Implicit Bias: From Social Structure to Representational Format*

Josefa Toribio

Received: 11/04/2017
Final version: 03/10/2017

BIBLID 0495-4548(2018)33:1p.41-60
DOI: 10.1387/theoria.17751

ABSTRACT: In this paper, I argue against the view that the representational structure of the implicit attitudes responsible for implicitly biased behaviour is propositional—as opposed to associationist. The proposal under criticism moves from the claim that implicit biased behaviour can occasionally be modulated by logical and evidential considerations to the view that the structure of the implicit attitudes responsible for such biased behaviour is propositional. I argue, in particular, against the truth of this conditional. Sensitivity to logical and evidential considerations, I contend, proves to be an inadequate criterion for establishing the true representational structure of implicit attitudes. Considerations of a different kind, which emphasize the challenges posed by the structural social injustice that implicit attitudes reflect, offer, I conclude, better support for deciding this issue in favour of an associationist view.

Keywords: implicit attitudes, associationism, logical sensitivity, social structures.

RESUMEN: En este artículo cuestiono la tesis de que la estructura representacional de las actitudes implícitas responsables del comportamiento implícitamente sesgado es proposicional—en lugar de asociacionista. De acuerdo con la propuesta criticada, si la conducta implícita sesgada puede ocasionalmente ser modulada por consideraciones lógicas y evidenciales, entonces la estructura de las actitudes implícitas responsables de esa conducta es proposicional. Cuestiono, en particular, la verdad de este condicional. Sostengo que la sensibilidad de las actitudes implícitas a consideraciones lógicas y evidenciales resulta ser un criterio inadecuado para establecer su verdadera estructura representacional. Consideraciones de otro tipo, que enfatizan los desafíos planteados por la injusticia social estructural que las actitudes implícitas reflejan, ofrecen, concluyo, un mejor apoyo para decidir esta cuestión a favor de una visión asociacionista.

Palabras clave: actitudes implícitas, asociacionismo, sensibilidad lógica, estructuras sociales.

* Versions of this material were presented to audiences at UCL, University of Granada, University of Barcelona, University of Graz and ECAP9. I would like to thank those in attendance on these occasions for their critical feedback. I would also like to thank two anonymous referees for this journal, for their helpful comments on an earlier draft. Research for this paper was supported by MINECO (Ministerio de Economía y Competitividad) under grant agreement FFI2014-51811, AGAUR (Agència de Gestió d'Ajuts Universitaris i de Recerca) under grant agreement 2014-SGR-81 and from the European Commission's H2020 programme under grant agreement H2020-MSCA-ITN-2015-675415.

Introduction

Most of us, despite sincerely and justifiably considering ourselves to be unprejudiced agents, consciously committed to egalitarianism in all its forms, are often surprised to discover that we harbour implicit attitudes that betray our unprejudiced, egalitarian explicit beliefs. Such implicit attitudes reflect constant exposure to stereotypical portrayals of members of, and items in, all kinds of different categories: racial groups, professions, women, nationalities, members of the LGBTQ community, moral and political values, etc. We may thus very well find that we are better and faster categorizers when e.g. female names are paired with family-oriented tasks or when negative words are paired with pictures of black faces even if we disavow sexism and racism—and even when we ourselves are female or black.

The expression 'implicit bias' is sometimes used to refer, not to the mental states responsible for certain behavioural outputs, but to the behavioural outputs themselves. Often, however, the expressions 'implicit attitude', 'implicit bias' and 'stereotype' are taken to be synonyms, and are used to refer to the mental states responsible for implicitly biased behaviour. I here follow this trend. Throughout the paper, my focus is on the mental states responsible for implicitly biased, often discriminatory, behaviour.

According to what I would like to call 'the Associationist View', implicit attitudes are associations, i.e., mental states typically connecting one or two concepts and a valence (usually negative) or two or more concepts, one of which (again usually, but not always) has a negative slant. The claim that implicit attitudes are associations is consistent with the idea that they are e.g. images or mental maps, but the main contrast to be drawn here is between associations and propositionally structured mental states. In particular, the claim conveys the idea that implicit attitudes are different from beliefs, despite their similarities and despite the fact that we can express many, if not all, our implicit attitudes in linguistic form. The Associationist View is particularly prominent among social psychologists.[1] Stressing the fact that these associations are mostly unconscious, here is an illustration: 'Implicit attitudes [stereotypes] are social category associations that become activated without the perceiver's intention or awareness when he or she is presented with a category cue' (Blair et al. 2001, 828). The Associationist View is also popular in philosophy. Jules Holroyd, a well-known scholar on this matter, for instance, writes (Holroyd 2012, 275): 'An individual harbours an implicit bias against some stigmatised group (G), when she has automatic cognitive or affective associations between (her concept of) G and some negative property (P) or stereotypic trait (T).' Philosophers who endorse what Brownstein (2016a) calls a Sui Generis Model of implicit attitudes (Gendler, 2008a,b 2011; Brownstein & Madva 2012; Madva 2012, 2016) are committed in one way or another to the Associationist View.

The mental structure of associations is taken to be that of a network formed on the basis of the mutual causal activations of the concepts and valences involved. The main characteristic of such networks is that the relation between associated nodes is not predication. Nodes in the network do not decompose into a truth-evaluable propositional structure. Associations are instead the result of reliable, reinforcing connections, as they are typically learnt by exposure. For instance, on the Associationist View, the widespread—albeit of-

[1] The standard usage of 'attitude' in this context is thus different to the standard usage of 'attitude' in philosophy, where attitudes are taken to be relations to propositions.

ten implicit—attitude that we normally express with the sentence "women are caring" is taken to consist of an associative structure where the concepts WOMAN and CARING are mutually activated due to our being constantly exposed to images of women in nurturing roles, as opposed to e.g. leadership ones.[2] On the Associationist View, conditioning and reinforcement are taken to be the two standard ways of forming associations. The strength of this type of associations between concepts or concepts and valences depends on the frequency and strength of the exposure.

The cognitive dynamics of associations is also different to that of propositionally structured representations. It is a process of spreading activation in a network formed on the basis of spatiotemporal contiguity. The transitions that fall under the category of thinking are not governed by syntactic rules, but by proximity and strength of connection. A typical associative transition can occur between any two (or more) representations regardless of their content. For instance, I can associate bunny rabbits and Easter just in virtue of having often experienced these items/events together. A typical inferential transition between propositions, by contrast, rely on the content and formal structure of the representations involved, as in e.g. a modus ponens transition from if p then q and p, to q. Dual-system theories (see e.g. Sloman1996; Kahneman 2011 or Evans and Stanovich 2013) reflect this dichotomy, as their basic tenet is the postulation of two very different types of cognitive processes—two very different reasoning systems. According to these theories, system 1, also called 'the associative system' is fast, intuitive, automatic and, of course, associative; it requires little cognitive capacity and is independent of endorsed truth-values. System 2 or 'the rule-based system' operates on propositionally structured mental representations through symbolic rules; it is slow, reflective and sequential; it often requires a large amount of cognitive capacity and it is sensitive to truth-values.

The Implicit Association Test (IAT) (Greenwald et al. 1998) and sequential priming, together with other tests,[3], have become classic tools for unmasking the degree to which we are subject to the tyranny of stereotypes and are widely used in Social Psychology, even though the relevance of response latencies as a good measure of implicit attitudes is not without criticism. Some theorists do question the interpretation of the scores from tests such as the IAT (see e.g. Karpinski & Hilton 2001; Olson & Fazio 2004; Rothermund & Wentura 2004), but the debate is still going on and responses to the criticisms are certainly widespread (see e.g., Banaji 2001; Greenwald, Nosek, Banaji & Klauer 2005).[4] The general

---

[2] Of course, there are also non-associative explanations of how WOMAN and CARING can be related. One could straightforwardly hold the connection to be propositional. Alternatively one could think of CARING as a prototypical feature of the concept WOMAN, as typically posited by a prototype theory of concepts. The example here is meant to illustrate just the type of structure that, according to the Associationist View, explains this common automatic activation.

[3] E.g., the Affect Misattribution Procedure (AMP) (Payne et al. 2005) or the Go/No-go Association Task (GNAT) (Nosek & Banaji 2001).

[4] See, in particular, the recent controversy over the meta-analysis of studies that link subjects' IAT scores and their actual discriminatory behaviour. Greenwald, Poehlman, Ulmann and Banaji (2009) argue for a strong link between these two variables. Oswald, Mitchell, Blanton, Jaccard and Tetlock (2013) question the link and focus on the influence of overt biases in the participants. Greenwald, Banaji and Nosek (2015) quickly replied to the Oswald et al. meta-analysis. Additional studies since then keep feeding the debate.

consensus in social psychology is that standard indirect methods are reliable means of getting information about a phenomenon characterized by its opacity to introspection and its frequent clash with our explicit beliefs.

Indeed, what is most characteristic of implicit attitudes is that they permeate our perception, actions and decision in an unconscious manner. They are seldom the objects of awareness or easily accessible through introspection, which makes them especially resilient to change—resilient, but not unchangeable. Implicit attitudes can be modulated. According to a long-standing view in social psychology, extinction and counterconditioning are the two major sources of change, since we acquire implicit attitudes mainly through conditioning and reinforcement. Some recent studies, however, seem to suggest that implicit attitudes can also be changed through evidential and rational considerations (Gawronski et al. 2005; Sechrist and Stangor 2001). These empirical results have recently been used to argue against the Associationist View and in favour of what Michael Brownstein (2016a) calls the Doxastic Model of implicit attitudes (Mandelbaum 2016). According to the Doxastic model, implicit attitudes are propositional states, i.e., either beliefs (De Houwer 2014; Egan 2011; Hughes et al. 2011; Mandelbaum 2013, 2016; Mitchell et al. 2009; Smith 2005, 2012) or states that fall short of being beliefs, but are, nevertheless, propositionally structured (Levy 2014).

The modulation of implicit attitudes as a way of assessing their representational structure is the focus of this paper. In particular, my aim is to discuss the conditional that moves from the claim that implicit biased behaviour can occasionally be modulated by logical and evidential considerations to the view that the structure of implicit attitudes is propositional (Mandelbaum 2016). Mandelbaum's defence of this conditional is an inference to the best explanation. I argue, by contrast, that the Associationist View can give a satisfactory account of this type of modulation and is to be preferred to Doxastic Model. My argument, however, remains neutral with regard to the truth of the converse of Mandelbaum's endorsed conditional: if implicit attitudes are *insensitive* to logical and evidential considerations, then they are associations. Most research in social psychology, especially the work of Gawronski and collaborators, is driven by this second hypothesis. In philosophy, Madva (2016), although sidestepping the issue of the *real* representational structure of implicit attitudes, argues for their *insensitivity* to logical form, thus making them different from beliefs. My focus, I insist, is just on the considerations that make the move from logical sensitivity to propositional structure appear plausible—not the other way around. The positive part of my argument moves away from issues about logical sensitivity to focus instead on methodological issues about what makes implicit attitudes the distinctive mental states that they are and how to best characterize them given such distinctive features. The idea, in a nutshell, is the following. Even if implicit attitudes can sometimes be modulated by evidential and rational considerations, this is not their most distinctive characteristic, and any inquiry into their nature that emphasizes this aspect will be off-target as far as best explanations go.

The paper is organized as follows. In Section 1, I offer some counterexamples to the *general* idea that the rational and evidential sensitivity of mental states—any type of mental state—is best explained by their having a propositional structure. I acknowledge that the counterexamples I offer do not constitute a knock down argument. The dialectic here is to cast doubt on the truth of the general conditional on which Mandelbaum's more specific one seems to depend. I then put forward, in Section 2, a more specific argument that targets directly Mandelbaum's (2016) analysis of the evidence he takes to be conclusive against

the Associationist View. I there question the inference Mandelbaum makes from behaviour modulation to attitudes' representational properties. Cases of modulation of implicitly biased behaviour through rational and evidential sensitivity, I argue, do not settle the issue of the propositional representational format of the implicit attitudes themselves. In Section 3, more positively, I argue that the Associationist View remains the loveliest explanation, in Lipton's (2004) sense, of the phenomenon of implicit attitudes, i.e., it provides the best understanding of their central characteristics. To highlight such characteristics, we have to move back to the social arena where pervasive structural social injustice tunes and fosters our minds.

1. Modulation of (other) mental states

As a first step in a strategy to defend the Doxastic Model of implicit attitudes, Mandelbaum (2016) sets to prove the truth of the following conditional: if implicit attitudes can be modified through rational argumentation (or any evidential considerations that involve sensitivity to logical structure), then, they are not associations; they are propositionally structured mental states. The final step of the positive proposal takes implicit attitudes to be just plain beliefs.

Mandelbaum offers abundant evidence that seems to support the claim that implicitly biased behaviour can indeed be sensitive to rational and evidential considerations. I will discuss some of these experimental results below. In this Section, I make a preliminary move and cast doubt on the plausibility of the general version of the conditional on which Mandelbaum's specific conditional seems to depend, i.e., I question the view that, for any type of mental state, sensitivity to logical and/or evidential considerations settles the issue of their representational structure being propositional. I will follow Mandelbaum in taking the conditional as an expression of an inference to the best explanation. So, what I aim to show in this Section is that, *in general*, for any type of mental state, their sensitivity to logical and/or evidential considerations is not necessarily best explained by assuming that they are propositions. I focus on two types of mental states: pain and disgust.[5]

The mental state we call 'pain' is typically considered to be a sensation, with no content and, *a fortiori*, no representational structure. Yet, the experience of pain can be modulated through evidential and rational considerations. Placebo and nocebo studies clearly show this. There is ample evidence of top-down modulatory circuits that strongly change the experience of pain to the point of completely eradicating its typical phenomenology. Patients who are told that they will receive an analgesic while undergoing e.g. the removal of a molar report to feel less pain after they are given a placebo injection. This phenomenon, called "placebo analgesia", has been demonstrated to work in roughly one third of the subjects involved in the relevant studies (see e.g. Beecher 1955). Similarly, telling patients that they

[5] It could be argued that, even if sensitivity to rational considerations is not a decisive criterion for establishing, in general, the representational format of mental states, it is decisive for establishing the representational format of implicit attitudes. But then, those who endorse this view owe us an explanation of what makes implicit attitudes so special and the answer cannot be that implicit attitudes are just attitudes, i.e., states with a propositional format, on pain of circularity.

will receive a drug that would increase their pain —a hyperalgesic drug— has also a top-down effect on the affective phenomenology of the experience, resulting in much higher pain scores (Ossipov et al. 2010, 3780).

It could, of course, be argued that philosophical theories of pain vary substantially and that the placebo and nocebo cases can be treated as counterexamples to the general conditional under review just in case pain could be uncontroversially recognized a sensation with no content. There are, after all, some representational accounts of pain and, on some of these accounts, the feeling of pain is understood as the representation of some sort of tissue damage.[6] My point, however, is not to offer a knock down argument, but to help undermine the connection between rational sensitivity and propositional structure with regard to a mental state, pain, which seems, at least *prima facie*, a bad candidate for having such a representational format, given its sensory nature. Even representational theories of pain, when the issue is how to best explain the modulation of pain illustrated by the placebo and nocebo cases, offer a different kind of explanation. Pain modulation is often accounted for by appealing to the cognitive penetrability of experience, i.e., the idea that verbal information about the composition of the drug has a causal effect on the phenomenology of the pain experience itself. Again, we may question the general plausibility of the cognitive penetrability thesis (see e.g. Firestone & Scholl 2016), and even if we do not, we still need to say a lot more about how cognitive penetration works. Nevertheless, my point remains: to account for the modulation of pain by verbal information by positing a propositional representational format does not seem to capture what is most characteristic of this mental state: its phenomenology.

Consider now one of our basic emotions: disgust. Studies on disgust as a food-related emotion show that the valence of olfactive experiences involving this emotion can be modulated by giving the subject information about the source of the food-odour. For example, if someone who likes cheese is asked to sniff the odour coming out of an opaque vial after being told that it contains cheese, she likes the odour. But she does not like it if she is told that the vial contains faeces, even if the odours are exactly the same (Rozin 1987, 24). Providing "evidence" about the source of a particular odour hence changes the affective phenomenology of the experience.

Pre-theoretically, we tend to think of *basic* emotions as feelings, as mental states with no content and hence no representational structure. If we endorsed this view, the above studies on disgust would be a clear counterexample to the general conditional under discussion, i.e., to the idea that modulation by evidential sensitivity entails propositional structure. However, as in the case of pain, philosophical accounts of emotions come in different varieties, some of which, so-called cognitivist theories, characterize emotions as judgments (Nussbaum 2001), as sets of beliefs and desires (Marks 1982) or as mixed states whose components are beliefs, desires, and feelings (Oakley 1992).[7] But, again, to endorse a propositional view of emotions based on their sensitivity to evidential considerations, as e.g. Mandelbaum (2013) does, seems to be methodologically dubious: it amounts to focusing on the

[6] See e.g. Martínez (2011) for the view that pain has imperative content or Bain (2013) for the view that pain has indicative content with an evaluative component.

[7] See also Mandelbaum (2013) for an analysis of these experimental results that invites viewing disgust as a propositionally structured emotion.

opposite of what is most characteristic about them: their resilience to be modulated by rational or evidential considerations.

The take home message is thus the following. The *general* link between sensitivity to rational and evidential considerations and propositional structure is, at least, questionable, especially for mental states whose most salient characteristic is precisely their resilience to change, as in the case of pain and emotions. For the most part, the representational accounts on the market of these types of mental states are motivated by considerations of a different kind—considerations that are neutral with respect to the representational format of the states.

In the next Section, I adopt a different approach. Instead of questioning the general strategy behind Mandelbaum's target conditional, I focus instead on the specific evidence that allegedly connects implicit attitudes' evidential sensitivity to propositional structure.

2. Rational sensitivity: the evidence and what follows from it

Let us summarize Mandelbaum's argument. The thesis he tries to refute is:

T: The implicit attitudes responsible for implicitly biased behaviour are mental states with an associative structure.

And this is how Madelbaum tries to falsify it.

P1: If the implicit attitudes responsible for implicitly biased behaviour are mental states with an associative structure, then implicitly biased behaviour can be changed or eliminated *only* by altering certain environmental contingencies, i.e., either by extinction or counterconditioning.
P2: Logical and evidential factors change or eliminate implicitly biased behaviour.
C1: The implicit attitudes responsible for implicitly biased behaviour are not mental states with an associative structure.
C2: The implicit attitudes responsible for implicitly biased behaviour are just beliefs.

In what follows I'll try to show that we do not need to accept C1 because P1 is false—i.e., although the argument is valid, it is not sound. Even if implicitly biased behaviour can be changed by factors other than extinction or counterconditioning, such as logical and evidential factors, we should not accept, for this reason, a propositional view of implicit attitudes. I will not offer an argument against C2 and simply assume (pace Levy 2014) that the transition from C1 to C2 is unproblematic, were C1 true.

Even though I have re-rewritten Mandelbaum's argument as a valid *modus ponens* argument, we should not forget that the real form of the argument is that of an inference to the best explanation, i.e., the claim Mandelbaum defends is that the apparent rational sensitivity of implicitly biased behaviour observed in some experimental set-ups is best explained by implicitly attitudes being propositionally structured. What thus needs to be shown is not just that the data could also be explained within an associationist framework, but also that this framework is, on the whole, a better explanation of implicit biases' distinctive properties than Mandelbaum's Doxastic account. My aim in the rest of the paper is hence two-fold. First, in this Section, I discuss two of the experiments Mandelbaum takes to be decisive to support the truth of P1 so as to isolate an unwarranted step in their interpre-

tation. I also suggest an alternative explanation for the apparent rational sensitivity of implicit attitudes that remains faithful to general associative principles. Second, in the next and final Section, I motivate why the Associationist view of implicit attitudes may still be the best explanation of their central features.

Before engaging with the first of the two tasks just mentioned, I would like to briefly discuss two other studies, which Mandelbaum does not consider. The reason for bringing them here is to offer a mixed view of P2, but also to introduce an element of caution when interpreting the experimental results surrounding implicit biases modulation. In the first, Kawakami and collaborators (2000) examine the effect of training in negating stereotypical associations related to skinheads and race on future behaviour. They show that repeatedly negating stereotype-congruent pairs of traits (by pressing a button labelled "NO" when exposed to stereotypical pairings, such as skinhead-hostile) and affirming stereotype-incongruent pairs of traits (by pressing a button labelled "YES" when exposed to counter-stereotypical pairings, such as skinhead-friendly) can later modify the influence of these stereotypes on a person categorization task, even for stereotypical traits that are not involved in the training. These results do initially support the plausibility of P2, given the essential role of negation in the training. However, even here we have to be careful. Kawakami et al. (2000) also conjecture that affirmation of counter-stereotypes may have an influence on the weakening of the initial associations, since they did not find that non-stereotypical traits were more often associated with the relevant categories after the training.

In a follow-up study, Gawronski and collaborators (2008) questioned precisely the idea that *just* negating stereotypical traits could be the relevant explanatory factor involved in the observed change of stereotype activation. In this study, the two tasks were separated, i.e., subjects either received training in negation of racial stereotypes or affirmation of non-stereotypical racial associations, but not both. Their results show that while there is a reduction in the activation of stereotypes after non-stereotypical affirmation training, when the training consists just in the negation of stereotypical-congruent racial traits, it has the opposite effect, i.e., it *enhances* instead of reducing the influence of stereotypes in subsequent person categorization and evaluation tasks. These remarkable results seem to suggest that simply being exposed to stereotype-congruent traits, even when the training consists in negating the association, may be enough to reinforce previously held implicit attitudes.[8] They also seem to confirm Kawakami et al.'s (2000) conjecture that the counterconditioning involved in the counter-stereotype pairing part of their study might have played an important role in the later reduction of stereotype association.

If Gawronski et al. (2008) are right, the truth of P2 would seem to shatter. At a minimum, as it often happens in social psychology, evidence is mixed. Interestingly, however, and moving now to the issue that concerns me here, even Kawakami et al. (2000) do not conclude from their findings that implicit attitudes have propositional structure. In fact, when addressing the issue of the processes responsible for the stereotype reduction activation found in their study, they offer three options, all of which remain faithful to a characterization of implicit attitudes as associations: (i) strengthening and weakening of category-trait associations; (ii) higher motivation to stop their previous associations and

[8] See Madva (2016) for an argument in favour of the insensitivity of implicit attitudes to the logical form of thoughts and information based on this kind of results.

(iii) a combination of the two (Kawakami et al. 2000, 884). Less surprisingly, Gawronski et al. (2008) also take their results to reinforce the associative nature of implicit attitudes. The best explanation of this reinforcement of implicit attitudes, even after training with negation of stereotype-congruent pairings, they claim, is that implicit attitudes are associations.

If we were to formulate Gawronski's view in terms of a conditional, however, it would be the converse of P1: if implicit attitudes are not sensitive to logical and evidential considerations, then they are associations. Needless to say, a conditional may be true while its converse is false. So, I do not take these results to be directly in conflict with Mandelbaum's approach. As I said, the purpose of this part of the discussion is to raise a cautious voice about the interpretation of experimental results involving changes in implicit biases. In what follows, I will discuss two other studies, which Mandelbaum does review in his (2016) paper. I aim to show that the evidence they provide does not conclusively establish that the representational format of implicit attitudes is propositional.

The first of the two studies I have selected (Gawronski et al. 2005) involves the measurement of both implicit and explicit attitudes vis-à-vis Cognitive Balance Theory (Heider 1958). According to Cognitive Balance Theory (CB henceforth) to maintain psychological stability, we form relationships that balance our attitudes, toward people, events, activities or ideas.[9] The classic example to illustrate CB's predictions is the formation of interpersonal relationships, the idea, for instance, that if A dislikes B and is told that B dislikes C, A would end up liking C. The mental transitions involved in adjusting our attitudes in this way seem to have an inferential character. According to Mandelbaum, if such transitions are inferential, then they operate on propositional structure. With that in mind, Mandelbaum (2016, 638) argues for a (contrapositive) version of P1:

> ... if we can find support for something like Balance Theory among implicit attitudes, we can be reasonably sure that implicit attitudes aren't partaking in an associative process but instead have some sort of logic operating over them.

Gawronski at al. (2005) do indeed look into the effects of CB on the formation and change of both implicit and explicit interpersonal attitudes. They do it by reproducing an office environment and examining the way in which participants form attitudes about the people in it. They first present participants with photos of their just acquired colleagues, i.e., unfamiliar individuals (CS1s), while pairing the photos with consistently positive or consistently negative statements. The target is to make participants form positive or negative attitudes toward these CS1s. Once these relationships have been established, in a second part of the experiment, participants are introduced to another set of people photographs. This time, however, the only information they have about these different unfamiliar individuals (CS2) is whether they are liked or disliked by the previous CS1s. After this second step, participants are, finally, given an affective priming task to evaluate their implicit attitudes toward both CS1s and CS2s.

---

[9] Similar principles run the work done in Cognitive Dissonance Theory (Festinger 1957), according to which we have an inner drive to hold all our attitudes and beliefs in harmony and avoid disharmony or dissonance among them. Both theories aim to give an account of the idea that we seek to eliminate inconsistent beliefs and attitudes and we do that by, sometimes, engaging in irrational or maladaptive behaviour.

As it was to be expected, if participant S formed a positive implicit attitude toward CS1 A, and CS1 A liked CS2 B, then S also reacted positively toward B. Gawronski and collaborators also found that if e.g., participant S formed a negative attitude toward CS1 A and A disliked CS2 B, then S reacted positively toward B—although this type of balanced triads were not obtained in a parallel but inversed order case, i.e., when S was first told, before getting any negative information about A, that A disliked B. Here is Mandelbaum's (2016, 639) interpretation of the results:

> [The hypothesis that implicitly biased behaviour is caused by some sort of associative process or structure] predicts that you should have enhanced negative reactions toward the CS2 because you a) are encountering the CS2 as yoked to negatively valenced CS1 and b) are activating another negative valence because you are told that the CS1 *dislikes* the CS2. I have no opinion on whether two wrongs make a right, but I'm confident that if you find two negatives making a positive, what you've found is a propositional, and not an associative, process.

The structure of the argument is as follows: If CS1 is associated with a negative valence and becomes associated with CS2, which also has a negative valence (associated with the word "dislike"), then the additional negative valence should increase the negative valence associate with CS2. If the propositional information that CS1 dislikes CS2 yields, in accordance with CB, a positive reaction toward CS2 despite the summing of negative valences, then such a transition seems to operate over propositionally structured mental states.

How plausible is this interpretation of the results? Let's grant, for the sake of the argument, that acquiring information about (a disliked) CS1's dislike for CS2 leads to a positive response toward CS2s. After all, I am not questioning P2 in Mandelbaum's argument.[10] According to P2, logical and evidential factors can change or eliminate implicitly biased behaviour. Yet, the inference from the fact that evidential factors can have an effect on implicitly biased behaviour to the claim that the structure of the attitudes themselves is propositional would work only if implicitly biased behaviour was *just* the result of implicit attitudes—that no other associative or non-associative factors were involved in our implicitly biased behavioural outputs. Yet, no one in social psychology would deny that all sort of factors—associative, non-associative and even non-attitudinal processes[11]—have to be taken into account when offering explanations of implicit attitudes' modulation (see e.g. Calanchini & Sherman 2013 and my discussion of the Quadruple Process Model below). It is perfectly consistent to maintain that implicit attitudes are associations and that factors other than counter-conditioning and extinction can modulate implicitly biased behaviour. So P1 is false.

My point, just to be clear, is this: since changes in implicitly biased behaviour are caused by all sort of factors, not just implicit attitudes, moving from (certain types of) changes in biased behaviour to specific properties of the attitudes themselves—without taking into account the role of the other factors involved in the production of biased behaviour—makes the propositional hypothesis loose plausibility.

---

[10] Even if evidence is less than conclusive here. In a previous study, Gawronski and Strack (2004) provided evidence in support of the idea that CB did not affect implicit attitudes.

[11] Non-attitudinal processes are (either associative or non-associative) processes that affect implicit task performance without affecting specifically the content of the attitude. They are "domain-general processes unrelated to specific attitude content" (Calanchini & Sherman 2013, 661).

There is plenty of evidence now about the poor correlation of IAT scores with other measures of implicit bias, including affective priming, with changes in real world biased behaviour. A non-negligible number of recent meta-analysis studies have made it clear that classic experimental measures of implicit biases are not reliably tracking individual's implicit attitudes or dispositions to biased behaviour due to fundamental methodological mistakes, such as a worryingly low test-retest reliability. Test-retest reliability (r) for any psychometric instrument ranges between 0 and 1. A value of $r = 1$ indicates that the test gives the same result when repeated at different time intervals. r is usually considered acceptable for a psychometric test when its value is 0.8. For the IAT, however, some studies (see e.g. Bar-Anan & Nosek 2014) report a test-retest reliability as low as $r = 0.4$.[12] The suspicion is that this methodological oddity is due precisely to the multiplicity of factors responsible for biased behaviour. And if this is the case, then poor correlation of IAT scores with other measures of implicit bias could be used to undermine the inference from changes in biased behaviour to changes in the attitudes themselves—changes used, in turn, to settle the issue of their representational format.

More important for my purposes here are the results of a recent meta-analysis focused on *modulation* of implicit attitudes, and not just predictability of implicitly biased behaviour. Forscher et al. (2016) run a thorough discussion of different experimental results involving all kinds of psychometric measures and plausibly question the idea that apparent changes of implicit biases in experimental set-ups give us reliable information about changes in real world implicitly biased behaviour. The general idea behind this work, and the reason I bring it up in connection with Mandelbaum's interpretation of Gawronski et al. (2005)'s experiments is that, even if we accept that certain procedures, e.g., verbal information, do in fact change implicit biased behaviour in certain experimental conditions, there is very little evidence showing that these changes replicate themselves in the world outside the lab. So, even if we accepted that experiments like the one above show changes in implicit attitudes, there may be little reason to think that such changes are replicable in the real world—and again, this may very well due to the multiplicity of factors involved in biased behavioural modulation. So, we should be cautious about taking these experimental results as the definite proof for endorsing a hypothesis about the propositional representational structure of the implicit attitudes. Indeed, evidence and rational considerations have generally proved to be poor instruments of change. As I will argue in the last Section, the main reason is that implicit attitudes are a reflection of unjust structures in our social environment and it seems unlikely that real changes can occur without changing first such social structures (*pace* Brownstein 2016b).[13]

---

[12] See also e.g. Carlsson and Agerström (2016). Unlike the debate between Oswald et al. (2013) and Greenwald et al. (2015), which focuses on the predictive power of psychometric instruments like the IAT with regard to predicted biased behaviour or on general methodological issues, Carlsson & Agerström focus on the behaviour itself, in particular, on disparate treatment (or discrimination) based on race.

[13] To be accurate, Brownstein does not completely deny this point. Instead, he points out some of the limitations of what he calls the "world-first strategy" and the "situationist", outwardly-focused ethics, i.e., the idea that to change implicit attitudes one has to change the world first or to seek/avoid situations which would potentially make us more biased. He argues in favour of three types of self-regulatory strategies based on the special relationships between social context (a term of art in his work) and implicit attitudes.

Here is the second experiment I would like to discuss from the evidence presented in Mandelbaum (2016). Sechrist and Stangor (2001) recruited 54 (26 male and 28 female) white undergraduate psychology students from the University of Maryland and divided them in two groups according to their high- or low-prejudice attitudes toward African American based on a preliminary testing session using the Pro-Black Scale (Katz & Hass 1988). The participants were told that the task was to assess people's views about different social groups. To this end, they were given the same Pro-Black Scale test again, i.e., a questionnaire with nine favourable and nine unfavourable stereotypical traits associated with African American. They were asked to rate their perception of the percentage of African Americans with those traits. Participants were also given feedback about other University of Maryland students' opinion on these matters. This information took the form of the average percentage of peers who agreed with the answers provided by the students in the experiment. This percentage was calculated by comparing their own responses with their peers' responses. Randomly after completing the questionnaire, each participant received one of two different types of feedback: half of the participants in both the high- and low-biased groups was told that 81% of the University of Maryland students agreed with them and the other half was told that only 19% of the students in their university agreed with them.

The way Sechrist and Stangor measure the participants' implicit attitudes toward African American was to confabulate some excuse to ask them to leave the room where the experiment took place and sit in a different room with seven chairs. An African American confederate was already sat in the seat closest to the door of that room. Implicit attitudes were measured in terms of how close or further away from the confederate the participant sat—a standard test for measuring racial attitudes. Participants who were highly biased against African Americans according to the initial test, after learning that their peers disagreed with them, sat closer to the African American than those in the same group who were informed that their peers agreed with them. Although the results were less significant in the case of participants who were low-biased, these students also sat closer to the African American confederate after learning that their peers agreed with them. If the feedback was that the majority of their peers disagree with them, they sat further away.

Mandelbaum presents these results as, again, a case in favour of a propositional view of the representational structure of implicit attitudes since students changed their implicitly biased behaviour as a result of evidence, in the form of the students' peers' opinion, and, if this occurs, according to Mandelbaum, then (the best explanation is that) their implicit attitudes have a propositional structure (the contrapositive of P1). Here is the relevant part of his analysis (Mandelbaum 2016, 642):

> [O]n a purely associative story, ... the high-prejudice person who receives negative feedback, finding out that his peers disagree with him, should now have his negative affect exacerbated. Yet this exacerbation of negative affect causes him to move closer, not further away from the experimental compatriot.

Mandelbaum thus takes disagreement to be a negatively valenced cue—the kind of cue that, following CB principles, has a negative influence on us. If associationism about implicit attitudes is true, Mandelbaum argues, the already negative attitude toward African

Americans held by students in the high-biased group should get reinforced upon learning about their peers' pronounced disagreement, so they end up sitting further away from the African American confederate. Yet, the behaviour of the high-prejudice person need not be explained by CB or dissonance theory principles acting over mental representations with a propositional structure. The effects of disagreement on the high-prejudice person could perfectly be the result of both adjustments within already held negative associations involving African Americans—with the disagreeing majority acting as an inhibiting element on an already activated association—and the regulatory action over these mental representations of an external factor subject to the principles of CB.

Social psychologists have known for a long time that all sorts of non-associative processes can influence subjects' performance on tasks involving measures of implicit attitudes. In particular, the so-called Quadruple Process Model (Sherman et al. 2008) alerts us to there being four different parameters behind the explanation of any performance on tasks of this kind. These parameters are: activation of associations, detection of correct responses, overcoming bias, and guessing. The presence (or absence) of preconscious goals has also been isolated as a non-associative, regulatory factor with similar functions (Moskowitz et al. 1999). Calanchini et al. (2014) apply this model to estimate the contribution of such processes to IAT performance in an effort to examine whether the influence is relatively general, i.e., not related to the content of the attitude to be measured, or attitude-specific. Of these four parameters, the most important for the purpose of this paper is the overcoming bias parameter—which also includes the motivation to do so and the detection of conflict that this process generates. Overcoming bias is typically invoked to explain how activation of associations and detection of correct responses interact when in conflict. In such cases, overcoming bias is invoked as a regulatory, conflict-resolving process that prevents already activated associations from influencing biased behaviour (Calanchini et al. 2014, 1286). A majority of peers disagreeing with you would naturally be taken to play exactly this role, namely the role of regulating the already established association between African American and a negative valence in light of what is taken to be (given the majority of disagreeing peers) the correct response.[14]

A closer look at the kind of evidence Mandelbaum examines thus undermines the link between logical / evidence sensitivity and propositional structure. If my argument here is sound, the fact that implicitly biased behaviour can sometimes be modulated by logical and evidential considerations (and not just by altering certain environmental contingencies) does not settle the issue of the representational format of the implicit attitudes themselves. Sensitivity to logical and evidential considerations falls short of establishing

[14] Tools from the Quadruple Process Model can also be used to weaken Mandelbaum's use of Briñol et al. (2009)'s results, which show argument quality to influence implicit attitudes. After presenting students with strong and weak arguments in favour of hiring more African American professors at universities, they were given an IAT test. The tests revealed that students who listened to the strong arguments showed more positive associations with black faces than those who were exposed to the weak arguments. Yet, these findings are, again, consistent with there being a regulatory role of non-associative factors over associations—especially the presence of the students' motivation to overcome their bias. It is revealing, in this sense, that the students who listened to the strong arguments were also told that "the integration policy was being considered for implementation at their own (vs. a remote) university and in the next academic year (vs. in 10 years)" (Briñol et al. 2009, 294).

the propositional representational structure of implicit attitudes.[15] We need to look elsewhere in our search for criteria that help decide between a Doxastic and an Associationist view of implicit attitudes. The place to look for, I shall argue next, is the world of social structures and structural injustice. We pick up features repeatedly linked together within our social environment, and the specific way in which we pick up such features in the case of implicit attitudes makes it more plausible to think about them as associations. Or so I shall argue next.

3. Bringing implicit attitudes back to the social sphere

If changes in implicit attitudes, as measured by standard psychometric tools, are not a reliable indicator of changes in real world behaviour, and if the mechanisms responsible for such behaviour are in all likelihood quite complex and heterogeneous, how are we to settle the issue of the representational structure of the implicit attitudes themselves? Perhaps we can't. Perhaps there is no conclusive way of determining the exact representational structure of these mental states (but see e.g. Huebner 2016). Perhaps, as Holroyd and Sweetman (2016) argue, there is no unitary cognitive phenomenon that falls under the concept of implicit attitude. Yet, on the idealized assumption that implicit attitudes are indeed a specific type of representation, the question of which hypothesis about its representational structure best explains the available evidence concerning its fundamental qualities still makes perfect sense. It still makes sense to ask, following Lipton (2004), which explanation is "the loveliest explanation"—the explanation that provides the most understanding of the phenomenon and is, for this reason, the likeliest explanation. The loveliest explanation of the representational structure of implicit attitudes is the explanation that best explains their *central features*. It would thus be a mistake—a methodological mistake with important consequences—to focus on a property of the phenomenon that appears to be rather peripheral: modulation by evidence.

A hypothesis that focuses on the influence of logical and evidential considerations on modulating implicitly biased behaviour fosters (even if it is not necessarily committed to) the thought that we could, through reflective deliberation alone, bring about social and self-reform in the face of prejudiced environments. Such a hypothesis promotes the sort of individualism that Sally Haslanger (2015) so insightfully isolates as promoting the concealment of structural elements of injustice and prejudice in our societies. It helps to shift the focus away from the most important property of the phenomenon: not just implicit attitudes' resilience to change in the face of evidence. Implicit attitudes share this type of resilience with other mental states that deserve properly the label of beliefs.[16] It helps to shift

[15] As I pointed out earlier (Section 0), my argument remains neutral as to the truth of a different conditional, namely, that if implicit attitudes are insensitive to logical and evidential considerations, then they are associations. A version of this conditional, which focuses of the alleged insensitivity of implicit attitudes to logical form, is defended by Madva (2016). This is also the conditional guiding most research in social psychology, especially anything involving Gawronski and collaborators.

[16] Those who endorse a doxastic view of implicit attitudes do acknowledge resistance, but typically explain it by appealing to a fragmented mind: explicit and implicit attitudes would be, according to them, stored separately, with little or no connection between storage modules. I cannot address here

the focus away from a type of resilience to change in the face of evidence that reflects and contributes to social injustice. Such stubbornness seems to arise from (i) the frequent gap between our implicit and explicit attitudes, and, also, and importantly, from (ii) the fact that the reproduction of socially and morally unjust structures often makes us more efficient epistemic agents. If we focus on these two uncontroversial properties of implicit attitudes, as they relate to their standard lack of malleability, the associative view becomes a much better hypothesis than the alternative doxastic view. I'll briefly address these two issues in turn.

About (i). Of course, our implicit and explicit attitudes may coincide. There are people who are openly racist, sexist or homophobic. And, of course, implicit and explicit attitudes can be related—with the strength of the relationship varying with context and subject matter. But what makes implicit attitudes theoretically interesting—and morally and politically damaging—is that they are often at odds with our explicit beliefs in such a way that we do not endorse their truth. It is thus difficult to change what we do not think we hold true. Furthermore, this feature has an important consequence that speaks in favour of an associationist view: we cannot *retract* from implicit attitudes. We cannot take back what we do not endorse to begin with. It is not just that implicit attitudes are regulated irrespective of truth-supporting considerations. This is also the case with some of our both conscious and unconscious beliefs. It is rather that implicit attitudes do not seem to be in the business of representing facts—they are not the kind of mental state that is constitutively subject to a norm of truth (e.g. Shah & Velleman 2005) or a norm of knowledge (see e.g. Williamson 2000).[17] They express, instead, an evaluative attitude toward their object.

About (ii). Indeed, implicit attitudes not only reflect entrenched social structures in which inequality and discrimination are pervasive, the fast and automatic ways in which they move us to act often grant us an epistemic advantage, especially under time constraints. We don't just represent genuine (unjust) regularities in our social environment, these representations also contribute to the perpetuation of such prevailing unjust features by often granting us epistemic gains.[18] Again, the presence of such epistemic benefits partially explains why it is so difficult to change implicit attitudes through rational considerations. At the same time, when what we want to explain is the very presence of such epistemic advantages, achieved in the face of our often contradictory explicitly held beliefs, a

---

the advantages and disadvantages of the fragmentation hypothesis as it applies to implicit attitudes. It seems to me, however, that fragmentation falls short of properly addressing the distinctive social role of implicit attitudes in light of their also being epistemically efficient representational structures (see below).

[17] For those who endorse a norm of truth, the thesis is that subjects can be said to believe that *p* only if they implicitly accept a normative judgment whose content is to believe that *p* only if *p* is true (e.g. Shah & Velleman 2005). Williamson (2000) defends instead that belief is constitutively regulated by the knowledge norm, i.e., one must: believe *p* only if one knows *p*.

[18] Gendler (2008b, 2011) is right in pointing out the difficult normative dilemma that this poses for us. The more we know about our social environment, the more likely it is that we act in biased and prejudiced ways. So we have to either ignore genuine socially relevant regularities, thus incurring into epistemic costs, or we have to deploy extra cognitive energy in suppressing and controlling the readily available information that we get from our social environment.

view of implicit attitudes as associative structures offers a straightforward, lovelier explanation than a doxastic view. Associationism allows us to draw a picture of the properties associated with a certain category such that even when not all them are properties that all its members possess, it is often epistemically advantageous to act as though they were. From an associative point of view, it is also easier to explain why we are capable of recognizing members of a category based on incomplete or distorted information and also capable of generalizing knowledge acquired about a certain category to members of a similar one. An associative view of implicit attitudes explains why it is so difficult to be both epistemically virtuous and epistemically efficient by characterizing them as evaluative structures with great developmental and evolutionary adaptability but little sensitivity to truth, hence their resilience to change through logical and evidential considerations.

Despite this resilience, implicit attitudes do change. Yet, when they do, it is usually through the acquisition of strategies to fight a biased social environment (see e.g. Devine et al. 2012)[19] and, above all, through changes in the very social structures that cause the biases to begin with. A classic example is Dasgupta and Asgari (2004)'s field study comparing the change in implicit attitudes with regard to gender and leadership of two groups of women after one year in college. The first group attended an all women's college, where most faculty and administrators were women, and was hence exposed to lots of counter-stereotypical female roles. The second group attended a co-educational school and was exposed to a comparatively lower number of women in leadership positions. Implicit gender bias was hugely reduced in the women who attended the all-women's college by their sophomore year; they were also faster at automatically associating women with counter-stereotypical qualities regarding leadership. Implicit gender bias in the women who attended the co-educational college was, by contrast, much stronger after the first year. Interestingly, neither group of women showed any change in their explicit beliefs about women vis-à-vis leadership. In both groups, women remain convinced that women were better carers than leaders.

This study illustrates how implicit attitudes' underground sovereignty over our egalitarian views is the result of our getting tuned, in fast, automatic and appropriately sensitive ways to prevailing features of our social environment and it is most efficiently changed through changes in such environment. I agree with Haslanger that the source of the problem with implicit attitudes is structural rather than individual, and that the analysis and the strategies to understand and fight against them cannot afford to ignore structural properties of our social environment.[20] Even so, as Haslanger herself suggests, just focusing on structural social factors would not provide the whole picture either. When the task is to offer the best possible explanation about the representational structure of implicit attitudes, the Associationist View seems better tuned to the anti-individualist approach Haslanger correctly demands. For associationism is wedded to the idea that changes in the thinking patterns that characterize implicit attitudes are mainly promoted by changes in experiential context, so that no change is likely in the absence of social change.

---

[19] From a philosophical point of view, this is also the strategy suggested by Madva (2016).

[20] Advocates of the doxastic view may very well agree with this. The disagreement comes in when assessing the impact of this social fact on the sort of considerations that lead to settle arguments about the representational format of implicit attitudes.

I have tried to clarify what the representational structure of implicit attitudes might be in light of the available evidence and given how we use the ordinary concept of implicit attitudes. My theoretical toolbox is not rooted in the classic terms of conceptual analysis. Even so, if, following Haslanger (2006), we were to distinguish between descriptive and ameliorative projects in philosophy, my approach in this paper primarily responds to a descriptive inquiry. Yet, if I am right that (i) occasional modulation by logical and rational considerations does not settle the issue of implicit attitudes' representational structure and (ii) that viewing implicit attitudes as associations best explains the fact that implicit attitudes change predominantly through changes in the social environment, then this associationist view also illustrates a scenario in which the descriptive and the ameliorative projects coincide. For such an associationist view also allows us to arrive at a concept of implicit attitudes that helps us to better pursue our goal of social justice by highlighting that it is largely, if not only, through changes in the social environment that we can modulate the pernicious influence of implicit attitudes in our lives.

REFERENCES

Bain, David. 2013. Pains that don't hurt. *Australasian Journal of Philosophy* 92(2): 1-16.

Banaji, Mahzarin R. 2001. Implicit attitudes can be measured. In Henry L. Roediger, James S. Nairne, Ian Neath, and Aimée M. Suprenant, eds., *The nature of remembering: Essays in Honor of Robert G. Crowder*, 117-149. Washington, DC: APA.

Bar-Anan, Yoav and Brian Nosek. 2014. A comparative investigation of seven indirect attitude measures. *Behavioural Research* 46: 668-688.

Beecher, Henry K. 1955. The powerful placebo. *Journal of the American Medical Association* 159(7): 1602-1606.

Blair, Irene, Jennifer E. Ma and Alison P. Lenton. 2001. Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *J. of Personality and Social Psychology* 81 (5): 828-841.

Briñol, Pablo, Richard E. Petty and Michele J. McCaslin. 2009. Changing attitudes on implicit versus explicit measures: What is the difference? In R. Petty, R. Fazio, and P. Briñol, eds., *Attitudes: Insights from the New Implicit Measures*, 285-326. New York: Psychology Press.

Brownstein, Michael. 2016a. Implicit bias. *The Stanford Encyclopedia of Philosophy*. Zalta, E.N. (Ed.) <http://plato.stanford.edu/archives/spr2016/entries/implicit-bias/>.

—. 2016b. Implicit bias, context, and character. In Michael Brownstein and Jennifer Saul, eds., *Implicit Bias and Philosophy. Volume II: Moral Responsibility, Structural Injustice, and Ethics* Oxford: OUP.

Brownstein, Michael and Alex Madva. 2012. Ethical automaticity. *Philosophy of the Social Sciences* 42(1): 67-97.

Calanchini, Jimmy and Jeffrey W. Sherman. 2013. Implicit attitudes reflect associative, non-associative, and non-attitudinal processes. *Social and Personality Psychology Compass* 7/9: 645-667.

Calanchini, Jimmy, Jeffrey W. Sherman, Karl Christoph Klauer and Calvin K. Lai 2014. Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin* 40(10): 1285-1296.

Carlsson, Rickard and Jens Agerström. 2016. A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology* 57: 278-287.

Dasgupta, Nilanjana and Shaki Asgari 2004. Seeing is believing. Exposure to counterstereotypical women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Psychology* 40(5): 642-658.

De Houwer, Jan 2014. A propositional model of implicit evaluation. *Social and Personality Psychology Compass* 8(7): 342-353.

Devine, Patricia G., Patrick S. Forscher, Anthony J. Austin and William T. L. Cox. 2012. Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology* 48: 1267-1278.

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

Evans, Jonathan St. B. T. and Keith E. Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8: 223-241.

Festinger, Leon 1957. *A Theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fireston, Chaz and Brian J. Scholl. 2016. Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, doi: 10.1017/S0140525X15000965, e229

Forscher, Patrick S., Calvin Lai, Jordan Axt, Charles R. Ebersole, Michelle Herman, Patricia G. Devine and Brian A. Nosek. 2016. A meta-analysis of change in implicit bias. Open Science Framework. December 9. osf.io/awz2p.

Gawronski, Bertram and Fritz Strack. 2004. On the propositional nature of cognitive consistency: Dissonance changes implicit but not explicit attitudes. *Journal of Experimental Social Psychology* 40: 535-542.

Gawronski, Bertram, Eva Walther, and Hartmut Blank. 2005. Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology* 41, 618-626.

Gawronski, Bertram, Roland Deutsch, Sawsan Mbirkou, Beate Seibt and Fritz Strack. 2008. When "Just Say No" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, 44: 370-377.

Gendler, Tamar Szabó. 2008a. Alief and belief. *The Journal of Philosophy* 105(10): 634-663.

—. 2008b. Alief in action (and reaction). *Mind and Language* 23(5): 552-585.

—. 2011. On the epistemic costs of implicit bias. *Philosophical Studies* 156: 33-63.

Greenwald, Anthony G., Debbie E. McGhee and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74: 1464-1480.

Greenwald, Anthony G, Brian A. Nosek, Mahzarin R. Banaji and Karl Christoph Klauer. 2005. Validity of the salience asymmetry interpretation of the IAT: Comment on Rothermund and Wentura. *Journal of Experimental Psychology: General* 134(3): 420-425.

Greenwald, Anthony G., T. Andrew Poehlman, Eric Luis Uhlmann and Mahzarin R. Banaji. 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97: 17-41.

Greenwald, Anthony G., Mahzarin R. Banaji and Brian A. Nosek. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553-561.

Haslanger, Sally. 2006. What good are our intuitions? Philosophical analysis and social kinds. *Proceedings of the Aristotelian Society*, Sup. Vol. 80(1): 89-118.

—. 2015. Social structure, narrative and explanation. *Canadian Journal of Philosophy* 45(1): 1-15.

Heider, Fritz. 1958) *The Psychology of Interpersonal Relations*. New York: Wiley.

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

Holroyd, Jules and Joseph Sweetman. 2016. The heterogeneity of implicit biases. In Michael Brownstein and Jennifer Saul, eds., *Implicit Bias and Philosophy. Volume I: Metaphysics and Epistemology*. Oxford: OUP.

Huebner, Bryce. 2016. Implicit bias, reinforcement learning, and scaffolded moral cognition. In Michael Brownstein and Jennifer Saul, eds., *Implicit Bias and Philosophy. Volume I: Metaphysics and Epistemology*. Oxford: OUP.

Hughes, Sean, Dermot Barnes-Holmes and Jan De Houwer. 2011. The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record* 61(3): 465-498.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Karpinski, Andrew and James L. Hilton. 2001. Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology* 81: 774-788.

Katz, Irwin and Glen R. Hass. 1988. Racial ambivalence and American value conflict: correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology* 55(6): 893-905.

Kawakami, Kerry, John F. Dovidio, Jasper Moll, Sander Hermsen and Abby Russin. 2000. Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology* 78: 871-888.

Levy, Neil. 2014. Neither fish nor fowl: implicit attitudes as patchy endorsements. *Noûs* 49(4): 800-823.

Lipton, Peter. 2004. *Inference to the Best Explanation*. London: Routledge.

Madva, Alex. 2012. *The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency*. PhD dissertation. Columbia University.

—. 2016. Why implicit attitudes are (probably) not beliefs. *Synthese* 193: 2659-2684.

Mandelbaum, Eric. 2013. Against alief. *Philosophical Studies.* 165:197-211.

—. 2016. Attitude, association, and inference: On the propositional structure of implicit bias. *Noûs* 50(3): 629-658.

Marks, Joel. 1982. A Theory of Emotion. *Philosophical Studies*, 42: 227-242.

Martínez, Manolo. 2011. Imperative content and the painfulness of pain. *Phenomenology and the Cognitive Sciences* 10(1): 67-90.

Mitchell, Chris J., Jan De Houwer and Peter F. Lovibond. 2009. The propositional nature of human associative learning. *Behavioral and Brain Sciences* 32(2): 183-198.

Nosek, Brian A. and Mahzarin R. Banaji. 2001. The go/no-go association task. *Social Cognition*, 19(6): 625-666.

Nussbaum, Martha. 2001. *Upheavals of Thought: The Intelligence of Emotions*, Cambridge: Cambridge University Press.

Oakley, Justin. 1992. *Morality and the Emotions*, London: Routledge and Kegan Paul.

Olson, Michael A. and Russell H. Fazio. 2004. Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology* 86: 653-667.

Ossipov, Michael H., Gregory O. Dussor and Frank Porreca 2010. Central modulation of pain. *Journal of Clinical Investigation* 120(11): 3779-3787.

Oswald, Frederick, Gregory Mitchell, Hart Blanton, and Philip E. Tetlock. 2013. Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology Studies* 105(2): 171-192.

Payne, B. Keith, Clara Michelle Cheng, Olesya Govorun and Brandon D. Stewart. 2005. An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89: 277-293.

Rothermund, Klaus and Dirk Wentura. 2004. Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General* 133: 139-165.

Rozin, Paul and April E. Fallon. 1987. A perspective on disgust. *Psychological Review* 94(1): 23-41.

Sechrist, Gretchen and Charles Stangor. 2001. Perceived consensus influences intergroup behavior and stereotype v accessibility. *Journal of Personality and Social Psychology* 80 (4): 645-654.

Shah, Nishi and David J. Velleman. 2005. Doxastic deliberation. *The Philosophical Review* 114 (4): 497-534.

Sherman, Jeffrey W., Bertram Gawronski, Karen Gonsalkorale, Kurt Hugenberg, Thomas J. Allen and Carla J. Groom. 2008. The self-regulation of automatic associations and behavioral impulses. *Psychological Review* 115(2), 314-335.

Sloman, Steven A. 1996. The empirical case for two systems of reasoning. *Psychological Bulletin*, (119): 3-22.
Smith, Angela M. 2005. Responsibility for attitudes: activity and passivity in mental life. *Ethics* 115(2): 236-271.
—. 2012. Attributability, answerability, and accountability: In defense of a unified account. *Ethics* 122(3): 575-589.
Williamson, Tim. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Josefa Toribio is an ICREA Research Professor at the University of Barcelona. She previously held positions at the University of Sussex, Washington University in St. Louis, the University of Indiana, Bloomington, and the University of Edinburgh. Her current research focuses on the analysis of central topics in the philosophy of mind and the philosophy of cognitive science, with a special emphasis on the philosophy of perception and rationally responsive unconscious mental states such as implicit attitudes.

Address: ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain. Universitat de Barcelona. Department of Philosophy, Montalegre 6, Barcelona 08001, Spain. E-mail: jtoribio@icrea.cat