Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Virtually imagining our biases

Ema Sullivan-Bissett

To cite this article: Ema Sullivan-Bissett (2023) Virtually imagining our biases, Philosophical Psychology, 36:4, 860-893, DOI: 10.1080/09515089.2023.2184334

To link to this article: https://doi.org/10.1080/09515089.2023.2184334

Published online: 01 Mar 2023.

PHILOSOPHICAL PSYCHOLOGY
2023, VOL. 36, NO. 4, 860–893
https://doi.org/10.1080/09515089.2023.2184334

Routledge
Taylor & Francis Group

Virtually imagining our biases

Ema Sullivan-Bissett

School of Philosophy, Theology and Religion, University of Birmingham, Birmingham, England

ABSTRACT

A number of studies have investigated how immersion in a virtual reality environment can affect participants' implicit biases. These studies presume associationism about implicit bias. Recently philosophers have argued that associationism is inadequate and have made a case for understanding implicit biases propositionally. However, no propositionalist has considered the empirical work on virtual reality and how to integrate it into their theories. I examine this work against a propositionalist background, in particular, looking at the belief and patchy endorsement models. I argue that the results therein can only be accommodated by a model which recognizes structural heterogeneity, that is, one which allows for implicit biases being both associatively and non-associatively structured. My preferred view – that implicit biases are constituted by unconscious imaginings – allows for this, as well as for heterogeneity at the level of content (propositional and imagistic), a feature which also earn its explanatory keep in this context. I conclude that empirical work on virtual reality and implicit bias gives us a reason to prefer a pluralist model of bias, and that my unconscious imagination model, in its recognizing wide-ranging heterogeneity, is uniquely placed to accommodate the results of work on virtual reality and bias mitigation.

ARTICLE HISTORY
Received 31 March 2022
Accepted 18 February 2023

KEYWORDS
Implicit bias; associationism; propositionalism; doxasticism; unconscious imagination; virtual reality

1. Preliminaries

Recently, philosophers have been interested in virtual reality (VR). In particular, they have considered the metaphysics: whether virtual objects are fictional objects (Chalmers, 2017; McDonnell & Wildman, 2019, 2020) and how the ontological status of virtual objects bears on their intrinsic value (Mooradian, 2006); the epistemology: how virtual reality could contribute to knowledge (Cogburn & Silcox, 2014) and whether we could have only virtual bodies (Meijsing, 2006); the philosophy of perception: what it means to perceive a virtual object (Diodato, 2014); the aesthetics: implications of the virtual body in this domain (Diodato, 2012); and philosophical

CONTACT Ema Sullivan-Bissett e.l.sullivan-bissett@bham.ac.uk School of Philosophy, Theology and Religion, European Research Institute, University of Birmingham, Pritchatts Road, Birmingham B15 2SB, England

psychology: virtual reality and clinical implications for anorexia nervosa (Gadsby, 2019).

Meanwhile, *other* philosophers, also recently, have been interested in the attitudinal nature of implicit bias, in particular, what kind of mental construct implicit biases are and the kind of mental operations in which they partake. Working against the orthodox view in psychology of implicit biases as *associations*, these philosophers have argued that associationism is inadequate, given some experimental data on the behavior of implicit bias. They have thus moved to understanding implicit biases propositionally: Neil Levy (2015) in terms of *patchy endorsements* and Eric Mandelbaum (2016) in terms of *unconscious beliefs*.

Following David Chalmers, I will understand a virtual reality environment as one that is "immersive, interactive, [and] computer-generated" (2017, p. 312). In psychology, the areas of implicit bias and virtual reality have been brought together; studies have looked at how immersion and/or embodiment in virtual reality environments can affect participants' gender and race biases. However, discussion of these studies takes place against a presumption of implicit biases as associations between concepts, or between concepts and valences. Neither psychology nor philosophy has yet looked at these results against a background understanding of implicit biases as having propositional contents and being structured non-associatively. Here I argue that the empirical work on implicit bias and virtual reality can only be accommodated by an account of implicit bias which recognizes wide-ranging heterogeneity. My view (defended elsewhere, Sullivan Bissett, 2019) is that we should understand implicit biases in terms of *unconscious imaginings* which can be structured both associatively and non-associatively, and can also have both propositional and imagistic contents. It is uniquely placed to accommodate the recent work showing the inadequacies of pure associationism, as well as the experimental data generated from studies on virtual reality.

2. Associationism and propositionalism about implicit bias

To begin, we can understand implicit biases as "the processes or states that have a distorting influence on behavior and judgment, and are detected in experimental conditions with implicit measures" (Holroyd, 2016, p. 154). Implicit biases are fast and habitual, and can operate in the absence of agent awareness. Very roughly, implicit biases are posited as mental items which influence common micro-behaviors and discriminations, which cannot be tracked, predicted, or explained by a subject's explicit attitudes. They are thought to be inaccessible to consciousness, automatically activated, and prevalent among even those who identify as egalitarian. From this quick summary of implicit bias we can draw out a couple of key features which any

model thereof ought to be able to accommodate: their being mental items of which we are not aware, and their having a role in judgment and behavior.[1] The standard view of implicit biases characterizes them as *associations* brought about by the learning history of the subject (Levy, 2015, p. 803). In the presence of certain stimuli, stored associations between concepts and valences (e.g., *Black male* with negative valence) are activated. Some theorists also allow for associations between concepts (e.g., *woman* and *weakness*) (Mandelbaum, 2016, p. 630). On the associationist picture then, an implicit bias *is* an association between mental items (see Fazio, 2007; Gawronski & Bodenhausen, 2011; Rydell & McConnell, 2006).

Some authors working in this tradition have argued that the category of implicit bias exhibits great heterogeneity, and that any account worth its salt ought to accommodate this. In brief, biases are heterogeneous with respect to their contents (they are about different social groups), and with respect to their expression (even biases about the same social groups can vary in their manifestation as measured by different Implicit Association Tests (IATs) and in their behavioral expression) (see Amodio & Devine, 2006; for relevant experimental data, and Holroyd & Sweetman, 2016, for more general discussion). Some theorists have attempted to make good on this heterogeneity but have done so within the confines of an associative framework (for discussion see Sullivan Bissett, 2023, §5). For example, Jules Holroyd and Joseph Sweetman argue for a distinction between *semantic* and *affective* associations, Bruce Huebner (2016) argues for a variety of ways implicit biases get internalized, and Guillermo Del Pinal and Shannon Spaulding (2018) argue for heterogeneity at the level of encoding.

However, there has been a recent move away from associationism about implicit bias to understanding these mental items as *non-associatively structured* and having *propositional contents*. On such a picture, there is a specific relation between the constituents of bias absent on the picture of them as associations (Levy, 2015, p. 804). If a mental state has propositional contents, it has satisfaction conditions, whilst states with only associative contents do not (Mandelbaum, 2013, p. 199, fn. 1). So instead of understanding an implicit bias regarding *women* and *weakness* as an *association* between these concepts, propositionalists understand it as a propositionally structured mental representation with the content *women are weak*.

In the most robust defense of propositionalism, Mandelbaum overviews a host of empirical evidence which he argues cannot be accommodated by an associationist picture but is easily understood against a background of implicit biases as propositional (cf. Brownstein et al., 2019; Toribio, 2018). I will not relay his survey here, but to give an illustrative example: Mandelbaum reports on Bertram Gawronski and colleagues (2005) study of cognitive balance and implicit attitudes. Participants were introduced to a photo of an unfamiliar person (CS1), which was then paired with either

positive or negative statements (to set up the association between (CS1) and a particular evaluation). Then participants were introduced to a second photo of an unfamiliar person (CS1) and told either that (CS1) liked (CS2) or that (CS1) did not like (CS2). Subjects then underwent an explicit likability rating to gauge explicit attitudes toward (CS1) and (CS2), and an affective priming task in which they had to identify positive or negative words as such, having been primed with one of the two images (Gawronski et al., 2005, p. 621). The results showed that, for example, if (CS1) was paired with negative statements, and participants were told that (CS1) did not like (CS2), then participants liked (CS2). Reflecting on this Mandelbaum suggests that an associative theory would predict the opposite: negative valence + negative valence = negative negative valence (that is, an associative account would predict "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" (Mandelbaum, 2016, p. 639)). Mandelbaum sums up the implication of this by noting that "if you find two negatives making a positive, what you've found is a propositional, and not an associative, process" (2016; 639, cf. Levy, 2015, pp. 811–12). Further support for the claim that associative models are unable to account for many experimental findings is given by discussions of other work (for evidence that implicit biases are sensitive to argument strength, see Brinol et al., 2009; for evidence that implicit biases are adjustable in light of peer judgment, see Sechrist & Stangor, 2001).

For the purposes of this paper, I will take it that associationism is not a complete explanation of implicit bias, that is, empirical evidence has revealed that implicit biases should, at least in some cases, be modeled propositionally. Mandelbaum's case for this claim is, if not conclusive, at the very least extremely challenging for the prospects of a pure associationist model. The terms of the debate have thus shifted. It is for this reason that I do not spend time discussing whether associationism can accommodate the results from studies on virtual reality immersion and implicit bias. For suppose that it could. How interesting would that be when set against a background of the more foundational problems it faces? My project then is downstream of the question of whether pure associationism is adequate, and rather asks: if the propositionalists are right that it is not, can their preferred approach do the work when we look to a particular kind of evidence base (that of the influence of virtual reality immersion on implicit bias)? I will argue that it cannot. Given this, pure propositionalism is not adequate either. Let us then see if a hybrid approach, one which recognizes heterogeneity of structure, can do the work. Later we will see that my preferred model of bias, in its embracing of heterogeneity of structure as

well as contents, can accommodate the empirical results from virtual reality studies.

3. Empirical work on implicit bias and virtual reality

I turn now to a handful of studies on implicit bias and virtual reality, before looking at the belief and patchy endorsement models of implicit bias and how they might accommodate their results.

3.1. Implicit bias decrease following virtual reality immersion

I begin with two studies which found that implicit race bias decreased after immersion in a virtual reality environment and embodiment in a virtual avatar. Peck and colleagues (2013) were interested in whether embodiment could induce a body ownership illusion when the avatar was of a different race to the participant, and whether if so, the illusion could reduce negative implicit responses directed toward the race in which they were embodied. The study involved sixty female, light-skinned participants, split equally into four conditions: *light skin embodiment, dark skin embodiment, non-embodied dark skin, and alien skin (purple) embodiment*. In the embodiment conditions the participants wore a head-mounted display and a body-tracking suit, such that they looked down and ahead into a virtual mirror and saw a programmed virtual body, which moved synchronously with their own (Peck et al., 2013, p. 780). In the non-embodiment condition participants saw a virtual body reflected in a virtual mirror, which moved asynchronously with their own. This condition was designed to isolate the effects of embodiment from mere exposure to a dark-skinned avatar, and the alien skin condition was designed to isolate the effects of race from a strangeness effect (that is, the effect of an avatar having a *different* skin color).

Peck and colleagues found that scores on body ownership were significantly lower in the non-embodiment condition than in any of the three embodied conditions (within which there were no significant differences) (Peck et al., 2013, p. 783). They found that participants in the dark-skin embodiment condition had a lower level of implicit race bias as measured by the IAT[2] after the exposure compared with those embodied in light skin bodies, alien bodies, and those not embodied at all.

Domna Banakou and colleagues (2016) were interested in replicating the results of Peck and colleagues (2013) study, whilst also investigating whether the reduction in bias lasted for at least one week. Sixty female participants were encouraged to follow the movements of a virtual Tai Chi teacher of Asian appearance, whilst embodied in either a Black or a White virtual body, with visuomotor synchrony (there was no non-embodiment condition). All participants were given a race IAT one week before their first

exposure and one week after their final exposure. They were also given a questionnaire after each exposure to assess strength of body ownership and agency.

Banakou and colleagues found similar results one week after exposure to the results Peck and colleagues found immediately after exposure. The mean IAT results increased for those embodied in White avatars and decreased for those in Black avatars (Banakou et al., 2016, p. 6).

3.2. Implicit bias increase following virtual reality immersion

I turn now to two studies which found an *increase* in implicit bias following virtual reality immersion and embodiment. Victoria Groom and colleagues investigated the effects of imagining and embodiment on implicit racial bias.

White and non-white participants (distinguished by self-report) were asked to imagine themselves as a model whose photograph they were given, where models were categorized as unambiguously Black or White in experimental pre-screening. They were instructed to "imagine a day in the life of this individual as if you were that person", and then asked to imagine that the person was about to have a job interview (Groom et al., 2009, p. 238). They were told that they would be asked interview questions in a virtual reality environment, and that they should answer those questions as though they were the person in the picture.

Next participants wore a head mounted display and were told to turn 180 degrees. Those in the *embodied condition* saw a virtual mirror and were asked to confirm that they saw their avatar in the mirror, were told that the mirror image was of the person in the photograph, and that this would be how they appeared to others in the virtual environment. Participants in the imagined condition did not see their avatar in a virtual mirror; they saw a "window", which displayed a room identical to the one they were in. All participants were asked to move closer to the confederate and the distance was recorded. The confederate asked the participant to move closer again and the distance was again recorded. What followed was a series of interview questions. At the end participants exited the virtual reality environment, re-sat the IAT, and took questionnaires which included measures of explicit race bias.

Groom and colleagues were interested in testing three hypotheses: (1) embodiment would generate larger differences in implicit bias than mere imagining, (2) those assigned Black models would have less explicit and implicit bias than those assigned White models, and (3) those assigned Black models would have more explicit and implicit bias than those with White models. Hypotheses (2) and (3) are predictions of *perspective-taking theory* and *stereotype activation theory* respectively. Hypothesis (1) was supported, indeed, participants in the imagined condition "produced nearly identical

IAT scores regardless of whether the model was White or Black" (Groom et al., 2009, p. 242). This is a result which is described as "unanticipated" given the literature on perspective taking (Groom et al., 2009, p. 244).

Of (2) and (3), it was hypothesis (3) that was confirmed. That is, those participants embodied in a Black avatar exhibited *increased* implicit race bias compared with those embodied in a White avatar. Two explanations have been offered for this result (that is, for hypothesis (3) being confirmed over hypothesis (2)). First, body ownership was not measured, and the visuomotor synchrony only covered head movements (not full body), as well as the embodiment phase only lasting 60–75 seconds, significantly less than the embodiment phase in Peck and colleagues experiment which was eleven minutes (Peck et al., 2013, p. 785; see also Lopez et al., 2019, p. 1). Second, participants were in the context of a job interview, a situation in which race discrimination operates (Dasgupta, 2004). Peck and colleagues suggest that the increase in implicit race bias was due to being placed in a situation known for discrimination, rather than being due to embodiment (Peck et al., 2013, p. 785; see also Banakou et al., 2016, p. 8; Slater, 2017, p. 26; Schulze et al., 2019, p. 362). In an environment such as a job interview with a White confederate, the increase in bias might be explainable by appeal to stereotype activation, which could have "overwhelmed any positive effects of perspective-taking" (Lopez et al., 2019, p. 3). Following Sarah Lopez, we can understand stereotype activation as occurring when "features such as gender and race activate stereotypes about that group" (Bargh et al., 1996, p. 230; cited in Lopez et al., 2019, p. 3). The suggestion then is that when participants are embodied in a Black avatar, stereotypes held about Black people are triggered. It was the context which played a role in the confirmation of *stereotype activation theory* in this experiment, not the embodiment (which we might otherwise have expected to confirm *perspective-taking theory*).

Let us turn to a final experiment, run by Lopez and colleagues, who investigated the effects of immersion and embodiment on gender bias. They embodied twenty-four male participants in male or female avatars with full visuomotor synchrony, in which they carried out a Tai Chi task by mimicking the movements of a virtual Tai Chi master, for eight minutes. The avatars wore clothing associated with physical activity. Whilst humanoid in appearance, further aesthetic details of the teacher were kept to a minimum, and the skin was human but gender-neutral orange (Lopez et al., 2019, p. 5).

The participants embodied in female bodies displayed *higher* levels of bias against women after the exposure than those embodied in male avatars, whose levels of bias *decreased* (Lopez et al., 2019, p. 9). Levels of body ownership were similar regardless of avatar. Lopez and colleagues suggest that unlike with Groom and colleagues' study, the increase in bias cannot be

explained by the absence of full body visuomotor synchrony, since they had that in the experiment. Instead, they offer two explanations for their results. The first is that participants may have become frustrated with the task and attributed that to the most salient factor (gender). The second is that the environment was not a neutral one with respect to the social group they were interested in attitudes about, that is, stereotype activation could have occurred because of sport being a negatively stereotyped activity for women (Lopez et al., 2019, p. 9).[3]

3.3. Explaining implicit bias change from virtual reality immersion

The research investigating implicit bias mitigation through immersion or embodiment in virtual reality environments is in its infancy (Salmanowitz, 2016, p. 138), and so these studies are just the beginning of what is presumably a growing area of interest. However, it might already be thought that things do not look promising: two studies showing a mitigation of bias and two studies (plus the third mentioned in n. 3) showing a worsening of bias does not say much for the efficacy of this technique. I say two things to this. First, it is consistent with this small mixed bag of results that virtual reality immersion and embodiment reliably mitigate implicit bias. As we have seen from the explanations given of the results from Groom and colleagues', Lopez and colleagues', and Schulze and colleagues' studies, it may well be the lack of total visuomotor synchrony and the particular context of the immersion, which explains why virtual reality immersion was not mitigating. In the studies with full visuomotor synchrony and a context not negatively stereotyped for the social category in play, implicit biases were successfully mitigated. In addition, there is significant research showing a relationship between synchrony of various kinds and body ownership. If, as I suggest later, body ownership is key to implicit bias mitigation, we can perhaps be confident that inducing this in a virtual reality environment – an effect strongly linked to visuomotor synchrony – would deliver positive mitigation effects (indeed, this will play a crucial role in my case for modeling implicit biases on imagination). Second, even though the studies show different effects on implicit bias through immersion and embodiment, they all show effects nonetheless. The effect on bias (either mitigating or worsening) as a result of virtual reality immersion or embodiment is something those interested in the nature of implicit bias ought to be able to explain.

Consideration of what implicit biases *are* do not take place in discussions of these experimental results, since experimenters tend to assume associationism about bias. Earlier I referred to the empirical case against a pure associationism about bias, and noted that recognition of associationism's inadequacies had motivated a recent move to propositionalism (§2). So now

we can ask: is the data from the virtual reality studies consistent with propositionalist accounts? Can the recent move to propositionalism withstand this particular set of experimental results?

Next I discuss explanations of these results from the point of view of Mandelbaum's and Levy's respective accounts, and suggest that they face problems. Then I outline my imagination model which recognizes implicit biases as both associatively and non-associatively structured and as having both propositional and imagistic contents. I have argued elsewhere that this model is theoretically virtuous in myriad ways, and I will not repeat these arguments here. Rather, I will argue that it can accommodate the empirical results from studies on virtual reality without facing the problems of purely propositional models. Such results then give us new reasons to accept the imagination model.

4. Implicit bias as propositional and virtual reality

Let us turn to our two propositional models of bias, according to which implicit biases are unconscious beliefs[4] (Mandelbaum, 2016) or patchy endorsements (Levy, 2015). On its face, the claim that implicit biases are beliefs might sound surprising. On a traditional Cartesian way of thinking about the matter, beliefs are largely evidence-responsive, and propositions can be deliberated upon, and then taken up in belief, *or not*. Beliefs might be thought to be propositional states whose contents we take to be true, whose contents we take ourselves to be *committed to*, and it is also usually thought that it is not possible to hold conflicting beliefs. Given this, a natural reaction to one's first encounter with the claim that implicit biases are beliefs might be skepticism, or outright denial (given the kind of incongruent features these respective mental items are typically taken to have, it simply cannot be so!). Or at least, if implicit biases are beliefs, a traditional conception of belief may have to be sacrificed when we reflect on cases of implicitly biased egalitarians, where the content of one's explicit beliefs and implicit biases are in tension.

However, Mandelbaum's background understanding of belief departs from the more traditional understanding just outlined, such that this natural reaction is misplaced, or at least, loses its argumentative force. And here we need to spend a moment on the opposing *Spinozan* account. According to such an account of belief formation, *we believe any truth-apt proposition that we represent*. There is no gap between representing a truth-apt proposition, and believing it, "the act of understanding is the act of believing" (Gilbert et al., 1993, p. 222). In light of the putative worry that such an account would attribute conflicting beliefs to single subjects, the Spinozan doxasticist has at her disposal the idea of the mind as *fragmented* (à la Egan, 2008; Lewis, 1982; Stalnaker, 1984).

Although Levy endorses Mandelbaum's case for propositionalism on the grounds that implicit biases sometimes feature in content-driven transitions (2015, p. 816), he argues that they exhibit sensitivity and responsiveness to other mental representations which is too "patchy and fragmented" for them to be beliefs (2015, p. 800). With respect to inference, Levy argues against Mandelbaum's view by appeal to evidence showing that implicit attitudes often work non-inferentially. For example, John F. Dovidio and colleagues (1997) study showed that implicit bias against Black people was predictive of certain behaviors when interacting with a Black interviewer (i.e., less eye contact and more blinking) (Levy, 2015, p. 813). Levy argues that it is difficult to give an inferential account in terms of belief of what is going on in this case, and that a whole host of empirical evidence of this kind showing implicit attitudes' involvement in microbehaviors, puts pressure on the view that implicit biases are beliefs (2015, p. 813) (see Bessenoff & Sherman, 2000; Chen & Bargh, 1997; McConnell & Leibold, 2001; Wilson et al., 2000).

Levy's view is that implicit biases are *patchy endorsements*. The *endorsement* part is that a subject commits to the world being the way the proposition picks out, whilst the *patchy* part recognizes that implicit biases only respond to some sorts of evidence and only feature in some sorts of inference. Levy argues that this kind of state better matches the functional profile of implicit biases.

For the purposes of the following discussion, I will take it that both Mandelbaum and Levy are committed to implicit biases being formed and updated in a Spinozan way. That is explicitly the case for Mandelbaum's beliefs, whilst Levy's patchy endorsements respond to the world in a patchier way. However, this simplification is permissible for the following reason: any explanation of implicit bias change following virtual reality immersion or embodiment will require some level of automaticity. If implicit biases are patchy endorsements, this might be less reliable and more context-sensitive, but the presumption would have to be that the virtual reality studies create the right kind of context for implicit biases so constituted to be affected. We will see shortly that two of the three problems I raise for these models do not rely on a Spinozan background of propositional updating, and so if Levy's patchy endorsements are less Spinozan that Mandelbaum's beliefs, they still cannot play the required role in accommodating the results from virtual reality studies. The third problem I identify arises for the propositionalist in virtue of the endorsement of a Spinozan approach.

Let us turn then to how these accounts might explain the experimental results overviewed earlier. Recognizing that visuomotor synchrony is key to body ownership and bias change, we can suppose that a doxastic explanation of implicit bias change as a result of virtual reality immersion goes

something like this: *embodiment in a virtual avatar causes in a participant the formation of beliefs (conscious and unconscious) which affect the unconscious beliefs identified as implicit biases*. Now, it might be thought that this kind of explanation requires that virtual reality users form beliefs in extremely naïve ways. The doxastic explanation is thus implausible since there is little danger of acquiring false beliefs of the sort that, together with other beliefs, would be incongruent with implicit biases. As Chalmers points out, "given that the user knows they are using VR, they will not form the belief they are interacting with non-virtual objects in physical space. They will know full well that they are interacting with virtual objects in virtual space" (2017, p. 327). Similarly, we should not suppose that participants in the above experiments formed the beliefs that *they were in a job interview* or *had female/Black bodies*. Banakou and colleagues make this point when they note that immersed participants do not "in any way *believe* that their body has changed" (Banakou et al., 2016, p. 9).

However, this dismissal of the explanation fails to pay attention to the Spinozan notion of *belief* in play. Once this is attended to, then, on the face of it at least, it may not be implausible to think that participants do in fact form beliefs, so understood, while immersed in a virtual reality environment. A Spinozan account legitimizes being in the business of belief and makes palatable the idea that in virtual reality environments subjects form beliefs which can interact with their implicit biases.

The mechanism for patchy endorsements might look a bit different. Levy takes it that propositional models of implicit bias are committed to the idea that a subject takes there to be a determinate relation between constituents of an attitude (e.g., *women* and *weakness*), because she (unconsciously) endorses a proposition like *women are weak* (Levy, 2015, p. 805). For the patchy endorsement model then, I suggest an explanation for implicit bias change goes something like this: *embodiment causes in a participant the formation of propositions (conscious and unconscious) which the subject endorses, which affect the patchy endorsements identified as implicit biases.*

I turn now to two problems facing propositional models in this context, and then a third issue which is more serious for the doxastic account in particular.

4.1. Problem one: propositional representations

The first problem relates to virtual environments being apt to produce propositional representations that can be automatically up-taken to belief or patchily endorsed. This problem is offered in the spirit of speculation, and so I will be brief.

There are some reasons to think that virtual reality environments may not be the kind of environment in which even Spinozan propositional up-taking

might occur, *for some contents*. Of course, even a virtual reality environment is apt to produce plenty of propositions which can be represented and *thus* up-taken, such as *I am in a VR environment*, or *the VR body suit is uncomfortable*. These propositions may well be represented, and thus believed (or patchily endorsed), in the usual way. The virtual reality content itself may even produce representations like *that looks funny*, or *that's surprisingly realistic* which may also be up-taken to belief or patchily endorsed. There is a question though whether the virtual reality content is liable to produce what we might call *face-value* representations (i.e., those that would be true were the virtual reality environment the actual environment) which are up-taken to belief or patchily endorsed.

The phenomenology of mirrors is illustrative here. Chalmers suggests that as experienced users, our visual experience alters when interacting with a mirror, such that there's a "distinctive mirror phenomenology" (Chalmers, 2017, p. 331). I take this to mean that our background beliefs about how mirrors work orient us in such a way that what is presented at face value, is not up-taken to belief/patchily endorsed. Presumably the Spinozan about belief wouldn't take us to *believe* mirror representations, like *there is a duplicate of me* or *my written tattoo is backwards* (mirrors would be far scarier objects if such things were to occur).[5] Similarly the patchy endorsement theorist might balk at the idea that we endorse *all* propositional representations. And so this might suggest that in some contexts, i.e., those with associated particular phenomenologies, belief or patchy endorsement formation may not be as cheap as usual, which is to say, automatic Spinozan updating may be prevented.

Let us return to the virtual reality environment, and consider the potential implications of what Chalmers calls *the phenomenology of virtuality*, something subjects experience when they know they are in a virtual environment (Chalmers, 2017, p. 331).[6] The visual experience in a virtual reality environment is significantly and importantly different from the visual experiences we have in non-virtual environments. For example, Chalmers suggests that a user of virtual reality may "perceive virtual objects *as virtual*" (2017, p. 331). Relatedly, Roberto Diodato points out that virtual reality immersion results in "an unmistakable quality of experience which is different from what we hold to be 'real'" (2014, p. 48). We might wonder whether, just like with our interactions with mirrors, this distinctive phenomenology has implications for whether the environment is one in which propositional representations occur, which are automatically up-taken to belief or patchy endorsement. If it is not such an environment, then the propositionalist has no explanation for implicit bias change.

4.2. Problem two: which proposition?

The propositionalist is committed to there being a belief or patchy endorsement which is produced as a result of the virtual reality immersion. Assuming that virtual reality environments can produce propositional representations automatically up-taken to belief or patchy endorsement, identifying the content of the relevant proposition, and how that interacts with bias, creates a dilemma. On the one horn, (i) the propositional representations plausibly arising in virtual reality environments will not be ones that would affect bias, and on the other (ii) the propositional representations which might affect bias are not ones which would plausibly arise in virtual reality environments. Consider (i): those immersed will automatically believe or patchily endorse contents such as *I am in a virtual job interview* or *I have virtual Black skin*. The problem with this is that it's unclear that these contents would be able to affect bias (how could your *virtually* being in a job interview or your *virtually* having Black skin interact with your implicit biases about non-virtual Black people?). So it is one thing to make a case for uncritical belief- or patchy endorsement-forming mechanisms being such that even in virtual reality environments propositional representations are taken up, but the mostly likely candidate contents will not be appropriate ones for interacting with implicit bias.

Alternatively, consider (ii): those contents which could plausibly affect bias are not the kinds of contents which would be represented and thus up-taken in a virtual reality environment. [7] The kinds of content which could interact with implicit bias might include *I am Black*, or *I have a Black body*. Now of course, such beliefs are *consistent* with biased beliefs about Black people. But if we consider other beliefs subjects might have about themselves, and if we accept that implicit biases participate in (at least some kinds of[8]) inference, the propositions up-taken in virtual reality environments may have an effect on implicit biases. For example, a subject may have an implicit bias constituted by the proposition *Black people are dangerous*. If embodiment in a Black avatar immersed in a virtual reality environment tokens the proposition *I am Black*, and that interacts with the belief *I am not dangerous*, then implicit biases incongruent with the combination of newly represented propositions may be weakened or revised.

However, to say that it is contents like these which are represented is to go against the orthodox understanding of how users navigate and understand these environments, that is, they perceive them *as virtual*. The distinctive phenomenology of virtuality identified by Chalmers and echoed by Diodato is such that insofar as any propositional representation is going on, the contents of those representations will include the fact that what is being perceived are *virtual* objects.

In sum, if virtual reality environments are ones in which propositional representations are there for the up-taking by Spinozan mechanisms, a story needs to be told regarding either (i) how propositions such as *I have a virtual Black body* interact with implicit bias, or (ii) how propositions like *I am Black*, which could plausibly interact with implicit bias, get represented. The propositionalist needs either to opt for plausibility of contents taken up in virtual reality (*I am in a virtual job interview*) but then loses a story about how such contents might interact with bias. Or opt for a story about how newly represented contents might interact with bias (*I am Black*), but at the expense of plausibility that such contents would arise in the virtual environment.[9]

The problems of making plausible the idea that propositions can be up-taken in virtual environments at all, or difficulties (i) explaining why certain propositions would affect bias or (ii) how bias-affecting propositions arise, do not arise for the imagination model of bias, as we will see. That is in virtue of the heterogeneity it recognizes in the constituents of implicit bias. For a model exclusively in the business of propositional contents, these challenges are serious ones.

4.3. Problem three: longevity of effects

I turn now to a final problem for propositionalism, which is more serious for the belief model. This relates to Banakou and colleagues' finding that the decrease in implicit bias following virtual reality immersion lasted for at least one week. As I have already said, the kind of explanation the propositionalist will offer might go via embodiment in a virtual avatar causing one to represent propositions which can affect the unconscious beliefs or patchy endorsements constitutive of implicit bias. Notwithstanding the problem about candidate propositions and the phenomenology of virtuality above, a problem with this kind of explanation is that even on a Spinozan understanding of belief and patchy endorsement, the relevant newly formed attitudes (and their effects) cannot be expected to stick around. It is no part of the Spinozan position that propositions are up-taken quickly and unreflectively and then *cannot be revised or discarded*. It is only that these propositions are up-taken, and then revised in light of other beliefs or counterevidence. The participants in the studies know that they are in a virtual reality environment, and, presuming that they are not what Chalmers calls a "naïve user", the "background knowledge helps *orient* one to the perceived world, giving a global interpretation to what is perceived" (Chalmers, 2017, p. 330). Chalmers identifies this as *cognitive orientation*, and claims that non-naïve users will "act in ways that turn on interpreting themselves to be in VR" (2017, p. 331). So we should expect any propositions up-taken concerning endorsing the veridicality of what is

perceived – which affect or even replace the attitudes constituting implicit bias – to be fairly swiftly revised (given that these users are not naïve).

Perhaps it could be claimed that the Spinozan model predicts that attitudes formed in the virtual reality environment will not change during the experiment, after all, there is a constant input supporting them. This failure to revise beliefs or patchily endorsed propositions like *I have a Black body* perhaps could, on a Spinozan model, last long enough to get the participants to the post-virtual reality environment measures of implicit bias, which occur immediately after the virtual reality immersion. It is of course obvious that an effect might be maintained even if the cause is no longer present, and the propositionalist might well lean on this truism. However, it must be remembered that whatever is said here, also needs to be plausible for *up to a week*, since Banakou and colleagues' found that the positive (mitigation) effects of virtual reality on implicit biases were maintained a full week after immersion. If the Spinozan view can allow that propositional contents incongruent with biases are up-taken even in virtual reality environments the subjects know not to be real, and that the resulting attitudes are not quickly revised, it also needs to accommodate those attitudes not being revised *after a full week* in a normal, non-virtual environment. Once out of the laboratory environment, it is not the case that the participants continued to perceive, represent, and thus up-take propositions incongruent with their implicit biases, since they were not constantly exposed to perceptual information which might suggest to uncritical Spinozan mechanisms that they, for example, had a Black body. Rather, relevant counter-causes are present, i.e., those in the environment which produced the implicit biases in the first place.

Perhaps the propositionalist could say that the immersion in virtual reality was temporarily curative, such that a full week in a normal environment was not enough to generate back the implicit biases. However, this is implausible for a view which takes implicit biases to be formed and updated in a Spinozan fashion. In general, when it comes to implicit bias mitigation, the disappointing effects of real-world counter-causes in mitigation efforts is a well-recognized phenomenon. For example, Nilanjana Dasgupta notes that even techniques which reduce bias do so only in the short-term, since biases "will reflect whatever local environments [people] are chronically immersed in" (Dasgupta, 2013, p. 271). Indeed, "their very presence hints at their being not only generated but also maintained by culture" (FitzGerald et al., 2019, p. 9). Such observations have motivated some theorists to advocate for the prioritization of *structural change* in our bid to mitigate implicit bias. As Sally Haslanger has argued, if implicit biases come about from one's presence within certain social structures, then so long as such structures are maintained, it is a waste of time for individuals within those structures to correct for implicit bias (Haslanger 2015, p. 8).

This general observation about how implicit biases are produced is especially intractable for the propositionalist interested in mitigation, given that they think of implicit bias in Spinozan terms, that is, as being the kind of mental item which comes about merely from representing a truth-apt proposition.

Propositional models have it that implicit biases are highly malleable. Of course, malleability does not entail that they cannot persist; just because they can change doesn't in fact mean they will. But we'd need a reason to think that Spinozan mechanisms were behaving differently in this context, and that these attitudes can stand the test of time. Why think that? A case needs to be made that it is consistent with the Spinozan position that the attitudes formed in virtual reality are not revised. So even if the propositionalist can allow that propositions can be up-taken in virtual reality environments, and that the *right kinds* of propositions can be so up-taken, she needs to explain why those biases change long-term once the person is out of the laboratory.[10]

A final thing the Spinozan might do is to draw attention to the asymmetry between accepting a proposition (automatic) and rejecting it (effortful). Given that, leaving the laboratory environment might not be enough to get rid of the bias-mitigating beliefs or patchy endorsements formed in the virtual reality environment. If that's right, longevity of the new attitudes isn't a problem. However, that would only solve the problem if the week away from the virtual reality environment were spent in isolation, rather than back out in the world where opposing representations abound. It is one thing for these newly formed attitudes to not be *overwritten* or *rejected* (and for folk to hold contradictory attitudes, a prediction endorsed by the Spinzoan account of belief (Mandelbaum, 2014, p. 63)), it is quite another for those newly formed attitudes to shine through in an IAT a week after they were formed.

I mentioned at the start of this section that this issue is more severe for Mandelbaum's view than Levy's. That is because Levy takes it that patchy endorsements are limited in their responsiveness to evidence (in this case contrary representations). There is space in his account to identify virtual reality environments as ones apt to produce (the right kind of) propositional representations which interact with implicit bias, but then to say that the post-virtual reality normal environment is not one which we should expect to afford further interaction to undo the mitigation. However, that would be surprising. Presumably implicit biases arise in the first place due to our interactions with the real world. If the patchy endorsement theorist wants to accommodate longstanding mitigation effects by limiting the contexts in which these attitudes are formed or revised, it would then be difficult to retain a fairly natural story about why we have these attitudes in the first place.

To sum up: propositional models of bias face an initial problem in making a case for a virtual reality environment being proposition-apt; that is, being the kind of environment where propositions are there for the up-taking. If this can be overcome, the propositionalist would need to explain (i) why such propositions as *I have a virtual Black body* interact with implicit bias, or (ii) how propositions like *I am Black*, which could plausibly interact with implicit bias, get represented. If it can be said of users of virtual reality that they form attitudes which could increase or decrease their implicit biases, another problem arises, namely these attitudes (Spinozan as they are) cannot be expected to stick around. On the assumption that participants in these experiments are not naïve users of virtual reality, any unconsciously formed attitudes which come about from merely representing some propositional content while immersed, ought to be swiftly revised. Even if that charge does not stick (perhaps the Spinozan can say plausible things about why unconscious attitudes formed in a virtual reality environment can last until their implicit biases are re-measured), the propositionalist about implicit bias cannot explain why mitigation effects last a *full week* after immersion in the virtual reality environment.

I turn now to my model of bias, which I briefly overview before arguing that it is well placed to explain the data from virtual reality studies, and does not face the three problems of purely propositional models just outlined. Empirical work on virtual reality and immersion and implicit bias, then, allows us to formulate a new argument in favor of the imagination model.

5. Imagining our biases

My view has it that implicit biases are *constituted by* unconscious imaginings. In place of a robust account of what the imagination or its products are, I appeal to three features of imagination upon which there is "wide agreement" (Kind, 2016, p. 1). First, it is a primitive mental state, which is to say that it is irreducible to other mental states (cf. Langland-Hassan, 2012) like for example *perceiving, believing*, or *remembering*. Second, imaginings have representational content, that is, there is something which they are *about*. Third, imaginings are not connected to truth in the manner of, for example, belief[11] (Kind, 2016, pp. 1–3). Of course these features I have identified are not exhaustive of imagination or its functional role, but to add to the characterization would be to enter controversial waters in which I need not wade for the purposes of outlining my account. Rather, I pick out these three features to signal that I am signed up to a standard conception of the imagination.

I also distinguish two kinds of imagining on grounds of content: propositional imaginings and imagistic imaginings. The former have propositional contents (*there is a unicorn*) whilst the latter have imagistic contents (a

mental image of a unicorn). Of course, these kinds can have overlapping members (e.g., some propositional imaginings might involve mental imagery, Nanay, 2016, p. 132, n. 1).[12] *Unconscious imaginings* then are simply states with the three features upon which there is wide agreement, which can have propositional or imagistic contents, and which are tokened in a way as to be not available to introspection.

Let me quickly address a concern that may have occurred to the reader: the notion of imagination I have appealed to – in particular its being tokened unconsciously – is sufficiently revisionary as to increase the costs of buying into the model, even if it turns out that it is best placed to accommodate the empirical evidence from virtual reality studies. In reply I note that this is not a revisionary notion of imagination after all. The three features upon which there is wide agreement are neutral with respect to whether imaginings can be tokened unconsciously. If unconscious imagination does represent a departure from a standard view, that departure is not to be found in these three uncontroversial features. Another key thing to note: the idea of unconscious imagination is not mine, but has recently gained some currency (see e.g., Church, 2008, 2016; Goldman, 2006; Spaulding, 2016; Van Leeuwen, 2011). Elsewhere I more fully defend the claim that allowing for imaginings to be tokened unconsciously is not to endorse a revisionary notion of the imagination (Sullivan Bissett, 2019, §5). For now, let us go forward and see the work this state can do in an account of implicit bias.

My account has it that implicit biases are constituted by unconscious imaginings. One of the key differences between my model and previous ones defended in the literature, is that it can accommodate wide-ranging heterogeneity. As noted earlier (§2,) although many theorists have wanted to recognize heterogeneity, all extant accounts of the nature of implicit bias fall squarely into either associationism or propositionalism, and any heterogeneity posited remains within the boundaries of these respective frameworks. My account seeks to accommodate heterogeneity at the level of structure (i.e propositional – vs – associative) and content (propositional – vs- imagistic). That is, implicit biases can operate associatively (with two imaginings being associatively linked) or propositionally (with a single imagining). Implicit biases can also have propositional contents or imagistic contents. Importantly, the model is not one which employs strategic imprecision to prevent falsification, rather, it is constructed in the pursuit of extensional adequacy, and empirical evidence suggests heterogeneity of this kind. We will see later that the heterogeneity recognized by the imagination account is key to its success in accommodating the empirical work on virtual reality immersion and implicit bias.

It might be wondered at this point why I am seeking to model implicit bias on a folk psychological construct, rather than simply argue that the

empirical work on virtual reality immersion shows us that we need a pluralist approach with respect to implicit biases coming in both associative and propositional flavors. I note a couple of things here. First, merely getting to an account which includes biases being associatively and non-associatively structured is not going to be as explanatorily powerful as it otherwise might. For example, implicit biases having propositional contents may well explain a few things about their behavior (e.g., their participation in inference, as argued by Mandelbaum, 2016, pp. 636–7, 640), but it might not explain certain other features (e.g., their relationship to behavior), in a way which our existing knowledge of the functional role of some psychological constructs might. Furthermore, as we will see, mental imagery plays a role in some of my discussion of what is going on in virtual reality mitigation, and imagination is of course a natural vehicle for mental imagery. As the very least then, what we learn from virtual reality mitigation is that implicit biases ought to be modeled in a pluralist way, but in what follows I'll argue that imagination seems to be a good bet as to the vehicle, given the possible explanatory gains to be had by appeal to it.

I will now run through an example to see the various ways my model allows for implicit biases to be constituted. Our starting point is that to have an implicit bias is to unconsciously imagine certain things in response to stimuli. Let us consider an implicit bias regarding *women* and *weakness*. For biases structured associatively, the constituents of bias are associatively linked and do not stand in determinate syntactic relations. Against such a background, one of three things could be going on in the presence of certain stimuli, say, a woman. The first way of understanding implicit bias on my view is as associatively linked unconscious imagistic imaginings (i.e., an unconscious imagistic imagining of *woman* and an unconscious imagistic imagining of *weakness*) (as Toribio points out, understanding implicit biases as associations is consistent with thinking of the associated mental constructs as images (Toribio, 2018, p. 42)). Alternatively implicit biases could be understood as associatively linked propositional imaginings (i.e., an unconscious propositional imagining with the content *there is a woman* and an unconscious propositional imagining with the content *there is weakness*). Finally, were we in the realm of a more generalized negative bias against women (i.e., an affective rather than semantic bias) we could have an unconscious imagining (with either imagistic or propositional content) associatively linked with a negative valence.[13]

As we saw earlier, there has been a recent move to modeling implicit bias as non-associative, and empirical work suggesting that this is required, in at least some cases. If that is right, we should make room in our theory of implicit bias to understand the constituents of implicit bias non-associatively. Staying with the same example of an implicit bias regarding *women* and *weakness*, there are two ways my imagination model can capture

what form implicit biases could take against a non-associative background when presented with certain stimuli, say, a woman. A subject could have an unconscious imagistic imagining of a *weak women*, or an unconscious propositional imagining that *women are weak*. In the first case we have a single imagistic imagining (rather than an association between two such imaginings), and in the second case we have a single propositional imagining (rather than an association between two such imaginings). This last way of understanding the possible structure of implicit bias is where Mandelbaum's, Levy's, and my models look very similar and may share predictions.

My account then honors the heterogeneity within the category of implicit bias with respect to its structure (associative vs non-associative). It is uniquely placed to do so since other models take implicit bias to be *either* associative *or* non-associative. It also introduces an additional heterogeneity at the level of contents (propositional vs imagistic). It might be that further work could give us a more particular carving of the category along the lines of which kind of imaginings and processes are in play along the various subcategories. Another way to think about the contribution to the debate made by the imagination model is in terms of what it says about the relata and the relations of implicit bias. On propositionalism, the relata are propositional attitudes (beliefs, patchy endorsements) and the relations they enter into are logical/inferential. On associationism, the relata are concepts or valences, and the relation between them is associative. On the imagination model, heterogeneity is recognized with respect to both the relata (kinds of imaginings, valences) and the relations (propositional and associative processing).

I return now to the empirical work on virtual reality and implicit bias and argue that the imagination model can accommodate their results. As I have already mentioned, the heterogeneity recognized by the imagination model is key to its theoretical success *qua a model of implicit bias* more generally. What the model adds to this discussion in particular, and the reason it will be seen as friendly to the work on virtual reality, is the idea that some implicit biases have imagistic contents. We have seen reasons to think that purely propositional understandings of implicit bias will have a hard time accommodating the results of virtual reality studies. If, however, the implicit biases in play in these studies are constituted by imagistic unconscious imaginings, the explanation of bias change can be had.[14] In what follows I argue that this is indeed how we should understand what is going on in this context.

6. Virtually imagining our biases

I suggest that the illusion of body ownership prompted by virtual reality immersion is, broadly, the mechanism via which our implicit biases are affected. There is substantial evidence that the body ownership illusion is dependent on – or at least strengthened by – synchrony of various kinds.[15] To give a handful of representative examples from the literature, Valeria Petkova and H. Henrik Ehrsson had participants wear a head mounted display showing a video feed of a mannequin's point of view looking down at its body. A short rod was used to stroke the participant's abdomen (unseen by participant) in synchrony with the same strokes applied to the mannequin's abdomen. The control condition had the strokes administered asynchronously. In the visuotactile synchrony group only, participants reporting feeling like the mannequin's body was their own (Petkova & Ehrsson, 2008, pp. 2–3).

Another example comes from Lara Maister and colleagues who conducted two studies using the rubber hand paradigm. In the first, participants were in the synchronous or asynchronous visuotactile condition. In the former condition the rubber hand and the participant's own hand were stroked simultaneously in the same place, and in the latter condition the stimulation of the participant's hand and the rubber hand were offset by 180 degrees. Maister and colleagues found a significant difference in ownership of the hand between the synchronous and asynchronous conditions (with it being higher in the former (Maister et al., 2013, p. 174)). In the second experiment, there were four conditions: synchronous and asynchronous visuotactile for both dark-skinned and light-skinned rubber hands. Again, ownership scores were higher in the synchronous conditions, and there was no effect of hand color (Maister et al., 2013, p. 175). They also found that those in the dark-skinned synchronous visuotactile condition had more positive implicit attitudes toward Black people following the experiment.[16] In addition, strength of body ownership correlated with positive implicit attitudes toward Black people for those in the dark-skin conditions, with Maister and colleagues noting that "[c]hanges in body-representation may therefore constitute a core, previously unexplored, dimension that in turn changes social cognition processes" (2013, p. 176).

Returning to the studies overviewed earlier, recall that Peck and colleagues found that scores on body ownership were much lower in the condition in which participants were not embodied (i.e., no visuomotor synchrony) than in any of the embodied conditions (within which there were no significant differences) (Peck et al., 2013, p. 783). Banakou and colleagues did not have a control group with respect to visuomotor synchrony but reported that participants "tended to affirm the virtual body as their own" (Banakou et al., 2016, p. 5), and they found a reduction in bias for

those embodied in Black avatars. Groom and colleagues did not consider the question of body ownership, and they did not employ visuomotor synchrony, which, as above, is strongly related to body ownership. Their study found that implicit biases got worse. Lopez and colleagues measured body ownership but there was no asynchronous control group, and again implicit biases got worse. As we have seen, explanations for these latter results rest on the idea of stereotype activation triggered by the particular context, which, in Lopez and colleagues' study at least, might have trumped any positive effects of body ownership.

Banakou and colleagues suggest that body ownership leads to updates to the "multisensory representation of peripersonal space", but it also leads to "corresponding psychological updates", for example, changes to implicit bias (Banakou et al., 2016, p. 9). The imagination model integrates nicely with this idea. Consider first the studies which showed implicit bias mitigation following virtual reality immersion with visuomotor synchrony and reports of body ownership. When embodied in a Black avatar, an experimental participant may engage in certain imaginative activities, such activities may then make less accessible opposing cognitive, emotional, and behavioral representations in line with the unconscious imaginings which constitutes their implicit racial bias. One mechanism via which this might be achieved is suggested by Maister et al. (2015, p. 6). They say that the increase in perceived similarity between oneself and an outgroup member can lead to the generalization of positive self-like associations. In our terms, the suggestion is that generalizations of particular positive unconscious imaginings may be associatively linked with unconscious imaginings of outgroup members. Understood non-associatively, we might say that single unconscious imaginings prompted by body ownership make more accessible positive representations concerning members of certain groups.

It's possible that these imaginings are *constitutive of* body ownership, that is, these imaginings are prompted and that *just is* what it means to have a sense of body ownership. That would be consistent with the findings that ownership is reduced but not eliminated in contexts without synchrony but with stimuli which could nevertheless prompt certain kinds of imaginings. However, body ownership could equally precede these imaginings, be caused by them, prompt them, as well as be constituted by them. That does not matter for my account. The key point is that feeling ownership over a virtual body can be seen as a proxy for imaginative activities which interact with – and are constitutive of – implicit biases.

Let us turn to the studies which showed implicit bias increase following virtual reality immersion. In Groom and colleagues' study with limited visuomotor synchrony (covering only head movements), there are two things that might be going on. To explain the failure to mitigate implicit bias, we might suggest that the limited synchrony was insufficient to prompt

the body ownership illusion and the relevant imaginative states which might make less accessible opposing cognitive, emotional, and behavioral representations in line with the unconscious imaginings which constitute our implicit racial bias. To account for implicit biases getting worse, we can appeal to the context of the virtual reality immersion. This was one which is negatively stereotyped for Black people. We might expect then that such a context would prompt unconscious imaginings regarding the context of a job interview (a similar explanation can be had for Schulze and colleagues 2019 study, see n. 3). In Lopez and colleagues' study, where there was visuomotor synchrony which I suggest prompted the body ownership, we can say that unconscious imaginings concerning the incompetence of the avatar and the sports context (negatively stereotyped for gender) may have been triggered, overwhelming the would-be positive effects of body ownership. In both cases, negatively valenced (affective and content) imaginings may have been made more accessible (either associatively or non-associatively), strengthening the implicit biases regarding these groups.

Earlier I argued that there are three issues which face propositional models with respect to accommodating the results from virtual reality and implicit bias mitigation (§4). I now explain why my imagination model does not face these problems, and this gives us a reason to prefer it. I take the first two together, which are answered in the same way, before turning to the third.

6.1. Problems one and two: (which) propositional representations?

The first two problems can be summarized thus: is virtual reality the kind of environment where propositional representations occur, which can be automatically up-taken to belief or patchy endorsement? If it is, the propositionalist would need to explain either (i) why such propositions as *I have a virtual Black body* interact with implicit bias, or (ii) how propositions like *I am Black*, which could plausibly interact with implicit bias, get represented.

These problems are easily overcome by my view. The imagination model need not worry about virtual reality environments giving rise to propositional representations, nor specifying which propositions are up-taken in a way which would affect implicit bias. This is because some implicit biases are constituted by unconscious imagery devoid of propositional content.

This might feel like a bit of a cheat: although the imagination model *can* just fall back on some of its candidate constituents of bias, and say that the biases involved in cases of virtual reality mitigation are imagistic ones, are there any positive reasons to say this beyond it conveniently circumventing a problem for wholly proposition-based approaches? I think there are at

least two things to say which demonstrate that my appeal to imagistic imaginings in this context is not an ad hoc and theory-saving move.

Firstly, understanding implicit biases as so constituted in this context might gain its plausibility from reflection on the fact that virtual reality immersion is a highly visual phenomenon. As I have suggested, body ownership is key to implicit bias mitigation, and this is itself dependent on, or at least strengthened by, synchrony of various kinds. In all the studies in which this claim is made good on the synchrony includes the visual. If implicit bias mitigation is indirectly contingent on the presence or absence of visuomotor synchrony, we have reason to think that there's something imagistic at play.

The second thing to say in defense of appeal to non-propositionally constituted implicit biases comes from a comparison with two other kinds of bias mitigation technique: mental imagery exercises and engaging with video content. An example of the first comes from Irene Blair and colleagues' work on mental imagery, described as "the conscious and intentional act of creating a representation of a person, object, or event by seeing it with the 'mind's eye'" (Blair et al., 2001, p. 828). In one of five experiments, participants were either in the counterstereotype group, or the neutral (control group). In the first, participants were asked to imagine a strong woman, in the second, participants were asked to imagine a holiday. Blair and colleagues found that those in the counterstereotype group "produced a significantly lower level of the implicit stereotype than the participants who imagined a neutral event" (Blair et al., 2001, p. 831). In discussion they say that imagining a counterstereotypical exemplar "reduced the implicit stereotype by more than half, providing the first demonstration that mental imagery can have a powerful effect on implicit processes" (Blair et al., 2001, p. 831).[17]

In her discussion of these mitigation techniques, Natalie Salmanowitz suggests that the individual differences in imaginative capacity in the mental imagery exercises might explain why these exercises do not produce consistent results (Salmanowitz, 2016, p. 139). She further notes that watching videos as a way to mitigate bias meets the challenge of differences in imaginative capacity (because all participants are exposed to the same content), but that such an activity lacks the interactive component. Salmanowitz suggests that virtual reality "could simultaneously harness the benefits of mental imagery techniques and videos while circumventing their shortcomings" (Salmanowitz, 2016, p. 139). The idea then is that virtual reality immersion offers us the best of both worlds: sameness of content for all participants and interactivity.

So if we know that mental imagery exercises and exposure to certain video content individually mitigate implicit bias, and that we can understand virtual reality immersion as a mix of these two techniques, with

mental imagery constituting some cases of implicit bias, the success of virtual reality in mitigating it is what we should expect.

6.2. Problem three: longevity of effects

The third issue for propositional models concerned longevity of implicit bias mitigation. If problems one and two can be overcome, and adopting a Spinozan approach to propositional representations could allow for the formation of attitudes incongruent with implicit biases, I suggested that such attitudes ought to have been revised. Why is it that the attitudes formed in a virtual reality environment last for at least a week? (As must be said given that Banakou and colleagues found that a reduction in bias was maintained a full week after exposure to the virtual reality environment.) It would need to be the case that a full week in a normal environment was not enough to re-generate the previous levels of bias, but this is implausible given what we know about the quick generation of Spinozan attitudes. As I noted earlier, even though it is a well-observed fact that implicit bias mitigation is fairly short-lived, the problem here for the propositionalist is not based merely on that observation, but on that observation *paired with* the idea that mechanisms for implicit bias formation and updating are Spinozan. The problem then, is compounded.

The imagination model I endorse does not face a problem here. It is no part of unconscious imagination that every proposition or image represented is *thus* imagined (compare with e.g., the Spinozan theory of belief). So from the outset we can say there is no reason to expect the changes to implicit biases not to last *on the grounds that* contrary representations abound in the non-virtual environment. Of course, it could well be that unconscious *propositional* imagining is also Spinozan, indeed, I find this a plausible claim.[18] If that is right, we may well not expect implicit biases so constituted to be mitigated for very long. But we have already learned of some reasons to be skeptical that propositional bias is in play in the virtual reality context, which suggests that it is not implicit biases with propositional contents that are mitigated as a result of virtual reality immersion. If that is right, my model of bias can explain what is going on.

7. Propositionalism via imagining

Even if everything I have said in the previous sections stands, there is nevertheless a way that pure propositionalist models might try to accommodate the results of virtual reality studies, via imagination. As Mandelbaum says in his overview of his Spinozan account of belief, "[p]eople do not have the ability to contemplate propositions that arise in the mind, *whether through perception or imagination*, before

believing them" (Mandelbaum, 2013, p. 61, my emphasis). The propositionalist then might help herself to what I've said about imaginings being the vehicle for implicit bias change in the virtual reality context, whilst continuing to characterize implicit biases as Spinozan beliefs or patchy endorsements. The story might go like this: the imagination is involved in roughly the way I describe, but that's not the whole story; it is rather the stimulus which affects the formation and revision of unconscious beliefs or patchy endorsements. So the propositionalist need not say that there's some propositional content that's affecting the implicit bias, the imagistic imagining can do that work instead. That would solve the problems of specifying appropriate propositions (depending on how we understood the imagistic imaginary content), and would solve the longevity problem because unconscious imagistic imagining need not be Spinozan.

In response I note that the propositionalist would need to make a case for the relationship between unconscious imagery and belief being relevantly similar to the relationship between ordinary cases of perception and belief.[19] If the idea is that unconscious imagery can be the stimulus for unconscious proposition up-taking in a way parallel to how perceptual stimuli affect belief formation, that this occurs as such cannot be presumed. In addition, we would have a more complicated story on our hands: the propositionalist would accept all of the explanatory work done by imagining in this context, and instead of letting it do all of the explanatory lifting, she would add in Spinozan belief or patchy endorsement. As far as I can see, this additional component in our account of implicit bias and virtual reality immersion is one which does not pull its explanatory weight.

8. Concluding remarks

I began by noting that there has been a recent move to modeling implicit biases propositionally, a move motivated by empirical work suggesting that associationism could not accommodate certain empirical data. Pure associationism then, is inadequate. Thus my starting point was propositionalism. However, recent work on virtual reality immersion has shown implicit bias change as a result of embodiment in a virtual avatar, something that propositionalism is unable to explain, for three reasons. First, the distinctive phenomenology of virtuality may prevent Spinozan up-taking. Second, the kinds of representations there for the up-taking in the virtual reality environment are not the kinds of representations we should expect to mitigate bias. On the other hand, those representations we might expect to mitigate bias are ones we shouldn't expect to arise in a virtual reality environment. Third, even if the propositionalist can explain the mitigation, they cannot explain why that mitigation lasts for at least a week (given their background

endorsement of Spinozan updating). Pure propositionalism then, is also inadequate.

I argued then that we needed a hybrid view, one which recognized heterogeneity at the level of structure. The idea that implicit bias as a class should recognize heterogeneity is not new, although previous calls for such recognition have taken place within the confines of associationism, with the request being one of recognizing *different kinds of association*. I suggested that in light of the previously argued inadequacies of pure associationism, and the inadequacies of pure propositionalism revealed by virtual reality immersion mitigation, we needed a broader heterogeneity (that which recognized implicit biases in associative and propositional flavors). I argued that my preferred view of implicit biases as constituted by unconscious imaginings could accommodate such heterogeneity.

Another feature of my account which helps us in the particular context of virtual reality is that it is able to recognize a different kind of heterogeneity, one at the level of *contents*. So not only might implicit biases be structured associatively or non-associatively, so too might they have propositional or imagistic contents (something we can recognize if implicit biases are constituted by unconscious imaginings). I argued that the key to understanding implicit bias mitigation might be via implicit biases as imagistic, since a key predictor of mitigation was the body ownership illusion which was strongly mediated by visuomotor synchrony. If I am right that it is implicit biases with imagistic contents which are affected as a result of virtual reality immersion, then a surprising outcome of the work here is that we might better understand the nature and behavior of some implicit biases by better understanding the nature and behavior of mental imagery. That is, in the pursuit of implicit bias mitigation, we need to think really carefully about how unconscious imagery works, i.e., the conditions under which it is activated and affected. Suppose I am right that unconscious imagery is at least sometimes the mode of content implicit biases have, and suppose that it can stick around, and is not always lost from counter-conditioning or representation of new propositional contents. We would thus do well to investigate how it behaves, in the service of better understanding how to mitigate (at least some) implicit bias.

To conclude then: my account of implicit bias as constituted by unconscious imagination, in its recognition of wide-ranging heterogeneity in the set of mental constructs we capture as *implicit bias*, can accommodate both the recent data speaking against pure associationism, and also the data from studies on implicit bias and virtual reality. If that's right, we have reason to prefer it over its competitors, and learn also that further research on the role of mental imagery in some implicit bias might be key to the development of mitigation strategies.

Notes

1. Elsewhere (Sullivan Bissett, 2019) I argue that my preferred account can do this explanatory work. I won't repeat the details here but instead will focus on how this account is best placed to accommodate the results from empirical work on virtual reality immersion and implicit bias mitigation.
2. IAT results will be interpreted differently by associationists and propositionalists, and of course such theorists will also say different things about how implicit biases cause behavior more broadly. Take the relevant behavior to be a person's faster pairing of *women* to *arts subjects* than to *STEM subjects* in an IAT. An associationist might say that the person's implicit bias consists of an association between the concepts *women* and *arts*, and so upon seeing the woman stimuli, arts-related concepts are made more accessible. A propositionalist might say that the implicit bias consists of a representation with propositional content like *women are best suited to arts subjects*. The faster pairing of women stimuli with arts stimuli is explained by the person taking there to be a determinate relationship between these two ideas (see Levy, 2015, p. 805 for more).
3. A similar study was run by Stephanie Schulze and colleagues (2019). The experimenters were interested in the effect of gendered embodiment on implicit gender bias, and participants completed an IAT on gender and leadership before and after the embodiment experience. The virtual environment was a manager's office and male and female participants were embodied in either a male or female avatar (and so there were four experimental conditions). Following an orientation period, Caucasian virtual men and women came in and out of the office. Although the experimenters note that on average, participants did feel as though the virtual body was their own, scores of body ownership were lower here than in Peck's study (they speculate that less time embodied may explain this result, 2019, p. 373). In all but the female participant with male avatar condition, there was an increase in implicit gender bias against women after the embodiment. Schulze and colleagues suggest that the context of the embodiment (a manager's office) may have been responsible for the result (and they draw a comparison with the results from Groom et al. (2009) study on race in an interview context) (Schulze et al., 2019, p. 373).
4. Mandelbaum uses the term "implicit bias" to pick out *biased behavior* caused by an implicit attitude, and so strictly speaking his view is that implicit attitudes are unconscious beliefs. This is merely terminological and for ease of expression I have put his account in terms of it picking out the nature of "implicit bias" understood as a mental item.
5. If the Spinozan does want to claim something like this, the next problem I raise will nevertheless stand.
6. This need not be dependent on a certain irrealism about what is presented in experience – even if virtual environments were indistinguishable from their correlates in reality, a distinctive phenomenology could arise due to non-content related features. For example, Mohan Matthen argues that normal scene vision is "actuality committing" whilst seeing the same scene in a picture is not, with the latter involving a *feeling of presence* (Matthen, 2010, p. 114). Virtual reality might be more like normal scene vision in this respect, but the distinctive phenomenology might arise from the discordance between some approximation of a feeling of presence and background beliefs about one's actual location. Alternatively, Jerome Dokic and Jean-Remy Martin argue that we can experience a sense of reality without genuine perceptual experience, and suggest that in virtual reality environments we experience a genuine

sense of presence absent the sensory content present in ordinary perceptual conditions (Dokic & Martin, 2017, p. 302). This too could be part of the story for the phenomenology distinctive of virtual reality immersion.

7. It might be said that there is an unargued assumption in my point here, namely, that the propositionalist is committed to the idea that biases can be changed only by other *beliefs*. My discussion though is in line with the Spinozan account in the background, which has it that any proposition *represented* is thus believed. There is no gap to exploit between representation and belief, and so although we might equally talk in terms of mere *representations* mitigating bias, given the Spinozan background, that is no less committal than talking in terms of beliefs mitigating bias. Perhaps I might be pushed again: perhaps bias can be changed by *perceptions* rather than represented propositions. But as has been noted, non-naïve users of virtual reality will perceive their surroundings *as virtual*. A story would need to be told concerning how perceiving things *as such* could interact with bias.
8. The parenthesized material here is a nod to Levy's claim that patchy endorsements partake in only some kinds of inference (Levy, 2015, p. 816). I assume for the sake of argument that the virtual reality context would afford the right type, without that assumption it is unclear what the explanatory story would be for implicit bias change on the patchy endorsement model.
9. Alternative contents might be noted here. I have said that the most plausible propositions to arise would have contents tagging the objects perceived *as virtual*, and would thus not be propositions which we could expect to affect propositionally structured implicit biases. I also said that those propositions which could plausibly affect bias would not arise in the virtual reality environment. Are there any contents which might arise *and* influence bias? One suggestion might be *I am not so different from Black people* or *this is what it is like to be Black*. However, given that the subject perceives everything *as virtual*, in what sense is it the case that they are not so different from Black people? Or, in what sense is the case that *this* is what it is like to be Black? I think the closest thing that might arise is in fact *I'm not so [virtually] different from Black people*, or *this is what it is like to be [virtually] Black*. But these are not contents which we should expect to influence bias. I am grateful to Katherine Puddifoot for discussion.
10. Perhaps the propositionalist could say that memory of the virtual reality intervention might produce the right kinds of beliefs even after the intervention. I note two things in response. First, memories of an intervention may well be expected to affect bias, of course. But this possibility only lands us back on a version of the second problem: that which is remembered is either something not plausibly represented in the virtual environment (*I am Black*), or it is plausibly represented but not the kind of proposition we should expect to interact with bias (*I am virtually Black*). Second, if we allow memories of representations to mitigate bias, we should also allow for the fact that memories of everyday life (awash with the kinds of things which entrench bias in the first place) would count against such mitigation.
11. Kind understands this third feature as specifying the absence of a *constitutive* connection to truth. Elsewhere (Sullivan Bissett, 2017, 2018) I defend a contingent relationship between belief and truth, and so I have dropped the "constitutive". Whatever one makes of the strength of the relationship, the point is that belief is connected to truth in a way that imagining is not.
12. There is a question about whether there could be *purely* propositional imaginings, i.e., imaginings without imagery. It is beyond the scope of this paper to discuss this issue,

and so I will assume that purely propositional imaginings are possible (this is an approach I have defended elsewhere, see Sullivan Bissett, 2019).

13. There is no reason to rule out at this stage associations between different kinds of mental items (i.e., an imagistic imagining and a propositional imagining), but it is unclear to me what the empirical evidence would have to look like to motivate this possibility.
14. Of course, it might turn out that purely associative accounts could accommodate the empirical results in the same way – that is, by having the relevant relata as unconscious imagery standing in associative relationships. That may be so, but as I noted at the start of the paper, I take the work here to be downstream of the question of whether a purely associative account is adequate, taking my leave from recent work suggesting an answer in the negative.
15. It has been found that visuo*motor* synchrony is more likely to lead to the ownership illusion than e.g., visuo*tactile* synchrony, but when the synchronies are combined, the cessation of either kind equally lead to the loss of the illusion (Kokkinara & Slater, 2014). Schulze and colleagues also found that whilst a first-person perspective from the avatar was the most important factor for creating body ownership, visuomotor synchronicity was also key (Schulze et al., 2019, p. 362).
16. This might also be an empirical result which propositional models will have difficulty accommodating. I am grateful to Dan Cavedon-Taylor for suggesting looking at work on the rubber hand illusion.
17. These results are easily accommodated by the imagination model. We can be caused to imagine all sorts of things by the sexist, racist, and heteronormative culture many of us inhabit (Dasgupta, 2013, p. 240). When engaging in imaginative activities, like imagining a counterstereotypical exemplar, the existence of, or effects of, these implicit biases change. Indeed, Blair and colleagues note in their explication of mental imagery that it increases the "accessibility of related cognitive, emotional, and behavioral representations" (Blair et al., 2001, p. 829). In its doing so, it could make less accessible opposing cognitive, emotional, and behavioral representations in line with the unconscious imaginings, that is, in line with the target implicit bias.
18. As already noted, the propositionalist's case has rested, at least in part, on studies which show that implicit biases are highly malleable and formed very quickly. This is problematic in light of the longevity of bias change. My sense is that propositionally structured biases are indeed highly malleable, but imagistic biases might not be. If we have an account of implicit bias on which at least some have imagistic contents, those studies which show that some implicit biases have propositional contents and are highly malleable are consistent with some biases being otherwise constituted and having different conditions for formation and extinction. For theorists who take the set of implicit biases to be all or mostly propositional, they are tied to malleability in a way which makes longevity difficult to accommodate.
19. I write this response in terms of belief because patchy endorsement is a sui generis relatively recently posited state, and so we don't have an ordinary case with which to compare.

Acknowledgements

I acknowledge the support of the Arts and Humanities Research Council (*Deluded by Experience*, grant no. AH/T013486/1). I'm grateful to Neil McDonnell who invited me to speak at a workshop on Autonomy and Immersive Technology – it was in anticipation of

that event that I wrote this paper. Many thanks to audiences at the Open University, Queen's University Belfast, and the University of Durham. I am grateful to Chiara Brozzo, Dan Cavedon-Taylor, Josh DiPaolo, Anna Ichino, Louise Richardson, Michael Rush, Katherine Puddifoot, and Nathan Wildman for helpful comments on earlier versions of this paper. Finally, thank you to two reviewers for this journal for their comments which improved the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Arts and Humanities Research Council [AH/T013486/1]

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661. https://doi.org/10.1037/0022-3514.91.4.652

Banakou, D., Hanumanthiu, P. D., & Slater, M. (2016). Virtual Embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Frontiers in Human Neuroscience*, *10*, 1–12. https://doi.org/10.3389/fnhum.2016.00601

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition*, *18*(4), 329–353. https://doi.org/10.1521/soco.2000.18.4.329

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841. https://doi.org/10.1037/0022-3514.81.5.828

Brinol, P., Petty, R., & McCaslin, M. (2009). Changing Attitudes on Implicit versus explicit measures: what is the difference? In R. Petty, R. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 285–326). Psychology Press.

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *WIREs Cognitive Science*, *10*(5), e1501. https://doi.org/10.1002/wcs.1501

Chalmers, D. (2017). The virtual and the real. *Disputatio*, *9*(46), 309–352. https://doi.org/10.1515/disp-2017-0009

Chen, M., & Bargh, J. A. (1997). Nonconscious behavioural confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology*, *3*(5), 541–560. https://doi.org/10.1006/jesp.1997.1329

Church, J. (2008). The hidden image: A defense of unconscious imagining and its importance. *American Imago*, *65*(3), 379–404. https://doi.org/10.1353/aim.0.0024

Church, J. (2016). Perceiving people as people: An overlooked role for the imagination. In A. Kind & P. Kung (Eds.), *Knowledge through imagination* (pp. 160–182). Oxford University Press.

Cogburn, J., & Silcox, M. (2014). Against brain-in-a-vatism: On the value of virtual reality. *Philosophy & Technology*, *27*(4), 561–579. https://doi.org/10.1007/s13347-013-0137-4

Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research*, *17*(2), 143–169. https://doi.org/10.1023/B:SORE.0000027407.70241.15

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: a decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology*, *47*, 233–279. https://doi.org/10.1016/B978-0-12-407236-7.00005-X

Del Pinal, G., & Spaulding, S. (2018). Concept centrality and implicit bias. *Mind & Langiage*, *33*(1), 95–111. https://doi.org/10.1111/mila.12166

Diodato, R. (2012). *Aesthetics of the virtual*. State University of New York Press.

Diodato, R. (2014). About virtual experience. some questions. *Metodo International Studies in Phenomenology and Philosophy*, *2*(2), 47–68. https://doi.org/10.19079/metodo.2.2.47

Dokic, J., & Martin, J.R. (2017). Felt reality and the opacity of perception. *Topoi*, *36*(2), 299–309. https://doi.org/10.1007/s11245-015-9327-2

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*(5), 510–540. https://doi.org/10.1006/jesp.1997.1331

Egan, A. (2008). Seeing and believing: Perception, belief formation, and the divided mind. *Philosophical Studies*, *140*(1), 47–63. https://doi.org/10.1007/s11098-008-9225-1

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603

FitzGerald, C., Martin, A., Bemer, D., & Hurst, S. 2019: 'Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review'. *BMC Psychology*. Vol. 7, article no. 29.

Gadsby, S. (2019). Manipulating body representations with virtual reality: Clinical implications for anorexia nervosa. *Philosophical Psychology*, *32*(6), 898–922. https://doi.org/10.1080/09515089.2019.1632425

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 44, pp. 59–127). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology*, *41*(6), 618–626. https://doi.org/10.1016/j.jesp.2004.10.005

Gilbert, D. T., Taforodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Attitudes and Social Cognition*, *65*(2), 221–233. https://doi.org/10.1037/0022-3514.65.2.221

Goldman, A. I. (2006). *Stimulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Groom, V., Bailenson, J. N., & Nass, C. (2009). The influence of racial embodiment on racial bias and immersive virtual environments. *Social Influence*, *4*(3), 231–248. https://doi.org/10.1080/15534510802643750

Haslanger, S. (2015). Distinguished lecture: Social structure, narrative, and explanation. *Canadian Journal of Philosophy*, *45*(1), 1–15. https://doi.org/10.1080/00455091.2015.1019176

Holroyd, J. 2016: 'What do we want from a model of implicit cognition?' *Proceedings of the Aristotelian Society*. (Vol. CXVI, no. 2, pp. 153–179).

Holroyd, J., & Sweetman, J. (2016). The heterogeneity of implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy: Metaphysics and epistemology* (Vol. 1, pp. 80–103). Oxford University Press.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy: metaphysics and epistemology* (Vol. 1, pp. 47–79). Oxford University Press.

Kind, A. (2016). Introduction: Exploring imagination. In A. Kind (Ed.), *The Routledge handbook of philosophy of imagination* (pp. 1–11). Routledge.

Kokkinara, E., & Slater, M. (2014). Measuring the effects through time of the influence of visuomotor and visuotactule synchronous stimulation on a virtual body owndership illusion. *Perception*, *43*(1), 43–58. https://doi.org/10.1068/p7545

Langland-Hassan, P. (2012). Pretense, imagination, and belief: the single attitude theory. *Philosophical Studies*, *159*(2), 155–179. https://doi.org/10.1007/s11098-011-9696-3

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs*, *49* (4), 800–823. https://doi.org/10.1111/nous.12074

Lewis, D. (1982). Logic for equivocators. *Nous*, *16*(3), 431–441. https://doi.org/10.2307/2216219

Lopez, S., Yang, Y., Beltran, K., Kim, S. J., Cruz Hernandez, J., Simran, C., Yang, B., & Yuksel, B. F. 2019. 'Investigating implicit gender bias and embodiment of white males in virtual reality with full body visuomotor synchrony'. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, Paper 557, (pp. 1–12).

Maister, L., Sebanz, N., Knoblich, G., & Tsakiris, M. (2013). Experiencing ownership over a dark-skinned body reduces implicit racial bias. *Cognition*, *128*(2), 170–178. https://doi.org/10.1016/j.cognition.2013.04.002

Maister, L., Slater, M., Sanchez-Vives, M. V., & Tsakiris, M. (2015). Changing bodies changes minds: Owning another body affects social cognition. *Trends in Cognitive Science*, *19*(1), 6–12. https://doi.org/10.1016/j.tics.2014.11.001

Mandelbaum, E. (2013). Against Alief. *Philosophical Studies*, *165*(1), 197–211. https://doi.org/10.1007/s11098-012-9930-7

Mandelbaum, E. (2014). Thinking is believing. *Inquiry*, *57*(1), 55–96. https://doi.org/10.1080/0020174X.2014.858417

Mandelbaum, E. (2016). Attitude, inference, association: on the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658. https://doi.org/10.1111/nous.12089

Matthen, M. (2010). Two visual systems and the feeling of presence. In N. Gangopadhyay, M. Madary, & F. Spicer (Eds.), *Perception, action, and consciousness: Sensorimotor dynamics and two visual systems* (pp. 107–124). Oxford University Press.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the implicit association test, discriminatory behaviour, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*(5), 435–442. https://doi.org/10.1006/jesp.2000.1470

McDonnell, N., & Wildman, N. (2019). Virtual reality: Digital or fictional? *Disputatio*, *11* (55), 371–397. https://doi.org/10.2478/disp-2019-0004

McDonnell, N., & Wildman, N. (2020). The puzzle of virtual theft. *Analysis*, *80*(3), 493–499. https://doi.org/10.1093/analys/anaa005

Meijsing, M. (2006). Real people and virtual bodies: How disembodied can embodiment be? *Minds & Machines*, *16*(4), 443–461. https://doi.org/10.1007/s11023-006-9044-0

Mooradian, N. (2006). Virtual reality, ontology, and value. *Metaphilosophy*, *37*(5), 673–690. https://doi.org/10.1111/j.1467-9973.2006.00460.x

Nanay, B. (2016). Imagination and perception. In A. Kind (Ed.), *The Routledge handbook of philosophy of imagination* (pp. 124–134). Routledge.

Peck, T. C., Seinfeld, S., Aglioti, S. M., & Slater, M. (2013). Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition*, *22*(3), 779–787. https://doi.org/10.1016/j.concog.2013.04.016

Petkova, V. L., & Ehrsson, H. H. (2008). If I were you: perceptual illusion of body swapping. *Plos One*, *3*(12), 1–9. https://doi.org/10.1371/journal.pone.0003832

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Salmanowitz, N. (2016). Unconventional methods for a traditional setting: The use of virtual reality to reduce implicit racial bias in the courtroom. *UNHL Rev*, *15*(1), 117–160.

Schulze, S., Pence, T., Irvine, N., & Guinn, C. 2019: 'The effects of embodiment in virtual reality on implicit gender bias'. In J. Y. C. Chen & G. Fragomeni (Eds.) *Virtual, Augmented and Mixed Reality. Proceedings of 11th Internal Conference, VAMR 2019* (pp. 361–374).

Sechrist, G., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. *Journal of Personality and Social Psychology*, *80*(4), 645–654. https://doi.org/10.1037/0022-3514.80.4.645

Slater, M. (2017). Implicit learning through embodiment in immersive virtual reality. In D. Liu, C. Dede, R. Huang, & J. Richards (Eds.), *Virtual, augmented, and mixed realities in education* (pp. 19–34). Springer.

Spaulding, S. (2016). Simulation Theory. In A. Kind (Ed.), *The Routledge handbook of philosophy of imagination* (pp. 262–273). Routledge Press.

Stalnaker, R. (1984). *Inquiry*. MIT Press.

Sullivan Bissett, E. (2017). Biological function and epistemic normativity. *Philosophical Explorations*, *20*(1), 94–110. https://doi.org/10.1080/13869795.2017.1287296

Sullivan Bissett, E. (2018). Explaining doxastic transparency: Aim, norm, or function? *Synthese*, *195*(8), 3453–3476. https://doi.org/10.1007/s11229-017-1377-0

Sullivan Bissett, E. (2019). Biased by our imaginings. *Mind & Language*, *34*(5), 627–647. https://doi.org/10.1111/mila.12225

Sullivan Bissett, E. (2023). Implicit bias and processing. In R. Thompson (Ed.), *The Routledge handbook of philosophy and implicit cognition* (pp. 115–126). Routledge.

Toribio, J. (2018). Implicit bias: From social structure to representational format. *Theoria*, *33*(1), 41–60. https://doi.org/10.1387/theoria.17751

Van Leeuwen, N. (2011). Imagination is where the action Is. *The Journal of Philosophy*, *108* (2), 55–77. https://doi.org/10.5840/jphil201110823

Wilson, T., Lindsay, S., & Schooler, T. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. https://doi.org/10.1037/0033-295X.107.1.101