Are the States Underlying Implicit Biases Unconscious? – A Neo-Freudian Answer

Beate Krickel

ABSTRACT

Many philosophers as well as psychologists hold that implicit biases are due to *unconscious* attitudes. The justification for this *unconscious-claim* seems to be an inference to the best explanation of the mismatch between explicit and implicit attitudes, which is characteristic for implicit biases. The unconscious-claim has recently come under attack based on its inconsistency with empirical data. Instead, Gawronski et al. (2006) analyze implicit biases based on the so-called Associative-Propositional Evaluation (APE) model, according to which implicit attitudes are phenomenally conscious and accessible. The mismatch between the explicit and the implicit attitude is explained by the *Cognitive Inconsistency Approach* (CIA) (as I will call it): implicit attitudes are conscious but rejected as basis for explicit judgments because the latter lead to cognitive inconsistency with respect to other beliefs held by the subject. In this paper, I will argue that the CIA is problematic since it cannot account for the fact that implicit attitudes underlying implicit biases typically *are* unconscious. I will argue that a better explanation of the attitude-mismatch can be given in terms of a Neo-Freudian account of repression. I will develop such an account, and I will show how it can accommodate the merits of the APE model while avoiding the problems of the CIA.

1 Introduction

Consider the following case of Juliet the philosophy professor:

> Juliet is a white American philosophy professor. She knows there is no scientific evidence for racial differences in intelligence, and she argues with sincerity for equality of intelligence, a view which also harmonizes with her liberal outlook on other matters. Yet Juliet's unreflective behaviour and judgements of individuals display systematic racial bias: When she gazes out on class the first day of each term, she can't help but think that some students look brighter than others — and to her, the black students never look bright. When a black student makes an insightful comment or submits an excellent essay, she feels more surprise than she would were a white or Asian student to do so, even though her black students make insightful comments and submit excellent essays at the same rate as do the others. This bias affects her grading and the way she guides class discussion. (Schwitzgebel, 2010, p.532) (Frankish, 2016, pp. 24–25)

The phenomenon described in this story is called 'implicit bias' and it currently receives a lot of attention from different areas in philosophy and psychology (as the recently published book series by Brownstein and Saul (2016a, 2016b) shows). This attention seems to be justified given that implicit biases like in the case of Juliet the philosophy professor are of great concern for obvious social and ethical reasons. But what exactly are implicit biases? And are they indeed due to *unconscious* mental states as many philosophers and psychologists assume?

For the purposes of the present paper, the first question which concerns the *ontology* of implicit bias, will be set aside (although the theories that I am going to discuss imply a specific answer to that question). For now, presupposing the less controversial part of the answer to the ontological question suffices: implicit biases are psychological dispositions that manifest in behaviors that "involve a deviation from norms of fairness"[1] (Frankish, 2016, p. 24), and that have a cognitive categorical basis, which is called 'implicit attitude' (Mandelbaum, 2016, p. 630). Furthermore, in this paper, I will be concerned only with cases of implicit biases where the implicit attitude does not match with what the subject would report if asked about her explicit view on the respective topic (i.e., her explicit attitude). This restriction is justified because, as I will spell out in more detail below, the main reason that led researchers to think that implicit attitudes underlying implicit biases are unconscious was that they detected preferences where people were not aware of having any, or even denied having them. Hence, they inferred that these people have unconscious preferences, i.e., implicit attitudes. Indeed, in cases where a reported preference matches the measured preference, there would be no need to postulate an additional unconscious implicit attitude. Since I am interested in the question of whether implicit attitudes are unconscious or not, I will be concerned only with cases of implicit attitudes where we, *prima facie*, have good reasons to believe that they are. Hence, in this paper, I will be concerned only with cases of implicit bias where the implicit attitude does not match the explicit attitude.

Philosophers disagree on whether the implicit attitudes that form the cognitive bases of implicit biases are themselves dispositions, preferences, beliefs, aliefs[2], belief-like, affective, or mental associations, and it remains unclear how all these concepts relate (e.g.: Are beliefs dispositions? Are attitudes propositional? Are they preferences? Associations? Are preferences affective states?). Setting all this aside, the focus of the present paper will be the second question: Given recent psychological findings, is it plausible to think of the cognitive bases of implicit biases, i.e. implicit attitudes, as *unconscious* mental phenomena?

Many psychologists as well as philosophers seem to hold that the implicit attitudes that underlie implicit biases are unconscious. Although never formulated explicitly, the underlying argument seems to be an inference to the best explanation of the mismatch between the explicit and the implicit attitude (Gawronski, Hofmann, & Wilbur, 2006, p. 487): the fact that

[1] I take this from Frankish (2016) who adds that, in the present context, fairness is taken to be a norm of rationality as well as a social norm. I admit that this is less clear as it could be. Still, for present purposes an intuitive understanding suffices; at least it should be intuitively clear how Juliet's behavior deviates from norms of fairness.

[2] The term 'alief' is due to Tamar Gendler (2008) and refers to mental states that are associative, automatic, arational, affect-laden and action generating (2008, p. 641).

Juliet does not report a negative attitude towards her black students while still behaving in an unfair way towards them is best explained by the assumption that Juliet is not consciously aware of her negative attitude towards her black students. Despite the wide acceptance of the *unconscious-claim*, recently, doubts have been raised. Based on empirical findings, psychologists, such as Gawronski et al. (2006) and Hahn et al. (2014), argue that subjects can indeed introspect their implicit attitudes underlying their implicit biases, and that they are, thus, consciously aware of them. To account for the fact that implicit attitudes are conscious, Gawronski introduces the so-called *Associative–Propositional Evaluation (APE)* model. According to this model, roughly, implicit attitudes consist in associations between concepts such as AFRICAN AMERICAN and HOSTILE or STUPID that give rise to affective reactions that can be felt and are introspectively accessible. The mismatch between the reported attitude and the implicit attitude, Gawronski explains based on what I will call the *Cognitive Inconsistency Approach* (CIA) (Gawronski, 2012): implicit biases arise if an affective reaction induces a propositional evaluation, such as 'African Americans are hostile/stupid' or 'I dislike African Americans,' that is rejected due to its being cognitively inconsistent with other beliefs held by the subject such as 'African Americans are a disadvantaged group' and 'Negative evaluations of disadvantaged groups are wrong' (Gawronski, 2012, p. 662). While the subject rejects the propositional evaluation, the association, and thus, the affective reaction remain and keep influencing the subject's behavior.

In this paper, I will argue that, although the APE model is promising, the CIA is problematic. First, according to Gawronski (2012), not only the implicit attitude itself is conscious but also the process of re-establishing cognitive consistency. This is inconsistent with the fact that implicit attitudes underlying implicit biases *usually are unconscious*. This is indicated by the fact that people are usually surprised or shocked when they find out that they have implicit attitudes that do not match their explicit attitudes. Second, as I will show, the CIA is psychologically implausible.

I will show that we can avoid these problems if we explain the mismatch between explicit and implicit attitudes underlying implicit biases in terms of a *Neo-Freudian account of repression*. According to this account, implicit biases arise because the affective reactions induced by the implicit attitudes are not attended to and miscategorized in order to defend the preferred self-image, where this is taken to be due to an impulsive, non-deliberative act.

The paper will be structured as follows: in Section 2, I will explain what implicit biases are in more detail and I will motivate the unconscious claim. In Section 3, I will discuss

Gawronski's and Hahn's explanation of the attitude-mismatch, the CIA, and I will present the APE model based on which Gawronski and Hahn frame their explanation. I will show that the CIA fails. In Section 4, I will present an alternative explanation of the mismatch in terms of repression. Therefore, I will, in Section 4.1, explain what repression is taken to be in general. After that, in Section 4.2, I will introduce an analysis of the psychological mechanism of repression. Finally, in Section 4.3, I will apply my analysis of repression to the case of implicit bias and show how the mismatch between implicit and explicit attitudes can be explained based on the repression model. Section 5 will conclude with an answer to the question: Are the states underlying implicit biases unconscious?

2 What are implicit biases?

As already mentioned in the introduction, implicit biases are usually taken to be dispositions to behave or think in a certain unfair way that are grounded in *implicit attitudes*. In psychology, the term 'attitude' refers to evaluations, likings, or dislikings of different kinds of entities such as objects, persons, or groups of objects or people (e.g., 'I don't like cheese,' 'I like the Beatles') (Bohner & Dickel, 2011). Why *implicit*? Psychologists hold that attitudes can be either *explicit* or *implicit*. Explicit attitudes are said to be likings or dislikings that a person would report when prompted, and thus, that the person is consciously aware of. There is no consensus on what defines implicit attitudes. Some authors interpret the notions 'implicit' and 'explicit' in terms of the methods that are used to measure the attitude (Fazio & Olson, 2003). Explicit attitudes are measured by explicitly asking a subject to consciously think about her attitudes and report them. For example, subjects are asked how strongly they agree with statements such as "It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites" (see Henry & Sears, 2007). The most popular method for measuring implicit attitudes is the so-called *Implicit Association Test* (IAT). Roughly, in this test, for example, typical names or pictures of black or white people are combined with positively or negatively valenced words (such as 'good,' 'beautiful,' 'bad,' 'dangerous'). Subjects are asked to respond as quickly as possible (and without making too many mistakes) by pressing a button if a positive word *or* a white face is presented, and a different button if a negative word *or* a black face is presented on a screen (followed by a trial where one button is to be pressed if a black face or a positive word is presented, and another button if a white face or a negative word is presented). If it takes the subject longer to react to positive words when combined with black faces than when

combined with white faces this is interpreted as indicating that the subject has a negative implicit attitude against black people.

If we presuppose this method-based interpretation of the terms 'implicit' and 'explicit' one can characterize implicit biases in terms of a mismatch between the results of an explicit and a corresponding implicit test. Juliet, for example, would show a positive or neutral attitude towards people of color in an explicit test, but a negative attitude in the corresponding IAT. Indeed, as already mentioned in the introduction, it is this mismatch that leads many psychologists and philosophers to think of the attitudes underlying implicit bias as unconscious (Gawronski et al., 2006, p. 487). The underlying argument for this unconscious-claim seems to be an inference to the best explanation: the best explanation for why people show biases in a respective IAT or another implicit test, while not showing this bias in their explicit reports, is that these people are not consciously aware of the attitudes causing their behaviors in the implicit test.

Indeed, the idea that implicit attitudes are unconscious seems to be accepted by the vast majority of psychologists and philosophers. Here are a few examples:

> A widespread assumption underlying the application of indirect measures is that they provide access to unconscious mental associations that are difficult to assess with standard self-report measures (e.g., Bacchus, Baldwin, & Packer, 2004; Banaji, 2001; Bosson, Swann, & Pennebaker, 2000; Brunstein & Schmitt, 2004; Cunningham, Nezlek, & Banaji, 2004; Greenwald & Banaji, 1995; Jost, Pelham, & Carvallo, 2002; Phelps et al., 2000; Rudman, Greenwald, Mellott, & Schwartz, 1999; Spalding & Hardin, 1999; Teachman et al., 2001; Wilson, 2002). Specifically, it is often argued that self-reported (explicit) evaluations reflect conscious attitudes, whereas indirectly assessed (implicit) evaluations reflect unconscious attitudes. (Gawronski et al., 2006, p. 486)

> Although implicit and explicit attitudes have been distinguished along many dimensions (e.g., Bargh, 1994; Greenwald & Banaji, 1995; Gawronski & Bodenhausen, 2006) (…) awareness seems to be of particular significance, such that unconscious attitudes and implicit attitudes (or conscious attitudes and explicit attitudes) are often used as interchangeable terms (e.g., Bosson, Swann & Pennebaker, 2000; Cunningham, Nazlek & Banaji, 2004; Jost et al. 2002; Phelps et al., 2000; Rudman, Greenwald, Mellott & Schwartz, 1999; Quillian, 2008). (Hahn et al., 2014) (Italics in original)

> Unconscious forms of bias pose a problem for legal theory and practice because law's guiding model of human behavior assumes that sane, ordinary, adult behavior is the result of conscious and intentional decision-making. (Banaji, Bhaskar, & Brownstein, 2015, p. 183)

> The implicit biases that we will be concerned with here are unconscious biases that affect the way we perceive, evaluate, or interact with people from the groups that our biases "target". (Saul, 2013, p. 40)

> Examining the workings of implicit bias can illuminate a host of foundational issues in cognitive science, such as the entities that populate the unconscious mind, and how rationally responsive unconscious thought can be. (Mandelbaum, 2016, p. 629)

Furthermore, the assumption that implicit attitudes are unconscious fits with how the term 'implicit' is used in memory research:

> In cognitive psychology, individuals are said to display implicit memory for a prior event when their performance on some task shows evidence of their having been influenced by that prior event, even though they display no explicit memory for the event; i.e., they report no awareness of the event having occurred (Fazio & Olson, 2003, p. 302)

Despite its wide acceptance and its initial plausibility, the claim that implicit biases are due to unconscious mental states has recently come under attack. Several studies show that subjects indeed are aware of their implicit attitudes that underlie their implicit biases (for an overview see Gawronski (2006)). Most important for present purposes, a study by Hahn et al. (2014) suggests that implicit attitudes are indeed felt and can be accessed by means of introspection. In a series of studies, subjects were asked to predict their scores in different IATs measuring implicit attitudes toward five different social groups. Subjects had to closely look at pictures (that were later used for the IAT) and concentrate on their gut feelings. Then, they were asked questions like "How easy will this task be for you compared to this one?" or, more directly, "Will your true attitude/culturally learned association be more positive/negative towards black or white people?" Answers to these questions had to be given via different (discrete or continuous) scales, and were then translated into IAT-scores. Hahn et al. found that subjects were pretty accurate in their predictions regardless of whether implicit attitudes were described as true attitudes or culturally learned associations, regardless of how the question was formulated, and regardless of how much experience or explanation participants received before making their predictions. Furthermore, they suggest that the successful predictions indicated unique insight into implicit attitudes via introspective access to feelings induced by these implicit attitudes:

> [W]hen they were presented with the attitude targets, participants did in fact "feel" their affective reactions and reported on those reactions as their implicit attitudes, even though they might have invalidated those same responses as a basis for their explicit attitudes. (Hahn et al., 2014)

Now, the question arises: If implicit attitudes are consciously accessible, what explains the mismatch between the explicit and implicit attitude? Following Gawronski (2006), Hahn et al. argue that the best explanation for the mismatch between implicit and explicit attitudes does not lie in the fact that the former are unconscious:

> [A] person could be entirely aware of his or her implicit attitude, but not report it on an explicit attitude measure due to its inconsistency with other propositions. Implicit-explicit correlations reveal whether people consider their implicit attitudes valid bases for explicit attitudes, not whether they are aware of them. (Hahn et al., 2014)

What explains the mismatch between implicit and explicit attitudes is that subjects invalidate their felt implicit attitudes as the basis for explicit judgments. Gawronski and Hahn et al. do not take this to be a case of subjects simply lying about their true attitudes. Rather, they provide an explanation of the mismatch in terms of, what I will call, the *Cognitive Inconsistency Approach* ('CIA') which relies on a particular view of the ontology of implicit attitudes, the so-called *Associative–Propositional Evaluation* ('APE model') (Gawronski & Bodenhausen, 2006). I will present and discuss the CIA and the APE model in the next section.

3 Explanation of the Mismatch I: The APE Model & The Cognitive Inconsistency Approach

The so-called 'APE model' (*Associative–Propositional Evaluation*) (Gawronski & Bodenhausen, 2006) is a popular account of attitudes in psychology and constitutes an answer to the *ontological question* mentioned in the introduction—it makes assumption about what implicit and explicit attitudes *are*. According to this model, implicit attitudes are *associative evaluations*, whereas explicit attitudes consist in *propositional evaluations*. Associative evaluations are affective reactions that are induced by an associative mental network connecting concepts such as BASKETBALL and BOUNCING, or AFRICAN AMERICAN and HOSTILE. These associations are not sensitive to subjective truth ascriptions or any logical relations. In contrast to that, propositional evaluations are based on syllogistic inferences and have truth values (Gawronski & Bodenhausen, 2006, p. 694), like for example, "Basketball can be played in gyms" or "African Americans are hostile." The two kinds of evaluations are supposed to differ with respect to the underlying kind of information processing. Associative evaluations are immediate affective reactions. Propositional evaluations are made and tested in an reflective system that is superordinate to an associative store that transforms associative evaluations into propositional evaluations (Gawronski & Bodenhausen, 2006, p. 694). Then, the propositional evaluations are contrasted with further information that is available in the reflective system as to whether this information is consistent with the associative evaluation (Gawronski et al., 2006).

The APE model is promising because it can account for the empirical data presented in the previous section. Implicit attitudes (associative evaluations) induce affective reactions, which can be felt since they come with a certain phenomenology. Thus, implicit attitudes are consciously accessible *via* the affective reactions they induce. Still, the APE model is compatible with the fact that people's attitude-reports do not mention the implicit attitude. This is because sometimes the associative evaluation in incompatible with further information available in the reflective system, which leads to a rejection of the associative evaluation. I will call this alternative explanation the *Cognitive Inconsistency Approach* (CIA). Gawronski uses the following example to illustrate the central idea (Gawronski, 2012, 662): Imagine a person who comes to believe the following three claims, where the first belief is due to a negative affective reaction towards African Americans:

(1) I dislike African Americans.
(2) African Americans are a disadvantaged group.
(3) Negative evaluations of disadvantaged groups are wrong.

This set of beliefs is *cognitively inconsistent* (Gawronski, 2012). Since the "psychological need for cognitive consistency is as basic as hunger and thirst" (Gawronski, 2012, p. 652), the subject has to do something to red rid of the inconsistency. Gawronski argues that cognitive consistency can be re-established by rejecting any of the beliefs. Implicit biases arise when the subject resolves the inconsistency by rejecting (1) and thereby creates a mismatch between the propositional evaluation and the corresponding associative evaluation. The implicit bias occurs due to the fact that the negative affective reaction that led to first holding (1) persists even though (1) is rejected. How exactly is this process supposed to work? Most importantly, Gawronski argues that it is an entirely conscious process:

> In such cases [cases of cognitive inconsistency], the necessary reassessment of the activated information involves conscious awareness of the involved processing steps, such as the negation (i.e., reversal of the truth value) of a particular proposition or the search for information that resolves the inconsistency. (Gawronski & Bodenhausen, 2014)

Thus, according to Gawronski, implicit biases seem to arise as a consequence of four steps: first, the person has an affective reaction towards an object or a person. Second, the person generates a propositional evaluation in line with the affective reaction. Third, the subject realizes that the propositional evaluation is inconsistent with other beliefs she holds. Fourth, the propositional evaluation is consciously rejected and replaced by a different propositional evaluation. Still, the affective reaction remains.

The APE model is promising because it can explain many features of implicit biases. First, it can explain the automaticity of biased behavior (Andersen, Moskowitz, Blair, & Nosek, 2007; Holroyd & Sweetman, 2016). Biased behavior, such as Juliet's, is usually thought of as being triggered automatically without involving any kind of conscious deliberation. The associations that constitute implicit attitudes, according to the APE model, induce affective reactions that again trigger certain behaviors without the involvement of any kind of conscious deliberation. Second, it is plausible to assume that implicit attitudes need not be unconscious in order to count as *implicit*. Clearly, Juliet might become aware of the fact that she has a mental association between AFRICAN AMERICAN and STUPID. Still, she would keep her explicit liberal beliefs and does not become an explicit racist (this point has been noted by, for example, Frankish (2016), and Machery (2016)). Third, the APE model analyses implicit attitudes as non-propositional which seems to be suggested by empirical research (although this is a matter of debate among philosophers; see Madva (2015), and Mandelbaum (2016)). Fourth, the APE model can account for the empirical findings mentioned in the previous section showing that subjects have access to their implicit attitudes. Plausibly, implicit attitudes can be introspected via the affective reactions that they induce.

Still, Gawronski's and Hahn's explanation in terms of the CIA is problematic. First, even if Hahn et al.'s study succeeds in showing that subject can *in principle* introspect their implicit attitudes, it also suggests a further aspect: subjects were not conscious of their implicit attitudes *before* the experiment. Subjects were surprised or even shocked after becoming aware of their implicit attitudes. A participant of Hahn et al.'s study reports:

> I feel guilty because I think that I am an intuitive person. Yet, based on this test, it shows that if I go with my initial gut instinct about race and value judgments I am actually quite judgmental. (Hahn et al., 2014)

As this quote indicates, this participant was not aware of her implicit attitude until the experiment. Generally, people seem to be surprised or even shocked when they realize that they have implicit attitudes that conflict with their explicit attitudes. This is suggested also by a qualitative study by Hillard et al. (2013) in which the affective reactions to the IAT were measured. Hillard et al. report that 14 out of 32 subjects report being shocked after receiving their IAT-results, and 6 participants found the test "eye-opening" (Hillard, Ryan, & Gervais, 2013, p. 503). This at least suggests that most subjects are not aware of their implicit attitudes before taking the IAT.

The CIA predicts that implicit biases arise as the result of a *conscious* process of reestablishing cognitive consistency which requires a *conscious* detection and reflection of an

affective reaction, and which results in the rejecting of a previously formed belief in line with the affective reaction, while the affective reaction itself remains in existence (see above). Hence, if the CIA were correct, subjects who are implicitly biased in a way that does not match their explicit beliefs would have to be *consciously aware* of their implicit attitudes (their affective reaction). But if this were the case, subjects would not be surprised or shocked when they find out that they are implicitly biased. In other words, the CIA fails to account for the fact that implicit attitudes underlying implicit bias, although in principle consciously accessible, usually *are* unconscious.

Second, the act of re-establishing cognitive consistency as postulated by the CIA is psychologically implausible. It requires people to be able to simply decide to like someone or something although they know that they have a negative affective reaction towards that person or thing based on which they already formed the belief that they dislike that person or thing. This seems to be as impossible as holding that someone can simply decide to like bananas although she beliefs that she does not like them. Consider the following set of beliefs (where the first belief is due to an affective reaction):

1. "I dislike bananas."
2. "I do not want to eat things I dislike."
3. "Bananas are healthy."
4. "I want to eat everything that is healthy."

This set of beliefs is inconsistent because it implies that the person wants and does not want to eat bananas. Plausibly, the person will not cope with this inconsistency by deciding to stop believing that she dislikes bananas. It would be simply odd to think that this is what people in fact do or even *can* do. The person might be able to decide that she *should* like bananas, and thus, she might start to act in a way that makes eating bananas more pleasant for her (e.g. she could make delicious banana milk shakes, or she could use conditioning strategies). But surely, she cannot simply start believing that she likes bananas only by intellectually realizing that this would be the better thing to do. Plausibly, the inconsistency is dealt with not by rejecting any of the beliefs but by putting a lot of effort into letting the right belief cause her actions. That is, even if the person does eat a banana, she will still agree to all statements 1-4. She will say something like "I really do not enjoy eating this banana. But it is so healthy. Therefore, I eat it". Similarly, someone who has this set of beliefs but still refrains from eating bananas will say something like "I wished I would like bananas since they are so

healthy. But I really dislike them. Therefore, I do not eat bananas". But this strategy is not applicable in the cases of implicit biases since the having of an implicit bias implies that one does *not* have the explicit belief that corresponds to the affective reaction. Hence, in cases of implicit bias the first statement will be rejected by the biased person. But is there a psychologically more plausible story as to how this rejection is made?

In what follows, I will present an analysis of implicit biases that avoids the problems of the CIA. This analysis accepts the basic assumptions of the APE model but provides a different story of how the mismatch between explicit and implicit attitudes arises.

4 Explanation of the Mismatch II: Repression

The idea that the phenomenon of implicit bias can be explained in terms of repression has already been introduced into the debate by Wilson et al. (2000, p. 105). Wilson et al. argue that an implicit bias could be due to repression in cases where "feelings are kept out of awareness because they are anxiety-provoking" (Wilson et al., 2000, p. 105). Alas, Wilson et al. do not provide an explicit account of implicit bias in terms of repression. Hence, the idea was an easy target for criticism (Gawronski et al., 2006, p. 493) and, apart from that, has not received much attention in the literature. In the following sections, I will provide such an account.

4.1 What is repression?

Repression, as I will understand it here, is a psychological mechanism that has the function of defending the ego. Repression occurs when someone experiences what is called an *inner conflict*. This conflict consists in the incompatibility of the existence or disappointment of certain desires with the social norms internalized by the subject and the preferred self-image (as, for example, a child that is unavailingly struggling for the love of its parents; or a person that has homosexual desires but lives in a society punishing homosexuality). In order to get rid of the negative emotions (e.g., shame, guilt, sadness, anger, hate) and the negative impact on the self-image ('I am not loveable,' 'I am a sinner') resulting from the inner conflict, the subject represses the desire, relevant beliefs, emotions, or memories of the events that triggered the conflict (as for example, experiences of being abused by the parents).

There are different attempts to explain how repression works exactly (Billon, 2011; Boag, 2007; Brakel, 2009; Erdelyi, 2006; Hart, 1982). None of them is without problems, which is

why I will present my own analysis of the mechanism of repression.[3] Since the analysis of repression is not the focus of the present paper, I will not explicitly defend my approach against alternative accounts. It will suffice to show that it satisfies all criteria that an adequate approach to repression has to satisfy. Following Kihlstrom (2002), Wilson and Dunn (2004) provide such a list of criteria of adequacy (Wilson & Dunn, 2004, p. 495). They argue that the phenomenon of repression has the following features:

(1) People are motivated to keep thoughts, feelings, or memories outside of awareness;
(2) the attempt to keep material out of awareness is itself an unconscious process;
(3) people succeed in removing the undesired material from consciousness;
(4) the material, once removed from consciousness, still exists in memory and continues to influence people's thoughts, feelings, or behavior; and
(5) the material is recoverable; i.e., people can become aware of it if the repressive forces are removed

According to Wilson and Dunn's list, repression is a motivated, but unconscious act of removing material from consciousness. It is *motivated* because repression has the purpose of protecting the ego. It is *unconscious* because the subject does not know that she is repressing. It is an act of *removing* material from consciousness since the material was conscious at some point when it created an inner conflict. The conflict is resolved by rendering the material unconscious. The removed material, although unconscious, *remains* in existence – it is not deleted but remains as an unconscious mental state. A central assumption of psychoanalysis is that repressed mental states *influence* the behavior of the subject by, for example, causing neurotic symptoms or other less pathological behaviors. A further central conviction of psychoanalysis is that, at least typically, repressed states are *recoverable*. Otherwise psychoanalytic therapy would be useless. In the following section, I will present an analysis of the mechanism of repression that accounts for these criteria of adequacy.

4.2 A mechanism for repression

In order to make sense of repression in a way that accounts for the five criteria of adequacy, I will make use of four central concepts: *phenomenal consciousness*, *attention*, *control of action*, and *categorization*. Note that each of these four concepts is itself subject of a debate in the philosophy of mind and in psychology. Here, I will only introduce these notions very

[3] See Billon (2011) and Boag (2007) for overviews of different approaches to repression and their problems.

basically. Apart from that, an intuitive understanding suffices for present purposes. I will first explain briefly what these four concepts imply. Then, I will explain how we can make sense of repression with help of these concepts.

Phenomenal consciousness. I will use the terms 'phenomenal consciousness' and 'p-consciousness' to refer to the phenomenal, what-it-is-like character of mental states, such as feelings, emotions, moods, and pains. The account of repression that I am going to present, roughly, takes repression to consist in different ways to render particular p-conscious states unconscious. Therefore, the remaining three concepts will be characterized with regard to how they relate to p-conscious states.

Attention. A mental state can be unconscious in the sense that one does not pay attention to it. A mental state can be more or less conscious depending on the *degree of attention* it gets. For example, the keyboard of my computer is only in the periphery of my attention while I am typing (Dingler, Schmidt, Bakker, Hausen, & Selker, 2016). Phenomenal states can be in the periphery of attention as well. For example, you might have a strong headache which was first in the focus of your attention but after a while, when you started working, became part of the periphery of your attention. Still, the headache could switch back into the focus of your attention, for example, if someone asked you whether you are feeling better. There is an ongoing debate in philosophy of mind as to whether there can be phenomenally conscious states that a subject is not paying attention to at all.[4] For present purposes it is not important whether there can be p-conscious states that do not get any degree of attention. It suffices to allow for p-conscious states that appear only in the periphery of attention. Phenomenal states that are never in the focus of attention are unconscious in a further sense: since, as I will explain below, correct categorization requires attention, these phenomenal states cannot be categorized correctly. A low degree of attention will lead to miscategorization or enables only very broad categorizations.

Categorization. A p-conscious state can be unconscious in the sense that it is categorized incorrectly or only in a very broad way. For example, I might be afraid of the long journey that I am about to make but miscategorize the feeling as a stomachache, or I might only broadly categorize it as a weird nervous feeling (Billon, 2011, p. 15). Miscategorization can happen along different dimensions. First, the *type* of the state can be incorrectly categorized. Is the p-conscious state hate, sadness, depression, or a stomachache? Second, the intentional object of the p-conscious state (the object it is about) can be incorrectly identified. Is the intentional object the mother, some other person, or is there no intentional object at all? For

[4] For arguments in favor of the view that phenomenal consciousness can arise without attention, see Aru and Bachmann (2013) and Block (2013). For objections, see Schlicht (2012).

example, a p-conscious feeling might be fear of death or fear of a dog. Third, the categorization concerns the valence of the p-conscious feeling. Valence might not be a binary feature (pleasant vs. painful) but may involves degrees and mixtures (Colombetti, 2009). A p-conscious state can feel rather pleasant but mixed with painfulness, as for example, the feeling you have when you are about to leave home for an exciting journey. Plausibly, a categorization along this dimension cannot be incorrect. Fourth, the owner of the state of a p-conscious can be identified correctly or incorrectly. Under normal conditions, people do identify themselves correctly as the owner of there p-conscious states. But in rare pathological cases the owner is misidentified (as for example in cases of so-called *thought insertion* (Mullins & Spence, 2003)). Erdelyi describes a case, *Case N.*, where a patient comes to believe that "it is not he but Nixon who is the homosexual [i.e., has homosexual feelings]" (Erdelyi, 2006, 506). Hence, patient N. misidentifies Nixon as the owner of his feelings. A merely broad categorization of a p-conscious state occurs when only the valence and the owner are correctly identified without a specification of the type, and the intentional object of the state.

As already indicated above, any kind of categorization (correct or incorrect) of a p-conscious state presupposes that the state is in the focus of attention at least temporarily (Boag, 2006, p. 513). In other words, if a p-conscious state remains in the periphery of attention and never gets into the focus of attention, it cannot be categorized. For example, even in order to form the thought 'I have a weird negative feeling' you have to focus your attention at least shortly onto the feeling. In order to realize that this feeling is a certain type of emotion, further higher-level cognitive processes are necessary. For example, you have to think about what might have caused the feeling. This is not possible by only focusing on the feeling itself. You also have to access you memory and find the event that induced the feeling.

C*ontrol of action*. A p-conscious state can be unconscious in the sense that the way it causes behaviors is not under the control of the agent. More specifically, p-conscious states can control actions in different ways. First, a p-conscious state can figure in *deliberative* thought that leads to a decision that leads to an action. For example, you might realize that you have a negative feeling when thinking about a certain option for action ("I should better not do this since I have a bad feeling regarding this option") based on which you decide not to choose this option. Second, a p-conscious state can *impulsively* lead to an action. Frijda et al. characterize impulsive actions as follows:

> Actions are considered 'impulsive' when and because they are not preceded by deliberation or the conscious representation of some action goal. An impulsive action is thus defined as a non-deliberate action that serves the purpose of rendering one's relation

> to the object, event, or state of the world more pleasant or less unpleasant. (Frijda, Ridderinkhof, & Rietveld, 2014, p. 1)

In this sense, impulsive actions are not arbitrary but are purposive without demanding for conscious goals or intentions. Impulsive actions are triggered by perceiving or thinking about an object, person or event, which is pleasant or unpleasant for the subject and have the purpose of causing a change in the state of the world such that it becomes more pleasant for the subject.

In a nutshell, a phenomenally conscious state can be unconscious in the sense that it is not in the focus of attention, in the sense that it is categorized incorrectly along at least one of the dimensions 'type,' 'intentional object,' 'valence,' or 'owner,' or it is categorized only very broadly. Finally, a p-conscious state can be unconscious in the sense that it causes actions impulsively without the deliberative control of the agent. With the help of these four concepts, I will now present an analysis of the mechanism of repression:

a. *Inner conflict:* Repression is triggered by an inner conflict between the occurrence of a feeling F and the explicit desire not to have F, where the desire is a consequence of the internalized social norms, or the self-image.
b. *Repression*:
    i. *Impulsive shift of attention:* The inner conflict triggers an impulsive act of shifting attention away from F.
    ii. *Habitualization:* After multiple occurrences of F, F triggers i. without inducing an inner conflict due to habitualization.
    iii. *Miscategorization*: If F is only in the periphery of attention, F is disposed to be miscategorized by the subject.

In order to illustrate this account, take, for example, a person who is in love with her best friend's partner (Hart, 1982). This person experiences an inner conflict between her affections for her best friend's partner and the desire not to feel this affections (1). This desire is a consequence of her self-image of being a good friend and the internalized social norm that friends do not fall in love with their best friends' partners (1). This inner conflict induces a negative feeling (e.g., shame) that triggers an impulsive shift of attention away from her feelings (2i.). Since the person now does not pay attention to her affections anymore, the conflict is not consciously induced anymore, and therefore, the shame is not induced. Furthermore, since she does not pay attention to her feelings, her affections will be (if at all)

only in the periphery of her attention which disposes these feelings to be miscategorized by the subject (2ii). The subject will, for example, note that she feels strangely exited when seeing her best friend's partner which she will (if at all) explain by, for example, generally being in a good mood. Since the impulsive shift of attention is triggered every time she notices her feelings for her best friend's partner, at some point, these feelings will trigger the shift of attention immediately due to habitualization. At this point, the inner conflict will not be induced anymore. The idea of repression involving some kind of habiltualization is known in the neuropychoanalytic literature. For example, based on their research on suppression (which is, roughly, voluntary repression), Andersen and Green suggest that "if retrieving diversionary thoughts becomes habitual, inhibition may be sustained without any intention of avoiding the unwanted memory" (Anderson & Green, 2001, p. 368). The habitualization results in an association between the p-conscious state and the thoughts/behaviors that are used to shift away attention.

Based on assumptions 1. and 2i.-iii., we can account for the five criteria of adequacy listed above. First, according to the present model, repression is *motivated* in the sense that it has the purpose of avoiding an undesired feeling. Second, based on the terminology used here, repression can be characterized as *unconscious* in the sense that (i) it occurs not deliberatively but impulsively, (ii) it occurs without demanding attentional resources – the subject need not pay attention to the process of repression itself (note that this does not exclude the fact that the subject, at least at first, has to pay attention to the p-conscious state F). (iii) A typical feature of repression is that the subject does not correctly categorize the process of repression *as repression* (Boag, 2006, p. 514). Third, the model accounts for the idea that repression consists in the *removal* of unwanted material from consciousness. At first the p-conscious state is conscious in the sense that it is felt, at least briefly in the focus of attention, and at least imprecisely correctly categorized. Later, the p-conscious state is unconscious in the sense that it does not get attention anymore and it is miscategorized. The complete *removal* of material from consciousness is successful, according to the present model, if the connection between trigger and repression behavior becomes automatized. Depending on the strength of the association between the trigger and the behavior, the subject will become unable to focus on the feeling that triggers the conflict. Fourth, the present approach assumes that repression does not result in a deletion of the p-conscious state. Rather, when not paying attention to a feeling, and when miscategorizing a feeling, the feeling remains, and can still influence the subject's behavior. For example, the person who fell in love with her best friend's partner will still behave in specific ways towards the partner that indicates her feeling for him/her. Still,

fifth, the repressed material is recoverable. People can get to know the relevant concepts; they can learn to pay attention to their behavior and feelings, and they might come to recognize a certain pattern in their behavior and thus identify it as repression behavior. For example, the person in our example might at some point realize that her feelings are due to her affection towards her best friend's partner, and she might come to consciously feel the resulting conflict, and she might realize that she tends to repress it. All this might be rather difficult for her and often this recovery requires professional help (for example, a psychoanalytic psychotherapy). Still, even if the subject learns to pay attention to her repression behavior and her feelings, and even though she might come up with the right categorization of the p-conscious state and the repressive behavior, this does not imply that she thereby is able to deliberatively control her behavior. This issue of control comes in in two forms: first, since the repression itself is triggered impulsively, the subject might not be able to prevent the repressive acts even though she is in principle aware of her repressive behavior. Secondly, the repressed feelings trigger certain other behaviors (e.g., avoiding being alone with the best friend's partner, being extra nice to the best friend) that the subject might not be able to control even after learning about her repression.

I will not explicitly defend this approach to repression here. There might be less complex accounts that still satisfy the criteria of adequacy. For present purposes, it suffices to see that this is *one* possible way to make sense of repression.

4.3 The mismatch between implicit and explicit attitudes results from repression

The mismatch between implicit and explicit attitudes that gives rise to implicit biases can be analyzed in terms of repression in the following way: Take again our example of Juliet the philosophy professor. For reasons not discussed in the present paper, Juliet has a negative implicit attitude towards people of color (i.e., certain mental associations between, for example, AFRICAN AMERICAN and CRMINAL and STUPID). This implicit attitude, induces a negative feeling (e.g., fear; the feeling might also be something like a feeling of superiority) every time Juliet sees a person of color. At some point (plausibly as a child)[5], Juliet was aware of this negative feeling. Young Juliet might have found that people of color are scary, for example. As she grew older, she learned that it is wrong to think that people of color are criminals, that they are dangerous or stupid. She identified with being a liberal,

[5] There is scientific evidence that children under the age of 10 are aware of the affective reactions induced by their implicit attitudes, and that they form the corresponding explicit attitudes (Baron & Banaji, 2006; Newheiser & Olson, 2012). This changes at around the age of ten, where explicit attitudes become more liberal, but the scores in the relevant IATs still indicate the presence of an implicit bias.

open-minded person. Still, the negative affective reactions were induced when interacting with people of color. This induced an inner conflict in Juliet. She was scared by people of color but she did not want to be. Juliet felt shame, and this shame automatically led her shift her attention away from her fear that she felt when interacting with people of color. She was only vaguely aware of these feelings. She might have felt slightly unpleasant when interacting with people of color, what she explained by her being stressed by her work, by her being nervous in front of class, or by her bad mood. Since Juliet had been shifting away her attention from her negative feelings towards people of color, the feelings and the attention shift became connected which led to the fact that Juliet now does not feel the conflict anymore when she encounters people of color. Nonetheless, the implicit attitude, and the induced feelings remain and keep influencing Juliet's behavior. Depending on the strength of the repression (i.e., the strength of the association between the trigger and the repression behavior), Juliet is in principle capable of becoming aware of her implicit attitude by paying attention to her feelings and her behavior, and by applying the right concepts. However, this will not necessarily lead to an elimination of the biased behavior since the feeling will remain unconscious in the sense that it will still trigger impulsive actions (for example, when Juliet guides discussions in the classroom). Avoiding the biased behavior, again, might take a lot of effort and might require professional help.

Based on this analysis of implicit bias, we can avoid the problems afflicting Gawronski's CIA. As argued in Section 3, the CIA is problematic because it assumes that implicit biases arise as the result of a conscious process of re-establishing cognitive consistency. This would require the subject to be conscious of the affective reaction, the corresponding evaluation, the inconsistency, and the rejection of the evaluation. By postulating this conscious process, the CIA cannot explain why people that show implicit biases are usually surprised and shocked when they find out about their implicit attitudes. On the basis of the repression account, we can explain this phenomenon: people are surprised and shocked because, as a matter of fact, they were not aware of their implicit attitudes before the experiment due to repression. And they were not aware of these attitudes for a reason: because the corresponding affective states conflicted with certain desires implied by their self-image, or internalized social norms. Hence, due to the fact that the experiment draw attention to the feeling the inner conflict became conscious again, and thereby induced negative feelings (that show by, e.g., being shocked, or being angry).

Second, the repression account does not require subjects to simply decide to like someone or something while knowing that they do not like that person or thing. The reason is that the

present model does not take implicit biases to result from cognitive inconsistency that has to be resolved by consciously rejecting a belief. Rather, implicit biases are due to impulsive actions that aim at avoiding a violation of a desire by shifting attention away from the feeling that induces the conflict.

Finally, based on the repression model of implicit bias, one can make several empirical predictions based on which one can validate the model. The repression model predicts that biased subjects will shift their attention away from feelings that are induced by stimuli they are negatively biased against. The repression and the APE model, and thus the CIA, make the same prediction that biased subjections will have affective reactions if presented with a stimulus they are negatively biased against. Based on CIA, one would predict that subjects are aware of these feelings, know that these feelings indicate that they have a certain preference, but that this preference is rejected because it is inconsistent with other beliefs. The repression model predicts that subjects are *not* aware of these feelings since they will automatically shift their attention away from these feelings. If at all, they will not make a connection between the feeling that is induced by the stimulus and the stimulus. They will come up with alternative explanations. In a nutshell, based on the repression model, we can make the following predictions:

1. Stimuli a subject is biased against, will induce negative (or positive) feelings in the subject.
2. Subjects are not aware of these feelings.
3. Subjects will automatically shift their attention away from these feelings.
4. Subjects will come up with alternative explanations for these feelings.

Statement 1 is a prediction based on which alone one cannot distinguish between the CIA and the repression model, since both accounts would make this prediction. In contrast to that, prediction 2 is incompatible with CIA since this model states that implicit biases arise due to a conscious act of re-establishing cognitive consistency, where the inconsistency arises due to a belief that is formed based on a feeling (see first objection against the CIA). Note that prediction 2 is compatible with the fact that the feelings are accessible *in principle*. It just states that the default is that people are *not* aware of these feelings (unless something happens that forces them—with sufficient strength—to pay attention to their feelings). If predictions 3 and 4 turned out to be true, this would speak in favor of the repression model directly since it would support the core assumptions of the model.

5 Conclusions

Are the mental states underlying implicit biases unconscious? According to the model presented in this paper, implicit attitudes that give rise to implicit biases are unconscious in three senses: first, they are unconscious because they automatically trigger certain kinds of behaviors without the control of the subject. Second, they are unconscious because people do not pay attention to the feelings induced by their implicit attitudes. Third, since these feelings are in the periphery of attention at best, subjects tend to categorize them incorrectly. The main claim of this paper was that implicit attitudes are unconscious in the second and third sense for a reason: they are repressed in order to protect the subject's preferred self-image and in order to avoid conflicts with internalized social norms. Repression was described as an impulsive act of shifting attention that gets triggered by a negative feeling induced by an inner conflict between a p-conscious feeling (that is first correctly categorized and attended to) and the explicit desire not to have such a feeling, where the desire is derived from the preferred self-image and/or the internalized social norms. At some point, the p-conscious feeling will directly induce a shift of attention due to habitualization. This shift of attention leads to the feelings being disposed to be miscategorized.

References

Andersen, S., Moskowitz, D. B., Blair, I. V, & Nosek, B. A. (2007). Automatic thought. In *Social psychology* (2nd ed., pp. 133–172). Guilford Publications.

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, *410*(6826), 366–369. http://doi.org/10.1038/35066572

Aru, J., & Bachmann, T. (2013). Phenomenal awareness can emerge without attention, *7*(December), 1–2. http://doi.org/10.1068/p3127

Banaji, M. R., Bhaskar, R., & Brownstein, M. (2015). When bias is implicit, how might we think about repairing harm? *Current Opinion in Psychology*, *6*, 183–188. http://doi.org/10.1016/j.copsyc.2015.08.017

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes. *Psychological Science*, *17*(1), 53–58. http://doi.org/10.1111/j.1467-9280.2005.01664.x

Billon, A. (2011). Have we vindicated the motivational unconscious yet? A conceptual review. *Frontiers in Psychology*, *2*(SEP), 1–20. http://doi.org/10.3389/fpsyg.2011.00224

Block, N. (2013). The Grain of Vision and the Grain of Attention, *1*, 170–184. http://doi.org/10.1002/tht.28

Boag, S. (2006). Can repression become a conscious process? *Behavioral and Brain Sciences*,

29(5), 513–514. http://doi.org/10.1017/S0140525X06239116

Boag, S. (2007). Realism, Self-Deception and the Logical Paradox of Repression. *Theory & Psychology*, *17*(3), 421–447. http://doi.org/10.1177/0959354307077290

Bohner, G., & Dickel, N. (2011). Attitudes and Attitude Change. *Annual Review of Psychology*, *62*, 391–417. http://doi.org/10.1146/annurev.psych.121208.131609

Brakel, L. A. (2009). *Philosophy, Psychoanalysis and the A-rational Mind*. OUP Oxford. Retrieved from https://books.google.de/books?id=axCVNwAACAAJ

Brownstein, M., & Saul, J. (2016a). *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. OUP Oxford. Retrieved from https://books.google.de/books?id=TJQDDAAAQBAJ

Brownstein, M., & Saul, J. (2016b). *Implicit Bias and Philosophy: Metaphysics and Epistemology*. Oxford University Press. Retrieved from https://books.google.de/books?id=Mi7oCwAAQBAJ

Colombetti, G. (2009). Appraising valence. *Journal of Consciousness Studies*, *12*(8–10), 103–126. http://doi.org/10.1016/j.amjmed.2012.04.013

Dingler, T., Schmidt, A., Bakker, S., Hausen, D., & Selker, T. (2016). *Peripheral Interaction*. Springer. http://doi.org/10.1007/978-3-319-29523-7

Erdelyi, M. H. (2006). The unified theory of repression. *The Behavioral and Brain Sciences*, *29*(5), 451–499. http://doi.org/10.1017/S0140525X06009113

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their Meaning and Use. *Review Literature And Arts Of The Americas*, 297–327. http://doi.org/10.1146/annurev.psych.54.101601.145225

Frankish, K. (2016). Playing double: implicit bias, dual levels, and self-control. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology* (Vol. I, pp. 23–46). Oxford: Oxford University Press.

Frijda, N. H., Ridderinkhof, K. R., & Rietveld, E. (2014). Impulsive action: emotional impulses and their control. *Frontiers in Psychology*, *5*(June), 1–9. http://doi.org/10.3389/fpsyg.2014.00518

Gawronski, B. (2012). Back To the Future of Dissonance Theory: Cognitive Consistency As a Core Motive. *Social Cognition*, *30*(6), 652–668. http://doi.org/10.1521/soco.2012.30.6.652

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. http://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2014). The associative-propositional evaluation model: Operating principles and operating conditions of evaluation. In *Dual-process theories of the social mind.* (pp. 188–203). Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc11&NEWS=N&AN=2014-08812-013

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499. http://doi.org/10.1016/j.concog.2005.11.007

Gendler, T. S. (2008). Alief and Belief. *The Journal of Philosophy*, *105*(10), 634–663.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology. General*, *143*(3), 1369–92. http://doi.org/10.1037/a0035028

Hart, W. D. (1982). Models of repression. In R. Wollheim & J. Hopkins (Eds.), *Philosophical essays on Freud* (pp. 180–202). Cambrdige University Press. http://doi.org/10.1017/CBO9780511554636

Henry, P. J., & Sears, D. O. (2007). The Symbolic Racism 2000 Scale. *Political Psychology*, *23*(2), 253–283. http://doi.org/10.1111/0162-895X.00281

Hillard, A. L., Ryan, C. S., & Gervais, S. J. (2013). Reactions to the implicit association test as an educational tool: A mixed methods study. *Social Psychology of Education*, *16*(3), 495–516. http://doi.org/10.1007/s11218-013-9219-5

Holroyd, J., & Sweetman, J. (2016). The Heterogeneity of Implicit Bias. In M. Browstein & J. Saul (Eds.), *Implicit Bias and Philosophy: Metaphysics and Epistemology* (pp. 80–103). Oxford University Press. http://doi.org/10.1007/s13398-014-0173-7.2

Kihlstrom, J. F. (2002). No need for repression. *Trends Cognitive Science*, *6*(12), 502.

Machery, E. (2016). De-Freuding Implicit Attitudes. In M. Browstein & J. Saul (Eds.), *Implicit Bias and Philosophy: Metaphysics and Epistemology* (pp. 104–129). Oxford University Press.

Madva, A. (2015). Why implicit attitudes are (probably) not beliefs. *Synthese*, *193*(8), 2659–2684. http://doi.org/10.1007/s11229-015-0874-2

Mandelbaum, E. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous*, *50*(3), 629–658. http://doi.org/10.1111/nous.12089

Mullins, S., & Spence, S. A. (2003). Re-examining thought insertion. *The British Journal of Psychiatry*, *182*(4), 293 LP-298. Retrieved from http://bjp.rcpsych.org/content/182/4/293.abstract

Newheiser, A.-K., & Olson, K. R. (2012). White and Black American children's implicit intergroup bias. *Journal of Experimental Social Psychology*, *48*(1), 264–270. http://doi.org/10.1016/j.jesp.2011.08.011

Saul, J. (2013). Implicit Bias, Stereotype Threat, and Women in Philosophy. In *Women in Philosophy* (pp. 39–60). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199325603.003.0003

Schlicht, T. (2012). Phenomenal consciousness, attention and accessibility, *11*(3), 309–334. http://doi.org/10.1007/s11097-012-9256-0

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493–518. http://doi.org/10.1146/annurev.psych.55.090902.141954

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–26. http://doi.org/10.1037/0033-295X.107.1.101