, J. (Eds.). 2016. *Implicit Bias & Philosophy: Volumes I & II*. Oxford: Oxford University Press.

Carruthers, P. forthcoming. Implicit versus explicit attitudes: Differing manifestations of the same representational structures? *Review of Philosophy and Psychology* DOI: 10.1007/s13164-017-0354-3.

Chris, S. A. & Woodyard, H. D. 1973. Self-perception and characteristics of premanipulation attitudes: A test of Bem's theory. *Memory & Cognition* 1(3): 229-235.

Cooley, E., Payne, B. K., Loersch, C., & Lei, R. 2015. Who owns implicit attitudes? Testing a metacognitive perspective. *Personality and Social Psychology Bulletin* 41(1): 103-115.

De Houwer, J. 2014. A propositional model of implicit evaluation. *Social and Personality Psychology Compass* 8: 342-353.

Descartes, R. 1988. *Descartes: Selected Philosophical Writings*. J. Cottingham, R. Stoothoff, & D. Murdoch (Ed. & Tr.). Cambridge, UK: Cambridge University Press.

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.

Dunham, Y., Baron, A. S., & Banaji, M. R. 2008. The development of implicit intergroup cognition. *Trends in Cognitive Sciences* 12(7): 248-253.

Egan, A. 2008. Seeing and believing: Perception, belief formation, and the divided mind. *Philosophical Studies* 140(1): 47-63.

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

Fazio, R. H. 1990. Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology* 23: 75-109.

Festinger, L. 1957. A theory of cognitive dissonance. Palo Alto, CA: Stanford University Press.

Freud, S. 1949. *An Outline of Psycho-Analysis*. New York: W. W. Company.

Gawronski, B., Hofmann, W. & Wilbur, C. 2006. Are "implicit" attitudes unconscious? *Consciousness and Cognition* 15: 485–499.

Gawronski, B. & Payne, B. K. (Eds.). 2010. *Handbook of implicit social cognition: Measurement, theory, and applications*, New York, NY: Guilford Press.

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. S. 2008. Alief and belief. *Journal of Philosophy* 105(10): 634-663.

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

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553–561.

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

Goldhill, O. 2017. The world is relying on a flawed psychological test to fight racism. *Quartz*. 3 December. Online: https://qz.com/1144504/the-world-is-relying-on-a-flawed-psychological-test-to-fight-racism/

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. 2014. Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3): 1369-1392.

Haidt, J. & Hersh, M. A. 2001. Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology* 31(1): 191-221.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. 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.

Holroyd, J. 2012. Responsibility for implicit bias. *Journal of Social Philosophy* 43(3): 274-306.

Holroyd, J. 2015. Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition* 33: 511-523.

Johnson, C. 2013. Everyone is biased: Harvard professor's work reveals I barely know my own minds. *The Boston*

Globe. Online: <http://www.boston.com/news/science/blogs/science-in-mind/2013/02/05/everyone-biased-harvard-professor-work-reveals-barely-know-our-ownminds/7x5K4gvrvaT5d3vpDaXC1K/blog.html>, accessed Feb. 2016.
King, M. & Carruthers, P. 2012. Consciousness and moral responsibility. *Journal of Moral Philosophy* 9: 200-228.
Kriegel, U. 2009. *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford University Press.
Levy, N. 2014a. Consciousness, implicit attitudes and moral responsibility. *Noûs* 48(1): 21-40.
Levy, N. 2014b. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.
Levy, N. 2015. Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs* 49(4): 800-823.
Lewis, D. 1982. Logic for equivocators. *Noûs* 16(3): 431-441.
Machery, E. 2016. DeFreuding implicit attitudes. In Brownstein, M. & Saul, J. (Eds.), *Implicit Bias & Philosophy: Volume I* (pp. 104-129). Oxford: Oxford University Press.
Mandelbaum, E. 2014. Thinking is believing. *Inquiry* 57(1): 55-96.
Mandelbaum, E. 2016. Attitude, inference, association: On the propositional structure of implicit bias. *Noûs* 50(3): 629–658.
Madva, A. 2016. Why implicit attitudes are (probably) not beliefs. *Synthese* 193(8): 2659–2684.
Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. 2001. Taking a look underground: Detecting, interpreting and reacting to implicit racial biases. *Social Cognition* 19(4): 395–417.
Nier, J. A. 2005. How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations* 8: 39–52.
Nisbett, R. E. & Wilson, T. D. 1977. Telling more than I can know: Verbal reports on mental processes. *Psychological Review* 84: 231-259.
Nosek B. A. 2007. Implicit-explicit relationships. *Current Directions in Psychological Sciences* 16(2): 65-69.
Payne, B. K. 2001. Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology* 81: 181-192.
Prinz, J. J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University Press.
Quilty-Dunn, J. 2015. Believing in perceiving: Known illusions and the classical dual-component theory. *Pacific Philosophical Quarterly* 96 (4): 550-575.
Ranganath, K., Smith, C., & Nosek, B. 2008. Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology* 44: 386-396.
Rosenthal, D. M. 2005. *Consciousness and Mind*. Oxford: Clarendon Press.
Saul, J. 2013. Unconscious influences and women in philosophy. In F. Jenkins & K. Hutchison (Eds.), *Women in Philosophy: What Needs to Change?* (pp. 39-60). Oxford: Oxford University Press.
Schwitzgebel, E. 2010. Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91(4): 531–553.
Sullivan-Bissett, E. 2015. Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition* 33: 548-560.
Strawson, G. 1994. *Mental Reality*. Cambridge, MA: MIT Press.
Uhlmann, E. L., & Cohen, G. L. 2005. Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* 16: 474-480.