sciendo

The Underlying Mechanisms of Implicit Bias

A Multifaceted Account

Sergio Cermeño-Aínsa
University of Girona (UdG)

doi: 10.2478/disp-2023-0012 BIBLID: [0873-626X (2023) 70; pp. 267–97]

Abstract

This paper delves into the dark side of the mind. I am interested in the underlying mechanism of implicit bias. Implicit biases are implicit attitudes that conflict with our informed beliefs, leading to prejudiced feelings, stereotyped thoughts and ultimately discriminatory behaviours. Despite receiving much attention in philosophy and psychology, the mechanism by which these biases are formed is controversial. In this paper, I review the proposed models—association-based, belief-based and mental imagery-based—and suggest that these models can be reconciled in a multifaceted account. My argument relies primarily on debiasing interventions. If implicit biases are triggered by only one mechanism, experimental manipulations based on other mechanisms will fail to modify implicitly biased behaviour. However, manipulations involving each of the proposed models successfully modify biased behaviour. Thus, implicit bias has a multifaceted character, *i.e.*, all three models have something to say in explaining the mechanisms hidden in the implicitly biased behaviour.

Keywords

analyticity; David Lewis; mereology; universalism; unrestricted composition

1 Introduction

In the late 1800s, Sigmund Freud propagated the idea that the unconscious mind—attitudes, feelings and thoughts of which we are unaware—can profoundly influence behaviour. Today, unconscious attitudes that trigger unintentional behaviour are called implicit attitudes. The term "attitude" is differently employed in psychology and philosophy.

In psychology, it typically refers to mental states that denote preferences (*e.g.*, like or dislike), while in philosophy, it refers to propositional mental states (*i.e.*, mental states that are thought to bear a relationship to a proposition).[1] There are explicit and implicit attitudes. Explicit attitudes are mental states accessible to introspection, can be verbalized, and do not influence behaviour automatically. In contrast, implicit attitudes are mental states opaque to introspection, cannot be verbalized, and influence behaviour automatically. This paper focuses on implicit bias, a particular form of implicit attitude.[2] An implicitly biased decision, action, or thought reflects implicit features of cognition that distort or influence subjects' behaviour.

For its frequency and social impact, implicit biases have received much attention. They comprise three open philosophical spaces. First, there are ethical implications, *e.g.*, consequences for our understanding of agency, responsibility, and how we ought to act (Holroyd [2012]). Second, there are epistemological worries (Gendler [2011]), *e.g.*, whether implicit bias gives rise to a new form of scepticism (Saul [2012]). And third, there are metaphysical issues, or questions about their nature and their underlying cognitive processes, *e.g.*, whether they operate in the absence of agent awareness (Holroyd [2015]) or what kind of mental representation is responsible for them (Gendler [2008a, 2008b], Mandelbaum [2016], Nanay [2021]). In this paper, I set aside the first concerns and focus on the third.[3]

People identify weapons much faster when primed with pictures of black faces compared to pictures of white faces (Payne [2001]); subjects with a strong implicit bias against African Americans tend to smile less and cut off conversations sooner when interacting with them (McConnell and Leibold [2001]); gender information can automatically influence judgement: when you think of a nurse, you think of a woman, when you think of

---

[1] Some philosophers defend the idea that attitudes in general and implicit attitudes, in particular, are not, properly speaking, mental states, but dispositional states: character traits (Machery [2017]) or situations (Payne *et al.* [2017]). I cannot see any inconsistency in claiming that implicit attitudes are dispositional states with a corresponding mental state. In this paper, I assume that implicit attitudes comprise mental states and focus my discussion on which kind of mental states they are.

[2] By implicit attitudes, I mean the broad class of attitudes that are opaque to introspection and escape direct control. I understand implicit bias as a subset of attitudes characterized by having prejudicial and stereotypical content. In contrast, many other implicit attitudes have innocuous content—an individual can hold a stereotype without having negative attitudes.

[3] For a philosophical introduction to implicit bias, see Brownstein and Saul [2016a, 2016b].

a mechanic, you think of a man (Banaji and Hardin [1996]); racial and ethnic minorities and women are subject to less accurate medical diagnoses, curtailed treatment options, less pain management, and worse clinical outcomes (Chapman *et al.* [2013]). Examples of this sort are countless and can affect many different social areas (racial, gender, age, weight, LGBTQ, disability and many more), take many different forms (perceptual, judgemental, emotional, motivational or action moved), and be caused by diverse mental tendencies (it can be driven by pattern-seeking, by the natural propensity to take cognitive shortcuts or by social and cultural influences). This versatility should lead us to think of implicit bias as formed by a heterogeneous rather than a singular structure, as a multifaceted rather than a unique mental state, and as a multidimensional rather than a specific process (Holroyd and Sweetman [2016], Johnson [2020]).

What kind of mechanism causes the biased behaviour? What happens between the information that triggers the bias and the biased behaviour? Literature usually takes two routes: it is an associative process (Gendler [2008a], [2008b], Madva [2016a]) or a reflective one (De Houwer [2014], Hughes *et al.* [2011], Mandelbaum [2016], Mitchell *et al.* [2009]). The associationist view posits that implicit bias is based on pure associations without involving reflective processes. The reflective view states that implicit bias is a cognitive process driven by reflective thinking rather than a pure associative one. Recently, some philosophers have proposed an alternative position, suggesting that the mechanism that triggers biased representations is driven by mental imagery (Nanay [2021], Welpinghus [2020], Sullivan-Bissett [2019]).

In this paper, I argue that it is possible to defend all three approaches at once.[4] I contend that associations, beliefs, and mental imagery are all factors that contribute to the formation of implicit biases, each present to varying degrees across different instances. I reached this conclusion by analysing the mechanisms employed to undo biased behaviour. The argument goes like this:

1. If biased behaviour is represented by pure association, then counterconditioning and extinction are the best procedures to disengage the bias.

2. If biased behaviour is represented by propositionally structured beliefs, then belief-based mechanisms are the best procedures to disengage the bias.

---

[4] One can also be sceptical about the very existence of implicit biases. I do not explore this possibility here; I assume that although implicit bias circulates in mysterious and uncertain terrain, their existence is unquestionable, and time makes them observable and predictable.

3. If biased behaviour is represented imagistically, then imagery-based mechanisms are the best procedures to disengage the bias.

4. Empirical findings show that all three debiasing mechanisms work, and when applied together, better and longer-lasting results are obtained.

Therefore, implicit bias is not induced by a single representational mechanism, but by many.

Defending this argument is the aim of this paper. Here is the plan. Firstly, I explain the three models (sect. 2): the association-based (sect. 2.1), the belief-based (sect. 2.2), and the imagery-based (sect. 2.3). Subsequently, I carve out an argument for compatibility (sect. 3). Some possible caveats are initially put in order (sect. 3.1). Next, I review some evidence on the interventions employed to modify implicit bias: the association-based interventions (sect. 3.2), the belief-based interventions (sect. 3.3) and the mental imagery-based interventions (sect. 3.4). Finally, I show that when all three models are taken together (or at least two of them) interventions to counteract implicit bias obtain better and more durable results (sect. 3.5). This evidence is, I conclude, enough to seriously contemplate the multifaceted account suggested here (sect. 4).

2 Implicit bias: mere associations, beliefs or mental imagery

What implicit biases are, how they are formed, or how they affect behaviour is a matter of practical and theoretical interest. Many prejudices, stereotypes and discriminative behaviours are substantiated because of them. The prevalence of unwanted behaviour across many areas of human life (such as health care, law enforcement, employment, criminal justice and education) suggests that mental processes outside one's conscious awareness influence behaviour. From a practical perspective, researchers are interested in looking for a solution to problems caused by unintentionally biased behaviour, *i.e.*, how to change the automatic processes that produce changes in the behaviour. From a theoretical perspective, the distinction between automatic and deliberative mental processes encourages researchers to theorize about the nature of the mechanism responsible for implicit pro-

cesses.[5]

There are two ways to pick out the mechanism causing these representations. The first way focuses on the kind of relationship between the content of the concepts, thoughts or perceptions that trigger biased behaviour. In particular, whether or not the implicit biased representation is sensitive to semantic content and logical form. On the one hand, an implicitly biased representation is sensitive to semantic content when the co-activation of concepts or thoughts that lead to a biased representation is mediated by the meanings of those concepts or thoughts; otherwise, it will not. On the other hand, an implicitly biased representation is sensitive to logical form when the co-activation of concepts or thoughts that lead to a biased representation is mediated by the logical structure of those concepts or thoughts (*e.g.*, operators like negation or conditional); otherwise, it will not. The second way focuses on the mechanism that modifies implicit behaviour. In this case, it is assumed that the type of manipulation employed to undo subjects' implicit behaviour offers clues about the type of representations underlying implicit bias. Let's examine the proposed models in detail.

2.1 Implicit bias as an association-based representation

According to the dominant view, implicit biases are associations that arise from subjects' learning history (Levy [2015: 803]); the learned association between concepts or between concepts and valences causes the activation of such associations. This is *pure associationism*. A basic form of association claims that the frequency with which an organism has come into contact with Xs and Ys in one's environment determines the frequency with which thoughts about Xs and Ys will arise together in the future. This coactivation operates in the absence of agent awareness and, for some, it brings out a particular psychological kind (Holroyd [2016: 154–5]).

Indeed, many social cognitive scientists consider implicit biases to be acquired through some form of associative learning whose repetition and robustness trigger an associative

[5] Dual-process theories are at the core of this distinction. According to dual-process theorists, an associative mechanism describes a non-reflective, fast, automatic and non-conscious process; conversely, when reflective, slow, deliberately processed and conscious, the process is non-associative. However, this may not be entirely accurate since some cognitive processes can be unconscious, fast and automatic without being associative (Mandelbaum [2016: 647]). To avoid confusion, this paper restricts the notion of "associative" to the kind of connections between concepts, thoughts or stimuli that produce representations that are specific habitual propensities emerging automatically from agents' learning history.

structure. Associations can be structured as stimulus-response (when the bell rings, the dog salivates) or as stimulus-stimulus (when the bell rings, food is present). In the latter case, learning to associate two stimuli involves the co-activation of two different representations. The modification of associative learning requires associating the conditioned stimulus with an unconditioned stimulus or with the absence of stimuli. The first is counterconditioning: when the bell rings (stimulus), instead of food (conditioned stimulus), an electric shock (unconditioned stimulus) is presented. The second is extinction: when the bell rings (stimulus), neither food (conditioned stimulus) nor any stimulus is presented. Both are different ways of inhibiting a response or undoing an association. Thus, taking the association-based approach involves that the best mechanism for changing biased behaviour is changing environmental contingencies. The best way to get rid of the association between BLACK MALE and DANGEROUS is by eliminating the contingencies produced by the association between BLACK MALE and DANGEROUS. The prevailing view then states that the acquisition of implicit attitudes is processed through conditioning and reinforcement and elimination through counterconditioning and extinction.

These sorts of associations are direct associations. Direct because the association transits directly between concepts or between thoughts, without other mental processes interfering in such transition. Unlike other mental mechanisms that require more steps to connect concepts or thoughts, direct associations involve only one step; one concept is directly and automatically associated with another concept to produce a biased representation. The following hypothesis supports direct associationism:

> (ASSOCIATION) Implicit bias is represented by some sort of associative structure that is sensitive neither to semantic content nor to the logical form of the agent's thoughts and perceptions and, therefore, the best way to counteract it is (assuming no influence of other factors) via counterconditioning or extinction.

According to (ASSOCIATION), implicit bias represents a genuine mental process that occurs without the intervention of any other mental process. No doubt, (ASSOCIATION) stands out by its simplicity since it postulates just one mental process—the ability to relate two concepts causally. In such cases, implicit biases are, in contrast to beliefs, insensitive

to both the semantic content and the logical form of the concepts or thoughts associated.[6] Thus, a direct association between *e.g.*, salt and pepper, is just that, an association and nothing else. There is no specific semantic content connecting the concepts—they are just paired. A representative example of direct associations is subliminal priming. When subjects are subliminally exposed to certain associations, the causal connection between concepts is established without the intervention of any other internal process, *i.e.*, the connection is reinforced independently of the semantic content and the logical form in which concepts are connected. These mental states do not, for many, fit into the most common psychological categories and, therefore, they could determine a particular psychological category.[7]

To support (ASSOCIATION), theorists often appeal to the mechanisms of modification of implicit bias. If the deactivation or reorientation of implicit bias is achieved via the dishabituation of the biased associations or the elimination of certain environmental contingencies (*i.e.*, via counterconditioning or extinction), then such associations must follow a mechanism like the one described in (ASSOCIATION). Much of this paper is devoted to analysing this way of thinking. The key point for now is to emphasize that this is a strategy that defenders of (ASSOCIATION) often resort to.

Accordingly, the (ASSOCIATION) defenders emphasize that the type of representation involved in implicit bias differs from that of beliefs and other mental states. Marking this difference is critical, and although (ASSOCIATION) theorists do not always agree on the factors that lead to such differentiation, they agree that implicit biases must constitute

---

[6] A further reason usually employed by the defenders of (ASSOCIATION) to distinguish direct associations from beliefs is that in contrast to the latter, the former does not respond to their truth-value; *i.e.*, while beliefs respond to changes in the truth value of their contents, direct associations only respond to changes in habit or conditioning (Gendler [2008b: 566]). For example, if I believe that P, I believe that it is true that P, my beliefs respond to the truth-value of their contents. But if I find reasons showing that P is not true, then I will stop believing P. In contrast, a direct association is not sensitive to the truth value of its contents, it just happens.

[7] Gendler [2008a], for example, calls these mental states aliefs. An alief is a mental state with three associatively linked components: a representational component ("the representation of some object or concept or situation or circumstance, perhaps propositionally, perhaps non-propositionally, perhaps conceptually, perhaps non-conceptually"), an affective component ("the experience of some affective or emotional state") and a behavioural component ("the readying of some motor routine") ([263–4]). The three components are co-activated by some features of the subject's environment, and its automatization makes this co-activation very difficult to break.

a particular kind of mental state.

2.2 Implicit bias as a belief-based representation

Just as (ASSOCIATION) requires a single step to connect concepts and draw up a representation that ultimately triggers a biased behaviour, other models suggest that implicit biases require further elaboration. Being automatic, unconscious or fast is not equivalent to being associative (see fn 5); logical processes can also operate automatically, fast and out of awareness, so to assess the nature of implicit bias, it is also possible to resort to propositional structures (Mandelbaum [2016]). The connection between two concepts may be, therefore, caused by some kind of unconscious propositionally structured belief. More than a direct association between BLACK MALE and DANGEROUS, implicit racists may represent an unconscious belief with the structure BLACK MALES ARE DANGEROUS. Let's call (BELIEF), to these occurrences (the doxastic model of implicit bias). (BELIEF) is sustained by the following hypothesis:

> (BELIEF) Implicit bias is represented by some sort of belief-based structure connected via propositional attitudes that are necessarily sensitive to both the semantic content and the logical form of the agent's thoughts and perceptions, and, therefore, the best way to counteract it is (assuming no influence of other factors) via logical and evidential interventions.

In a first approximation, (BELIEF) may sound like an over-intellectualization of the phenomenon. According to this view, BLACK MALE and DANGEROUS are connected in a way that involves some sort of unconscious reasoning. This unconscious reasoning hides a belief whose content is opposite to the content of the held belief (when consciously generated). Proponents of (BELIEF) have in mind a constrained notion of belief, a belief formed through a simple propositional form, barely elaborated but structured through some kind of logical form. In this sense, implicit biases are like imperfect cognitions, or better, like insufficiently elaborated reasoning. One of the most prominent advocates of this position is Eric Mandelbaum. Mandelbaum [2016] provides strong evidence supporting the idea that implicit biases are sensitive to the semantic content and the logical form of agents' thoughts and beliefs. For example, when subjects are prompted to dislike person A (by implicit clues), and they are told that person A dislikes person B, subjects tend to like person B. Subjects are internally biased by the propositional structure 'my enemy's enemy is my friend' (Mandelbaum [2016: 10-11], from Gawronski *et al.* [2005]). There is a logical form

in this biased movement since subjects automatically derive a positive from two negatives by managing a conditional structure; *i.e.*, there is a thought formed by two premises (X is my enemy, Y is X's enemy) and a conclusion (Y is my friend). But a pure associative account would predict an increased negative valence for Y, since rather than a conditional structure, it posits a mere association where valences are instead accumulative. This example clearly contradicts this view.[8]

This leads Mandelbaum to seriously consider (BELIEF) as the hidden mechanism of implicit bias. But central to his account is the idea that if implicit bias can be eliminated through logical or evidential factors, then the implicit attitudes responsible for biased behaviour cannot be mental states governed by a mere associative process but mental states following belief-structured rules. Once again, theorists turn to the mechanism that neutralizes biased behaviour to infer the mechanism by which it is activated.

Accordingly, (BELIEF) predicts a propositional structure for implicit bias. Implicit biases are propositions encoded in memory for subsequent unconscious and automatic activation when requested by the environmental situation. Considering this, some biased representations may be explained as unconscious propositionally structured beliefs.

2.3 Implicit bias as a mental imagery-based representation

Finally, other researchers have recently argued that the mental representation responsible for implicitly biased behaviours is neither a propositional attitude nor a mere association, but mental imagery. Mental imagery is defined as "perceptual processing that is not triggered by corresponding sensory stimulation" (Nanay [2021: 1]).[9] The connection between two concepts that trigger biased behaviour is, in this case, caused by some form of mental imagery. Let's name these occurrences (IMAGERY). (IMAGERY) is sustained by the

---

[8] For more examples, see Mandelbaum [2016]. For different interpretations of this and other examples, see Madva [2016a] and Toribio [2018]. For more recent literature showing the logical structure of implicit attitudes, see Kurdi and Dunham [2021], Kurdi *et al.* [2022] and Kurdi *et al.* [2023]. Kurdi and Dunham [2021], for example, analysed the sensitivity of implicit evaluations to logical structure. Through five experiments, these researchers found that implicit attitudes, just as explicit ones, are sensitive to logical structures (in particular, conditional statements), thereby supporting propositional theories. Thanks to an anonymous reviewer for bringing this literature to my attention.

[9] These mental states may be voluntary or not, conscious or not, action-oriented or not, and emotionally charged or not (see Nanay [2021: 8–9]). In the context of implicit bias, however, the mental imagery state is expected to be involuntary, unconscious, action-unoriented and emotionally charged.

following hypothesis:

> (IMAGERY) Implicit bias is represented by some sort of imagery-based structure connected via unconscious mental images that are sensitive to semantic content but insensitive to the logical form of the subject's thoughts and perceptions and, therefore, the best way to counteract it is (assuming no influence of other factors) by manipulating mental images.

In this case, implicit attitudes resemble perception more than previously thought—they are represented iconically. When someone implicitly connects BLACK MALE with DANGEROUS, there is an unconscious mental image of a BLACK MALE IN A THREATENING SITUATION (or similar) whose iconic representation produces a generalization that contradicts background beliefs. This suggests that implicit bias is only attainable when some form of perceptual (or *quasi*-perceptual) representation is involved. Implicit biases are, in sum, image-like representations.

The appeal to mental images is strongly based on their difference from beliefs and direct associations. Unlike beliefs, mental imagery is insensitive to logical form, and unlike direct associations is sensitive to semantic content (Nanay [2021], Sullivan-Bissett [2019]).[10] On the one hand, implicit biases are sensitive to semantic content. For example, when subjects affix a label of 'Sucrose' to one bottle and a label of 'Sodium cyanide' to another bottle, even knowing that both bottles contain sugar, they prefer the beverage labelled 'Sucrose' (Rozin *et al.* [1990], see also Mandelbaum [2013]). Here, there is a clear connection between the 'Sodium cyanide' label and the property of being dangerous, hardly explainable by content-insensitive associations. On the other hand, implicit biases are insensitive to the logical form. Following the sugar example, when the labels attached are 'not poison' and 'safe', subjects are reluctant to take the bottle labelled as 'not poison' even when they know that both bottles contain sugar. In this case, the information is not only processed in a way that is sensitive to the semantic contents of the representations but also insensitive to the logical form; particularly, this content is blind to the negation

[10] Though I take the Nanay [2021] and Sullivan-Bisset [2019] accounts as parallel, this is not exactly true. For instance, whereas for the former implicit biases are not propositionally structured, the latter leaves this possibility open.

(see Rozin *et al.* [1990], see also Levy [2015]).[11] So, implicit biases are represented neither by direct associations nor by beliefs, but by something in between.

Finally, the supporters of (IMAGERY) also appeal to the interventions to neutralize biased behaviour. Again, if implicit bias can be modified by manipulating subjects' mental imagery, then this must be the mental state at stake in the acquisition, triggering and elimination of implicit bias (Nanay [2021], Sullivan-Bissett [2019]).

Accordingly, (IMAGERY) calls for an imagistic structure of implicit bias. Since biased representations are sensitive to semantic content but insensitive to logical form and can be eliminated via mental imagery interventions, they must be represented in an unconscious mental imagery way. Implicit bias can plausibly be explained as mental imagery-based.

3 Implicit biases as a combination of ASSOCIATION, BELIEF and IMAGERY

The theoretical discussion does not settle the debate on the mechanism behind implicit biases. Each option provides strong arguments and compelling examples to support their position. However, none provide a clear and decisive argument as to whether implicit bias is or is not sensitive to the semantic content and the logical form of the agent's thoughts and perceptions. At this point, the discussion is dangerously close to a dead end—there seems to be a dialectical dispute without clear progress at first glance.

All the proposals also agree on using the mechanism to counteract implicit bias as an argument to settle the issue. If the debiasing mechanism is a counter-representation of the activation and acquisition mechanisms, then there are strong reasons to analyse these interventions deeply. Therefore, whether the debiasing mechanism is driven by counter-conditioning, beliefs, or mental imagery is an empirical question that must be carefully addressed.

This reasoning pivots on the claim that implicit biases are acquired, activated and extinguished through similar mechanisms. In the following section, I address this issue and

[11] Levy [2015] offers this experiment as "one of a number of experiments that seem to demonstrate that non-conscious processes are blind to negation" [815]. Levy's account postulates a new mental state between beliefs and associations. He suggests that implicit attitudes are patchy endorsements: they are more than associations since they are sensitive to semantic content but not sensitive enough to be qualified as beliefs. Like Sullivan-Bisset, Nanay thinks Levy is too permissive with the propositional account (see Nanay [2021: 14–5]).

clarify why this assertion should be accepted. For the time being, my general suggestion is first and foremost inclusive; I contend that implicit bias is a heterogeneous and multifaceted mental state which can be acquired, activated and eliminated through various means. That is, (ASSOCIATION), (BELIEF) and (IMAGERY) can all provide valuable insights into the nature of implicit biases. The rationale behind this perspective is not only that the debiasing interventions succeed when each of the three mechanisms are involved, but also that their combination yields much better results. Thus, I propose that implicit bias is not represented in a single but in a multifaceted way.[12] The multifaceted account of implicit bias can be articulated as follows:

> (MULTIFACETED ACCOUNT) Implicit bias is represented by some sort of mental structure connected through minimal associations, unconscious beliefs and/or unconscious mental imagery and, therefore, it can be counteracted via counterconditioning, belief interventions and/or mental imagery interventions.

(ASSOCIATION), (BELIEF) and (IMAGERY) constitute the (MULTIFACETED ACCOUNT). That is, each of them provides a partial explanation of implicit bias. The idea is that implicit attitudes are multi-representational; the same input can be biasedly represented in multiple ways. The system, therefore, admits the storage of three distinct representations for the same object and these representations are supported by distinct mental and neural structures. The multi-representational view explains why debiasing interventions sometimes succeed or fail and why the multi-intervention works better and achieves

---

[12] Some authors also suggest that implicit bias does not originate from a single mental state. Del Pinal and Spaulding [2018], for example, point out that implicit bias can be encoded in various ways; through salient-statistical associations (associations) or dependency networks (beliefs). These accounts offer interesting clues as to how the beliefs underlying implicit biases can be represented following different logical forms. However, unlike the multifaceted account proposed here, they hold that each bias can only be represented through a single mechanism at a time. Holroyd and Sweetman [2016] also argue for important differences between implicit biases originating from gender, race, or ethnic groups, thus proposing that implicit biases are heterogeneous. However, these authors treat heterogeneity in a different way than I do. Just as they focus on their complexity and functional variability, I focus on the mechanisms from which implicit biases may be triggered. Finally, Byrd [2019] considers more comprehensive views, seeing implicit bias as the product of different interventions of associations or reflections, thus considering the debate on the nature of implicit bias in terms of associationism vs non-associationism. However, to my knowledge, no one has considered the combination of the three models as suggested here.

more enduring results. Perhaps by employing a single debiasing intervention (regardless of which one it is), investigators are breaking down one type of representation, keeping the others intact. For example, when researchers use counterconditioning to extinguish a biased behaviour, they are deactivating the representation that associates BLACK MALE with DANGEROUS, but the unconscious belief that represents BLACK MALES ARE DANGEROUS, or the unconscious mental imagery that pictorially represents A BLACK MAN IN A THREATENING SITUATION remain unaffected. The same pattern occurs when researchers employ other interventions. This explains why interventions that include the deactivation of all three representations tend to yield better and more durable results. This is likely not accidental.

I plan to show that if the extinction, mitigation or mere modification of the implicitly biased behaviour can follow the three different routes, counterconditioning (or extinction), belief-based and imagery-based interventions, then all three mechanisms (and not only one) may trigger the representation that causes the biased behaviour. This is a substantial claim, as supporters of a monolithic account generally arrogate the idea that manipulating bias requires the intervention of their preferred mechanism. For example, Gendler [2008b] claims that one way to redraw the implicit biased association is by "cultivating norm-concordant habits through actual rehearsal" [572], that is to say, by making "a conscious effort to behave in the ways that our commitments dictate, so that these patterns of behaviour become familiar and habitual" [573]. Mandelbaum [2016] also writes that "[i]f there are interventions that reliably work to counteract implicit bias and if these interventions are not reducible to extinction or counterconditioning, then we have evidence that the structure of implicit bias is not, after all, underwritten by associations" [637], and provides empirical data suggesting that "in order to explain the data, we need logical, propositional processes—inferences and not just associative transitions—working over propositional structures" [640]. Finally, Nanay [2021] also claims that "[a]mong the most efficient ways of manipulating implicit bias is with the help of manipulating mental imagery. For example, visualizing or putting ourselves imaginatively in the shoes of a member of another racial or gender group can reduce implicit bias significantly. And, crucially, the extent of this reduction correlates with the details and vividness of the imagery involved" [15].[13] In what follows, I show that these attempts to get this evidence to their

[13] Similarly, Sullivan-Bisset [2019: 643] claims that "[t]he imagination model can accommodate empirical findings regarding the effects of certain interventions on bias, and it can also speak to why such effects are short-lived".

side denote certain biases in the mentioned literature. In my view, these theorists show an exacerbated interest in bringing the evidence to their side; a more objective look at the literature indicates, however, that the mechanism at play takes, instead, a multifaceted form. The remainder of this paper explores this possibility in-depth. Before, I will examine some possible difficulties that my plot can encounter along the way.

3.1 Some difficulties and caveats to the plan of action

As a preliminary caveat, despite the proliferation of methods to reduce implicit bias, very little is known about their relative effectiveness. The persistence of these interventions over time is, for example, questioned. Although some studies have reported long-term reductions (see Devine *et al.* [2012]), others note that many of these reductions fade off relatively quickly (Lai *et al.* [2016]). Sometimes, just knowing about the presence of implicit bias and its potentially harmful effects on behaviour may prompt individuals to pursue corrective action (Green *et al.* [2007]), and sometimes encouraging people to avoid implicit stereotyping increases their bias (Payne *et al.* [2002]). Many studies report poor results in mitigating biased behaviour even when the results are evaluated in the short term or immediately after the intervention (FitzGerald *et al.* [2019]). There are, however, no theories explaining this disparity in debiasing interventions.[14] So, a more suitable explanation of the psychological background underlying implicit bias becomes a necessary instrument for the designing of future interventions.

The suggestion of this paper is clear: if all three mechanisms work to eliminate (or at least attenuate) the biased behaviour, then all three may be involved in the acquisition and activation of the biased representation. There are four considerations to point out here. First, one can be subtle and argue that the elimination mechanism does not necessarily coincide with the acquisition or activation mechanisms. An implicitly biased behaviour might be acquired, for example, through a direct associative route, but in turn, be activated by an imaginative-based mechanism and mitigated by a belief-based mechanism.

[14] In a recent meta-analysis review, Paluck *et al.* [2021] report several inconsistencies in most studies on prejudice reduction. They found significant replication failures—when the study size increases, the intervention effect drops critically—and signs of publication bias—prejudice reduction studies with favourable (statistically significant) results see the light more readily than those with unfavourable or ambiguous effects. These inconsistencies show that if we are to develop an adequate model of how implicit biases are acquired, maintained, triggered, and ultimately modified, much more theoretical and practical research is needed. The present paper aims to contribute new insights to develop such a comprehensive model.

However, this option strengthens the multifaceted account. Indeed, just as proponents of a monolithic account require that only one mechanism rules the acquisition, activation and elimination of implicit bias, the multifaceted account is consistent with the idea that acquisition, activation and elimination can follow different routes since a piece of each of the three mechanisms may be constitutive of the implicitly biased behaviour.

Secondly, one can also be methodologically suspicious as to whether changes in implicit measures in the lab lead to changes in explicit measures or behaviour under real conditions (Forscher *et al.* [2019]). Indeed, one must be cautious with experimental results; it is not always suitable to extrapolate what happens in the lab to real-world situations. The artificiality of some interventions and their replicability in the real world is, in fact, a constant point of contention in sociological and psychological studies: Can we endorse a particular model of implicit bias only from lab outcomes? Again, this would be relevant if we aim to endorse one model at the expense of repudiating the others. But my suggestion is inclusive; implicit bias may be caused by a multiplicity of different factors, represented in different modes, and modified by different mechanisms. Arguably, such inclusivity softens this concern. Furthermore, there are systematic reviews (see, for example, Fitzgerald *et al.* [2019]) focusing exclusively on interventions which are relatively easy to apply in real-life scenarios. These reviews collect data from very diverse areas (race, gender, sexual orientation, nationality, socio-economic status, or age) and apply to very different fields (social psychology, medical ethics, health psychology, neuroscience, education, death studies, LGBT studies, gerontology, counselling, mental health, professional ethics, religious studies, disability studies or obesity studies).

Thirdly, a word about the measurement instruments. To date, the most widely recognized measure of implicit biases remains the Implicit Association Test (IAT; Greenwald *et al.* [1998]). In a typical race IAT, participants are asked to categorize Black faces and positive words with one key and White faces and negative words with another key, and then the task is reversed: White faces and positive words and Black faces and negative words. The difference in average response time between the two conditions indicates the implicit preference for one category over the other. Recently, the IAT has undergone scrutiny. Some argue that it correlates poorly with other tests (Mandelbaum [2016: 631]), others that it shows poor test-retest reliability (Bar-Anan and Nosek [2014]), and others question its predictive validity (Oswald *et al.* [2013]). Despite this, the IAT is widely accepted as a measure of implicit bias, so I will not enter into these disputes. I assume IAT's general validity.

Lastly, one can be suspicious about the actual source of biased behaviours (whether

its roots are individual or collective) and consequently question the use of individualized interventions. Some theorists think that biases originate mainly from specific social structures, such as mass media, and interventions should focus on these structures rather than individuals. According to Haslanger [2015], institutional problems demand institutional solutions. I agree—a profound and serious re-evaluation of our social structures would be necessary to reverse discriminatory behaviours—, but this is not enough to abandon individualistic interventions. A thorough dialogue between collective and individualistic dimensions would be the most suitable approach to neutralize stigmatized and discriminatory behaviours. After all, there are no discriminatory behaviours far from the institutional jaws, but neither from the individuals who perpetuate them (for a profound analysis, see Madva [2016b]). So, I assume here that the individualistic approach to promoting debiasing interventions is, at least, part of the equation.

That said, and aware that these concerns call for a more exhaustive analysis, I explore the literature on debiasing interventions.[15] Among the varied interventions used to counteract implicit bias, I focus on those that I believe are most identifying of the models discussed here. Exposure to counter-stereotypical exemplars is, for example, the most typical intervention of direct association models since it appeals to the dishabituation through counterconditioning of the biased behaviour. Appealing to egalitarian values requires internal assessment and reasoning, making it a belief-based intervention. Lastly, imagining contact with the outgroup, which requires the pictorial representation of a specific experience, is a mental imagery-based intervention.

3.2 Implicit bias modified by pure counterconditioning

Reorienting our biased habits to put them in accordance with our reflective commitments requires, according to (ASSOCIATION), following a process of dishabituation. One way to modify implicit bias via counterconditioning mechanisms is through exposure to counter-stereotypical exemplars—the repeated exposure to instances that contradict biased associations. Given the biased association between A and B, the extinction would involve exposure to A and non-B or non-A and B. In this case, breaking the associative

[15] Much of the literature cited in this paper is taken from the systematic review conducted by Fitzgerald *et al.* [2019]. I rely on this work for two reasons. Firstly, because these interventions only include studies that apply to real-world contexts. Secondly, because it is subject to a very strict selection process (only 30 articles from 1931 initially reviewed are included).

process between categories and valences (positive or negative) that trigger the biased behaviour would depend on direct or subliminal exposure. Exposure to counter-stereotypical exemplars (exemplars that contradict the stereotype of the outgroup) is, in fact, one of the most promising techniques to counteract implicit bias, at least in the short-term overturning (Fitzgerald *et al.* [2019]).

Dasgupta and Greenwald [2001] provide one of the most cited examples of debiasing via counterconditioning. They tested whether repeated exposure to images of famous and admired African Americans as well as infamous and disliked European Americans moderate automatic racial attitudes. Researchers exposed participants to pictures of either admired Black and disliked White individuals (pro-Black condition), disliked Black and admired White individuals (pro-White condition), or non-racial exemplars (control condition). Participants completed an IAT immediately after exemplar exposure and 24 hours later. Results revealed that exposure to admired Black and disliked White exemplars significantly weakened automatic pro-White attitudes (see also Columb and Plant [2010]).[16] Joy-Gaba and Nosek [2010] replicated this study but reported weaker results. Like Dasgupta and Greenwald [2001], they assumed that exposure to admired Black individuals was the critical component of the manipulation. However, their findings indicated that when the positive exposure—specifically, exposure to admired Black exemplars— was isolated, the effects were substantially smaller. In contrast, focusing on exposure to negative White exemplars resulted in much stronger effects. They concluded that including negative exposure is a crucial condition for reliably shifting implicit bias. Leaving aside the ethical considerations involved in negative manipulations, it seems that more than the exposure to admired Blacks, the effects are significant due to the exposure to disliked Whites. More recently, Kurdi *et al.* [2023] replicated the original study of Dasgupta and Greenwald [2001] and found that exposure to counter-stereotypic exemplars (admired Black and disliked White) did not significantly change implicit racial evaluations. But when these researchers introduced contingency awareness—*i.e.*, becoming consciously aware of the association between the conditioned stimulus and the unconditioned stimulus— they found significant reductions in implicitly biased behaviour.

[16] Similar results were obtained when assessing gender and sexual orientation biases. Dasgupta and Asgari [2004] found that people were more likely to associate women with leadership after seeing admired female leaders without negative comparison exposure, and Dasgupta and Rivera [2008] found that people with low contact with gays and lesbians displayed more implicit antigay attitudes than those with high contact.

Crucially, they leave open the question of whether these results are more in line with associative or propositional perspectives, as we cannot be sure that inferential reasoning plays any role in these interventions. They conclude that implicit evaluations are malleable in response to a wide range of interventions—very close to the multifaceted account defended here.[17] McGrane and White [2007] examined the generalizability of these results to other contexts. Researchers assessed the explicit and implicit prejudice of both Anglo and Asian Australians. Using the exemplar exposure task, they found significant results in the attenuation of implicit prejudice in minority groups (Asian Australians) but not in majority groups (Anglo Australians). They conclude that the differences in the groups evaluated appear determinant; some interracial dynamics (*e.g.*, White and Black Americans) are reasonably unique, but when the intervention involves less salient inter-racial exemplars, other dynamics emerge.[18]

In these cases, the biased representation is deactivated through counterconditioning or dishabituation; only exposure to exemplars contradicting the biased behaviour is enough to moderate implicit biases. The implication of logical reasoning or mental imagery is, at best, limited. Although the strength of the reduction depends on varied factors, this literature shows that counterconditioning interventions work. So, if implicit bias can be attenuated via exposure to counter-stereotypical scenarios, then (ASSOCIATION) describes a plausible mechanism for generating and activating implicit bias. Therefore, there should be some kind of representation that associates the thoughts causing the biased behaviour in the way described by (ASSOCIATION).

3.3 Implicit bias modified by logical reasoning

What about debiasing through logical reasoning? As I pointed out earlier, Mandelbaum [2016] has employed this strategy to argue for (BELIEF). He presents evidence showing that implicitly biased behaviour is sensitive to rational interventions and resolves that the belief-based model underpins implicit bias. However, the success of some belief-based interventions is consistent with the idea that implicit attitudes are (at least occasionally) geared by direct associations—as we have just seen, debiasing via countercon-

[17] Thanks to an anonymous reviewer for bringing this literature to my knowledge.

[18] Of course, this review does not cover all the existing literature on debiasing via counterconditioning. For reviews, see FitzGerald *et al.* [2019], Forscher *et al.* [2016] and Lai *et al.* [2014]; for further explanation of debiasing experiments employing counterconditioning, see Madva [2017] and Byrd [2021].

ditioning works. So, it is wise to move away from such a strong conditional and temper the rationale; if verbal and logical training mitigates implicit bias, then implicit bias must be constituted, at least in part, by a certain kind of propositional structure. (BELIEF) can be supported when counteracting interventions involve appealing to egalitarian values, identifying the self with the outgroup, or through some forms of evaluative conditioning. All these interventions may be clustered as intentional strategies since some kind of internal assessment is required. Here, I focus on appealing to egalitarian values, perhaps the most distinctive reasoning-based intervention.

Blincoe and Harris [2009] evaluated prejudice-reduction interventions by comparing the relative effectiveness of concepts like cooperation, political tolerance or respect with a control concept (intelligence). Participants read paragraphs primed with the three concepts and subsequently were evaluated in the Race IAT. Participants in the cooperation condition scored significantly lower than those in the control condition, thus showing that appealing to egalitarian values is a fruitful way to reduce implicit bias. Similarly, Clobert *et al.* [2015] evaluated the impact of Buddhist concepts (*e.g.*, Buddha, Dharma, Sutra) on pro-social attitudes compared with non-religious concepts (*e.g.*, sun, flower, freedom) and other religious (Christian) concepts (*e.g.*, Christian, Mary, Mathew). Results showed that implicit activation of Buddhist concepts comparatively to Christian primes and neutral conditions increase pro-sociality and decrease prejudice. Castillo *et al.* [2007] studied the influence of multicultural training on implicit racial prejudice. After completing the Race IAT, some participants enrolled in multicultural counselling courses 3 hours per week for 15 weeks, while others enrolled in counselling foundations classes (the comparison condition). After multicultural counselling courses, the Race IAT showed decreasing implicit racial prejudice from the participants enrolled in the multicultural counselling course but not from those enrolled in the comparison condition. Accordingly, one way of reducing racial tension and perhaps other non-racial bias is by promoting nonprejudiced beliefs or egalitarian values. More recently, Kurdi and Durham [2021] analysed in three studies the success of reinterpretation in modifying implicit behaviour; in particular, they evaluated whether an 8-minute podcast can soften people's implicit opinion against rhino hunting. Notably, people have, in general, a strongly negative impression of these practises, both explicit and implicit. The podcast provides general information in favour of these practices, such as that only post-reproductive rhinos are hunted, that these rhinos hurt and even kill conspecifics and that the funds go to wildlife conservation ([689]). The researchers found significant positive shifts towards the rhino hunters, thus showing that reinterpreting the information leads to changes in both explicit and implicit behaviours. This study

provides clear evidence towards propositional models in modifying implicit attitudes (see also Mann and Ferguson [2017]).

This literature shows that logical reasoning interventions are effective; implicit bias can be modulated not only through counterconditioning but also through logical reasoning. Appealing to egalitarian values works, and it is mostly driven by logical reasoning; the implication of counterconditioning and mental imagery seems to be here too scarce, if not null. Furthermore, the most recent literature (Kurdi and Durham [2021]) demonstrates changes in implicit attitudes through reinterpretation (listening to a podcast providing counter-attitudinal information), even for highly negative targets. Thus, if implicit bias can be modified by appealing to logical reasoning, then (BELIEF) describes a plausible mechanism for representing the generation and activation of implicit bias. Implicit bias can, therefore, be represented in an unconscious propositional way.

3.4 Implicit bias modified by mental imagery

Finally, implicit bias can also be alleviated through interventions involving mental imagery mechanisms. Two claims sustain these interventions: first, that mental imagery elicits emotional and motivational responses similar to real experiences, and second, that mental imagery shares the same neurological basis and employs identical mechanisms (memory, emotion, and motor control) as perception (Kosslyn, Ganis and Thompson [2001]). Indeed, for some, one of the most efficient ways of counteracting implicit bias is by manipulating mental imagery (Nanay [2021]). These interventions often involve encouraging subjects to imagine contact with members of the outgroup or imagine themselves as part of the outgroup.

Blair, Ma and Lenton [2001] investigated the effect of mental imagery on moderating implicit bias. In one of five experiments, participants were divided into the counter-stereotype and neutral (control) groups. In the first, participants were asked to imagine a strong woman; in the second, to imagine a holiday. Researchers found a significantly lower level of implicit stereotypes in the counterstereotype group than in the neutral group (Blair *et al*. [2001: 831]). They conclude that as imagining a counter-stereotypical exemplar reduces the implicit stereotype, mental imagery can have a powerful effect on implicit processes. Turner and Crisp [2010] also investigated the effects of imagining intergroup contact in reducing implicit prejudice in two different fields: age and religion. They found that subjects who imagined contact with an elderly person showed less implicit bias in favour of the young over the elderly compared to participants in the control condition (study

1). Similarly, they found that non-Muslim subjects who imagined contact with a Muslim showed significantly less implicit bias in favour of Muslims compared to participants in a control condition (study 2) (for a meta-analysis of the imagined contact hypothesis, see Miles and Crisp [2014]).

Finally, one of the most promising techniques to mitigate implicit bias is through immersion in virtual reality environments.[19] For example, Peck *et al.* [2013] investigated whether body representation can change implicit attitudes by employing immersive virtual reality to induce illusions of ownership over a virtual body. Participants wore a head-mounted display, through which they would see a programmed virtual body that substituted their real body and a body-tracking suit that would mimic the movements of their real body (their virtual body would move in the same way). Subjects were arbitrarily distributed in four conditions (embodied-light, embodied-dark, non-embodied-dark or embodied-alien) and were administered a race IAT before and just after the experiment. The results showed a significant change in implicit attitudes. When participants were embodied in dark-skinned bodies, they had less implicit race bias than those embodied in light-skinned bodies and those not embodied at all. They conclude that putting yourself in the skin of a black avatar reduces implicit racial bias (see also Chen *et al.* [2021]). Also employing immersive virtual reality, Thériault *et al.* [2021] studied whether embodied perspective-taking is better at reducing implicit prejudice and enhancing empathy with the outgroup than mental perspective-taking. Just as mental perspective-taking involves an intentional process that requires remarkable effort for subjects to imagine themselves in the other's shoes, embodied perspective-taking leverages virtual reality to directly enter a visuospatial perspective of the other in a much more embodied and automatic way. Indeed, subjects get into the other's senses with little conscious effort. As expected, they found a much more significant reduction of prejudice in embodied perspective-taking in contrast to mental perspective-taking.[20]

Arguably, the predominating mechanism here is mental imagery; the implication of other mechanisms is, at best, limited. Thus, since mental imagery-based interventions are also fruitful, implicit bias can be modulated not only by counterconditioning and logical reasoning but also by mental imagery-based mechanisms. So, once again, if implicit bias

---

[19] For a recent defence of (IMAGERY) considering work on virtual reality immersion, see (Sullivan-Bissett [2023]).

[20] For reviews, see Pelletier *et al.* [2020] and Higgins *et al.* [2024].

can be mitigated by appealing to mental imagery mechanisms, then (IMAGERY) also describes a likely mechanism for generating and activating implicit bias. The nature of the representation that triggers the biased behaviour may, therefore, be unconscious mental imagery.

3.5 The overlapping signature of debiasing interventions

So far, I have clustered interventions to modify implicit bias into three groups: direct association, belief, and mental imagery. However, this does not seem to capture the authentic mental nature of most interventions; indeed, the interventions designed to attenuate implicit bias rarely invoke a single mental mechanism. Although interventions are often designed to break a particular biased representation, most of them affect different representations simultaneously. Counterconditioning, for example, can be reinforced by adding mental imagery, thereby breaking both the direct association representation and the mental imagery representation (Blair *et al.* [2001]). Indeed, just as the success of the exposure to counter-stereotypical exemplars (or situations) is bounded by the stimulus presentation, imagined exemplars (and situations) might evoke mental representations that go beyond mere exposure. The possibility of imagining contact with the outgroup or engaging in the other's perspective through imagination is endless; not only does it evoke features similar to real experience (concrete details, causal sequences, logical constraints, concomitant emotional arousal, and similar neurological characteristics) but it also increases the "accessibility of related cognitive, emotional, and behavioural representations" (Blair *et al.* [2001: 829]).

Some interventions can also be driven by associations or propositional processes, yielding significant results in both cases. Evaluative conditioning, for example, is the process by which pairing an attitude object with another valenced attitude object shifts the attitudes of the first object in the direction of the valenced object (for a review, see Jones *et al.* [2010]; for a meta-analysis, see Hofmann *et al.* [2010]). It seems that evaluative conditioning may exploit the associative and propositional sides of the story. De Houwer [2011], for example, suggests that evaluative conditioning may be the product of different mental processes. When conducted by propositional processes, it depends on awareness, unconditioned stimulus revaluation, conditioned stimulus post-exposures, attention, and instructions (De Houwer [2018]). But in cases where it is conducted independent of these preconditions, it might be driven by pure automatic (conditioning) processes such as the formation of holistic representations (see also Gast and De Houwer [2013]).

Other interventions use the identification of the self with the outgroup. Brannon and Walton [2013], for example, induced cues of social connectedness toward members of a stigmatized social group and found that these cues spark interest in the outgroup culture, and such interest contributes to reducing implicit prejudice against the outgroup. Similarly, Gündemir *et al.* [2014] showed that introducing a dual identity can help suppress the implicit pro-White leadership bias and allow the higher-level leadership possibilities of ethnic minorities by reducing discriminatory behaviour during promotion-related decision-making processes. Since identifying the self with the outgroup involves putting oneself in the other's shoes (mental imagery), pairing racial stimuli and valences (conditioning or counterconditioning), and the deliberative processing of conscious representations (beliefs), these interventions must be mediated, in more or less extent, by the three mechanisms.

This evidence no doubt promotes the multifaceted account defended here. Likely, when all the models deploy at once, the debiasing results in a more robust and persisting reduction over time. Consider Devine *et al.* [2012], who employed a cluster of varied strategies (Stereotype replacement, Counter-stereotypic imaging, Individuation, Perspective-taking and Increasing opportunities for contact) and achieved bias reductions 8 weeks after the interventions. Researchers taught participants these five strategies to employ in daily life to reduce their racial biases. The point is that these strategies involve all the models at stake: counterconditioning, reflection and imagination. For example, counter-stereotypic imaging, increasing opportunities for contact or perspective-taking, involves conditioning or counterconditioning negative associations, but also imagining oneself in the shoes of the other. And all these interventions, along with stereotyping replacement and individuation, involve the identification of the stereotypes, the identification of the counter-stereotype corresponding to a stereotype, focusing on individual instead of collective features, or imagining oneself interacting positively with negatively stereotyped people, which undoubtedly are representations that require conscious deliberation, and therefore, reflective processing (see Byrd [2021]).

4 Conclusion

With the empirical evidence on the table, we are ready to determine the inclusive view proposed here. The reasoning is clear: if implicit bias can be manipulated via counterconditioning, beliefs and mental imagery, and if taken together, the interventions yield much better and enduring results, then the mechanism behind implicit bias might be the prod-

uct of a combination of (ASSOCIATION), (BELIEF) and (IMAGERY). That is, according to the current evidence, the multifaceted is the most plausible account; implicit bias is not committed to any particular cognitive model but many. What should future research consider in light of these findings?

This discussion should encourage social psychology researchers to address implicit bias by designing debiasing interventions encompassing the three models. After all, when all the models go into action, the debiasing effects are stronger and long-lasting (Devine *et al*. [2012]). Even more, by adding or removing the effects exerted by the different models and measuring the effectiveness of each model alone or in conjunction with the other models, these convergent interventions would have the power to determine the strength of each model in different situations.

To conclude, implicit bias is not a unitary notion but a disparate one, and as such, it could be explained by different mechanisms. Most likely, our minds choose one model instead of the others depending on multiple factors. Perhaps the same implicit biased behaviour is conveyed by one or another type of representation depending on the particularity of the situations with which subjects confront in their everyday lives. The extent to which each model is appropriate to account for implicit bias depends on personal and situational factors, but ultimately, all the models proposed implicitly trigger biased behaviours.

Sergio Cermeño-Aínsa
University of Girona (UdG)
Departament de Filosofia
Edifici Sant Domènec II
Pl. Ferrater i Mora, 1
17004, Girona
sergio.cermeno@udg.edu

References

Andersen, Susan M., Moskowitz, Gordon B., Blair, Irene V., & Nosek, Brian A. [2007]. "Automatic thought". In *Social psychology (2nd edition)*, edited by E. T. Higgins & A. W. Kruglanski. New York, NY: Guilford: 133–72.

Banaji, Mahzarin R. & Hardin, Curtis D. [1996]. "Automatic stereotyping". *Psychological Science* **7**: 136–41.

Bar-Anan, Yoav. & Nosek. Brian A. [2014]. "A comparative investigation of seven indirect attitude measures". *Behavioural Research* **46**: 668–88.

Blair, Irene V., Ma, Jennifer E. & Lenton, Alison P. [2001]. "Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery". *Journal of Personality and Social Psychology* **81**: 828–41.

Blincoe, Sarai & Harris, Monica J. [2009]. "Prejudice reduction in white students: Comparing three conceptual approaches". *Journal of Diversity in Higher Education* **4**: 232.

Brannon, Tiffany N. & Walton, Gregory M. [2013]. "Enacting cultural interests: How intergroup contact reduces prejudice by sparking interest in an out-group's culture". *Psychological Science*. **24**: 1947–57.

Brownstein, Michael & Saul, Jennifer [2016a]. *Implicit Bias & Philosophy: Volume I, Metaphysics and Epistemology*. Oxford: Oxford University Press.

Brownstein, Michael & Saul, Jennifer [2016b]. *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Byrd, Nick [2021]. "What we can (and can't) infer about implicit bias from debiasing experiments". *Synthese* **198**: 1427–55.

Castillo, Linda G., Brossart, Daniel F., Reyes, C.J., Conoley, Collie W. & Phoummarath, Marion J. [2007]. "The influence of multicultural training on perceived multicultural counseling competencies and implicit racial prejudice". *Journal of Multicultural Counseling and Development* **35**: 243–55.

Chapman, Elizabeth N., Kaatz, Anna, & Carnes, Molly [2013]. "Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities". *Journal of general internal medicine* **28**: 1504–10.

Chen, Vivian H., Ibasco, Gabrielle C., Leow, Vetra J. & Lew, Juline Y. [2021]. "The effect of VR avatar embodiment on improving attitudes and closeness toward immigrants". *Frontiers in Psychology* **12**: 705504.

Clobert, Magali, Saroglou, Vassilis, Hwang, Kwang-Kwo. [2005]. "Buddhist concepts as implicitly reducing prejudice and increasing prosociality". *Personality and Social Psychology Bulletin* **41**: 513–25.

Columb, Corey & Plant, Ashby [2011]. "Revisiting the Obama effect: Exposure to Obama reduces implicit prejudice". *Journal of Experimental Social Psychology* **47**: 499–501.

Dasgupta, Nilanjana & Greenwald, Anthony G. [2001]. "On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals". *Journal of Personality and Social Psychology* **81**: 800–14.

Dasgupta, Nilanjana & Asgari, Shaki [2004]. "Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping". *Journal of Experimental Social Psychology* **40**: 642–58.

Dasgupta, Nilanjana & Rivera, Luis M. [2008]. "When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioural intentions". *Social Cognition* **26**: 112–23.

De Houwer, Jan [2011]. "Evaluative conditioning: methodological considerations". In *Cognitive methods in social psychology*, edited by K. C. Klauer, A. Voss, & C. Stahl. New York, NY, USA: Guilford Press: 124–47.

De Houwer, Jan [2014]. "A propositional model of implicit evaluation". *Social Psychology and Personality Compass* **8**: 342–53.

De Houwer, Jan. [2018]. "Propositional models of evaluative conditioning". *Social Psychological Bulletin* **13**: e28046.

Del Pinal, Guillermo D. & Spaulding, Shannon [2018]. "Conceptual centrality and implicit bias". *Mind & Language* **33**: 95–111.

Devine, Patricia G., Forscher, Patrick S., Austin, Anthony J., & Cox, William T. L. [2012]. "Long-term reduction in implicit race bias: A prejudice habit-breaking intervention". *Journal of Experimental Social Psychology* **48**: 1267–78.

FitzGerald, Chloê, Martin, Angela, Berner, Delphine & Hurst, Samia [2019]. "Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review". *BMC Psychology* **7**: 29.

Forscher, Patrick S., Lai, Calvin K., Axt, Jordan R., Ebersole, Charles R., Herman, Michelle, Devine, Patricia G. & Nosek, Brian A. [2019]. "A meta-analysis of procedures to change implicit measures". *Journal of Personality and Social Psychology* **117**: 522–59.

Gast, Anne & De Houwer, Jan [2013]. "The influence of extinction and counterconditioning instructions on evaluative conditioning effects". *Learning and Motivation* **44**: 312–25.

Gawronski, Bertram, Walther, Eva, and Blank, Hartmut [2005]. "Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information". *Journal of Experimental Social Psychology* **41**: 618–26.

Gawronski, Bertram, Hofmann, Wilhelm, & Wilbur, Christopher J. [2006]. "Are "implicit" attitudes unconscious?" *Consciousness and Cognition* **15**: 485–99.

Gendler, Tamar Szabo [2008a]. "Alief and belief". *Journal of Philosophy* **105**: 634–63.

Gendler, Tamar Szabo [2008b]. "Alief in action (and reaction)". *Mind and Language* **23**: 552–85.

Gendler, Tamar Szabo [2011]. "On the epistemic costs of implicit bias". *Philosophical Studies* **156**: 33–63.

Green, Alexander R., Carney, Dana R., Pallin, Daniel J., Ngo, Long H., Raymond, Kristal L., Iezzoni, Lisa I., & Banaji, Mahzarin R. [2007]. "Implicit bias among physicians and its prediction of thrombolysis decisions for Black and White patients". *Journal of General Internal Medicine* **22**: 1231–8.

Greenwald, Anthony G., McGhee, Debbie E., & Schwartz, Jordan LK. [1998]. "Measuring individual differences in implicit cognition: The implicit association test". *Journal of Personality and Social Psychology* **74** (6): 1464–480.

Gündemir, Seval, Homan, Astrid C., de Dreu, Carsten KW., & van Vugt, Mark [2014]. "Think leader, think White? Capturing and weakening an implicit pro-White leadership bias". *PloS one* **9**: e83915.

Haslanger, Sally [2015]. "Social structure, narrative, and explanation". *Canadian Journal of Philosophy* **45**: 1–15.

Higgins, Sarah, Alcock, Stephanie, De Aveiro, Bianca, Daniels, William, Farmer, Harry & Besharati, Sahba [2024]. "Perspective matters: a systematic review of immersive virtual reality to reduce racial prejudice". *Virtual Reality* **28**: 125.

Hofmann, Wilhelm, De Houwer, Jan, Perugini, Marco, Baeyens, Frank, & Crombez, Geert [2010]. "Evaluative conditioning in humans: A meta-analysis". *Psychological Bulletin* **136**: 390–421.

Holroyd, Jules [2012]. "Responsibility for implicit bias". *Journal of Social Philosophy* **43**: 274–306.

Holroyd, Jules [2015]. "Implicit bias, awareness and imperfect cognitions". *Consciousness and Cognition* **33**: 511–23.

Holroyd, Jules [2016]. "What do we want from a model of implicit cognition?" *Proceedings of the Aristotelian Society* **116**: 153–79.

Holroyd, Jules & Sweetman, Joseph [2016]. "The heterogeneity of implicit biases". In *Implicit Bias and Philosophy. Volume I: Metaphysics and Epistemology*, edited by Michael Brownstein and Jennifer Saul. Oxford: OUP: 80–103.

Hughes, Sean, Barnes-Holmes, Dermot & De Houwer, Jan [2011]. "The dominance of associative theorizing in implicit attitude research: Propositional and behavioural alternatives". *The Psychological Record* **61**: 465–98.

Johnson, Gabrielle M. [2020]. "The Structure of bias". *Mind* **516**: 1193–236.

Jones, Christopher R., Olson, Michael A., & Fazio, Russell H. [2010]. "Evaluative conditioning: The "how" question". *Advances in Experimental Social Psychology* **43**: 205–55.

Joy-Gaba, Jennifer A., & Nosek, Brian A. [2010]. "The surprisingly limited malleability of implicit racial evaluations". *Social Psychology* **41**: 137–46.

Kosslyn, Stephen M., Ganis, Giorgio, & Thompson, William L. [2006]. "Mental imagery and the human brain". *Psychology* **1**: 195–209.

Kurdi, Benedek, & Dunham, Yarrow [2021]. "Sensitivity of implicit evaluations to accurate and erroneous propositional inferences". *Cognition* **214**: 104792.

Kurdi, Benedek, Morris, Adam & Cushman, Fiery A. [2022]. "The role of causal structure in implicit evaluation". *Cognition* **225**: 105116.

Kurdi, Benedek, Mann, Thomas C. & Ferguson, Melissa J. [2022]. "Persuading the implicit mind: Changing negative implicit evaluations with an 8-minute podcast". *Social Psychological and Personality Science* **13**: 688–97.

Kurdi, Benedek, Morehouse, Kristen N. and Dunham, Yarrow [2023]. "How do explicit and implicit evaluations shift? A preregistered meta-Analysis of the effects of co-occurrence and relational information". *Journal of Personality and Social Psychology* **124**: 174–202.

Lai, Calvin K., Marini, Maddalena, Lehr, Steven A., Cerruti, Carlo, Shin, Jiyun-Elizabeth. L., Joy-Gaba, Jennifer A., Nosek, Brian A. [2014]. "Reducing implicit racial preferences: I. A comparative investigation of 17 interventions". *Journal of Experimental Psychology: General* **143**: 1765–85.

Lai, Calvin K., Skinner, Allison L., Cooley, Erin, Murrar, Sohad, Brauer, Markus, Devos, Thierry, ... Nosek, Brian A. [2016]. "Reducing implicit racial preferences: II. Intervention effectiveness across time". *Journal of Experimental Psychology: General* **145**: 1001–16.

Levy, Neil [2014]. "Consciousness, implicit attitudes, and moral responsibility". *Noûs* **48**: 21–40.

Levy, Neil [2015]. "Neither fish nor fowl: Implicit attitudes as patchy endorsements". *Noûs* **49**: 800–23.

Machery, Edouard [2017]. "Do indirect measures of biases measure traits or situations?" *Psychological Inquiry* **28**: 288–91.

Madva, Alex [2016a]. "Why implicit attitudes are (probably) not beliefs". *Synthese*, **193**: 2659–84.

Madva, Alex [2016b]. "A Plea for anti-anti-individualism: How oversimple psychology

misleads social policy". *Ergo* **3**: 701–28.

Madva, Alex [2017]. "Biased against debiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice". *Ergo* **4**: 145–79.

Mandelbaum, Eric [2013]. "Against alief". *Philosophical Studies* **165**: 197–211.

Mandelbaum, Eric [2016]. "Attitude, inference, association: On the propositional structure of implicit bias". *Noûs* **50**: 629–58.

Mann, Thomas C., & Ferguson, Melissa J. [2017]. "Reversing implicit first impressions through reinterpretation after a two-day delay". *Journal of Experimental Social Psychology* **68**: 122–7.

McConnell, Allen R., & Leibold, Jill M. [2001]. "Relations among the implicit association test, discriminatory behaviour, and explicit measures of racial attitudes". *Journal of Experimental Social Psychology* **37**: 435–42.

McGrane, Joshua A. & White, Fiona A. [2007]. "Differences in Anglo and Asian Australians' explicit and implicit prejudice and the attenuation of their implicit in-group bias". *Asian Journal of Social Psychology* **10**: 204–10.

Miles, Eleanor & Crisp, Richard J. [2014]. "A meta-analytic test of the imagined contact hypothesis". *Group Processes & Intergroup Relations* **17**: 3–26.

Mitchell, Chris, De Houwer, Jan & Lovibond, Peter [2009]. "The propositional nature of human associative learning". *Behavioural and Brain Sciences* **32**: 183–98.

Nanay, Bence [2021]. "Implicit bias as mental imagery". *Journal of the American Philosophical Association* **7**: 1–19.

Oswald, Frederick L., Mitchell, Gregory, Blanton, Hart, Jaccard, James & Tetlock, Philip E. [2013]. "Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies". *Journal of Personality and Social Psychology* **105**: 171–92.

Paluck, Elisabeth L., Porat, Roni, Clark, Chelsey S. & Green, Donald P. [2021]. "Prejudice reduction: Progress and challenges". *Annual Review of Psychology* **72**: 533–60.

Payne, B. Keith [2001]. "Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon". *Journal of Personality and Social Psychology* **81**: 181–91.

Payne, B. Keith, Lambert, Alan J. & Jacoby, Larry L. [2002]. "Best laid plans: effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons". *Journal of Experimental and Social Psychology* **38**: 384–96.

Payne, B. Keith, Vuletich, Heidi A., & Lundberg, Kristjen B. [2017]. "The bias of crowds: How implicit bias bridges personal and systemic prejudice". *Psychological Inquiry* **28**: 233–48.

Peck, Tabitha C., Seinfeld, Sofia, Aglioti, Salvatore M. & Slater, Mel [2013]. "Putting yourself in the skin of a black avatar reduces implicit racial bias". *Consciousness and Cognition* **22**: 779–87.

Pelletier, Petra & Drozda-Senkowska, Ewa [2020]. "Virtual reality as a tool for deradicalizing the terrorist mind: Conceptual and methodological insights from intergroup conflict resolution and perspective-taking research". *Peace and Conflict: Journal of Peace Psychology* **26**: 449–59.

Rozin, Paul, Markwith, Maureen & Ross, Bonnie [1990]. "The sympathetic magical law of similarity, nominal realism, and neglect of negatives in response to negative labels". *Psychological Science* **1**: 383–4.

Saul, Jennifer [2012]. "Skepticism and implicit bias". *Disputatio* **37**: 243–63.

Sullivan-Bissett, Ema [2019]. Biased by our imaginings. *Mind & Language* **34**: 627–47.

Sullivan-Bissett, Ema [2023]. Virtually imagining our biases. *Philosophical Psychology* **36**: 860–93.

Thériault, Rêmi, Olson, Jay A, Krol Sonia A, Raz, Amir [2021]. "Body swapping with a black person boosts empathy: Using virtual reality to embody another". *Quarterly Journal of Experimental Psychology* (Hove), **74**:2057–74.

Toribio, Josefa [2018]. "Implicit bias: From social structure to representational format". *Theoria* **33**: 41–60.

Turner, Rhiannon N. & Crisp, Richard J. [2010]. "Imagining intergroup contact reduces implicit prejudice". *British Journal of Social Psychology* **49**: 129–42.

Welpinghus, Anna [2020]. "The imagination model of implicit bias". *Philosophical Studies* **177**: 1611–33.