Forthcoming in Thompson, Robert (ed.) *The Routledge Handbook of Philosophy and Implicit Cognition.* [Please cite final version.]

Implicit Bias and Processing

This chapter will consider the kinds of processing involved in implicit bias, and how it relates to the debate on what kind of mental construct implicit biases are. I will begin by identifying what is meant by *implicit* bias and the *indirect* psychological measurement instruments designed to track it, before overviewing two streams of psychological research on implicit cognition. I will focus on dual process theories in this chapter, which recognize a distinction between the implicit and explicit, both with respect to processing and mental constructs. I will overview some versions of the canonical view in psychology of implicit biases as associations, before moving to recent empirical work, which has motivated an alternative, propositional understanding of implicit biases and the processes in which they partake. Next I motivate the need to recognize wide-ranging heterogeneity in the category of implicit bias, which accommodates various processes and mental constructs. Finally, I overview my preferred model of implicit biases as constituted by unconscious imaginings which is uniquely placed to accommodate this heterogeneity.

1. Two streams of attitude research

The term *attitude* is used differently in psychology and philosophy. In psychology it is used to pick out likings or dislikings, which are understood to be embodied in associations between objects and evaluations (Brownstein 2018: 264). In philosophy, the term is used to pick out a whole host of mental states, including beliefs, desires, and intentions (Brownstein and Saul 2016: 7). Talk of implicit *biases* then captures both kinds of ways of conceptualizing attitudes, since, as we will see, implicit biases have been understood both associatively and propositionally. If implicit biases are propositional, then there is a specific relation between their constituents in virtue of this, a relation which is absent if they are associations (Levy 2015: 804).

Michael Brownstein identifies two streams of research key to the development of the field of implicit social cognition, which he labels *True Attitudes,* and *Driven Underground* (2018: 266, see also Payne and Gawronski 2010). The True Attitudes stream came from cognitive theories of learning and selective attention, and recognizes two ways in which information is *processed*: automatic and controlled. However, this way of conceptualizing the landscape does not recognize two attitudes (one implicit, one explicit) that subjects might have towards a given target (e.g. a particular social group), but rather recognizes two kinds of information processing and measurement instruments – direct and indirect – where the latter is understood as helping us to get at a subject's *true attitude.*

One account of this kind is Russ Fazio's *Motivation and Opportunity as Determinants (MODE) model*. The driving thought behind this is that when subjects have motivation and opportunity to deliberate, they are able to respond to *direct* measures (that is, measures which allow for time and cognitive resources required for deliberation, and are based on self-report). However, when that control is taken away, *indirect* measures give us access to automatically processed information representative of one's true attitude, understood as an association in memory between an object and a subject's

evaluation of that object (Fazio 1990: 82). The strength of the association influences its accessibility and the likelihood that it will be automatically activated in response to certain stimuli (Fazio 1990: 92).

On the other hand, there is the *Driven Underground* stream according to which we have dissociated attitudes towards the same target object (e.g. members of a particular social group). Theories under this umbrella recognize the distinction between the implicit and explicit both in terms of processes and mental constructs. This stream of research grew out of work on implicit memory and gives *awareness* a key role in distinguishing the implicit from the explicit. In what follows I focus on this stream of research, since this is the space in which philosophers interested in the mental constructs and processing underlying implicit bias have operated.

2. *Implicit* bias and *indirect* measurement

To begin, we can understand implicit biases as 'the processes or states that have a distorting influence on behavior and judgement and are detected in experimental conditions with implicit measures' (Holroyd 2016: 154). Very roughly, implicit biases are posited as mental constructs which influence common micro-behaviours[1] and discriminations, which cannot be tracked, predicted, or explained by a subject's explicit attitudes (i.e. those attitudes a subject has introspective access to). For example, I might believe that men and women are equally adept at producing excellent philosophy, but when marking my students' work, if it is not anonymised, I tend to give female students less good marks than male students. My explicit attitudes will not explain that.

Implicit biases are also thought to be inaccessible to consciousness, automatically activated, and prevalent among even those who identify as egalitarian.[2] And, even though for any specific bias we care to name (e.g. concerning women and weakness) it might not be likely that any individual would harbor it, they are nevertheless likely to harbor *some* bias regarding that particular social group (i.e. women) (Holroyd and Sweetman 2016: 83, fn. 4).

In the literature on implicit bias, the term *implicit* is used to refer to at least four things: a psychological construct, a kind of measurement instrument, a set of processes (cognitive and affective), or a kind of evaluative behaviour (Brownstein 2019). For simplicity, and in honour of my particular interest in the topic, I will use the term *implicit* to characterize a particular kind of mental construct or process, where 'mental construct' is intended to be neutral between exactly what mental item implicit biases are (that is, neutral between

[1] Two notes of caution are called for here: some have argued on the basis of a meta-analysis of IAT results, that such results are a poor predictor of behaviour (Oswald and colleagues 2013, see Brownstein et al. 2019 for discussion). Others have argued that procedures designed to change results on implicit measures can do so (in a limited way, and in the short term), but that there is little evidence that such changes are reflected in behaviour (Forscher and colleagues 2019). Since this chapter is primarily about the processing underlying implicit biases, and less about the behavioral effects of them, I say no more about these points.

[2] In their review of over 2.5 million results from Implicit Association Tests (described below) across seventeen topics, Brian Nosek and colleagues report that '[i]mplicit and explicit comparative preferences and stereotypes were widespread across gender, ethnicity, age, political orientation, and region' (2007: 40).

implicit biases being associations, beliefs, imaginings, etc.). Another point of order is that the term *implicit bias* is used in at least two ways: to pick out a particular *mental construct* responsible for biased judgements and behaviours, or to pick out particular *judgements and behaviours* thought to be the result of certain implicit cognitions. This is a mere terminological issue; both ways of using the term recognize the existence of certain evaluative judgements and behaviours and mental constructs responsible for them. I will use the term *implicit bias* to pick out the mental construct involved in the generation of biased evaluative judgements, behaviours, and results on indirect psychological measures.

I now turn to overviewing a handful of these *indirect* (or *implicit*) measures,[3] which were designed to circumvent the shortfalls of *direct* (or *explicit*) measures when it came to assessing subjects' attitudes towards certain social categories (such as people not being honest about their attitudes, perhaps due to social desirability concerns). One way of measuring implicit biases is with an Implicit Association Test (IAT). IATs measure the speed at which subjects are able to pair two categories of objects (e.g. pictures of *old* and *young* faces) with, for example, pleasant and unpleasant stimuli (e.g. the words *wonderful* and *horrible*). The idea is that the speed and accuracy in the categorization performance of combinations of categories can give us insight into which categories a subject *associates* with one another (De Houwer et al. 2009: 347).

A second implicit measure is *semantic priming,* which is thought to assess a subject's strength of association between two concepts. One version of the semantic priming procedure presents two words in close succession (prime and target), where participants have to judge the second word (target), and their reaction time in doing so is recorded. One example of this class of measure comes from Mahzarin R. Banaji and Curtis D. Hardin, who were interested in the speed of judgements of gender-consistent or gender-inconsistent targets, and how that was influenced by a prime (Banaji and Hardin 1996: 136). In their experiment, subjects were exposed to an orientation symbol (+) (500ms), followed by a prime word (either male related, female related, gender neutral, or nonword) (200ms), then a blank screen (100ms), and the target pronoun appeared for as long as it took for a participant to enter a response (Banaji and Hardin 1996: 137). They found that when the target gender matched the prime gender, judgement was faster than when the target gender did not match the prime gender. Overall, Banaji and Hardin took their procedure to provide 'evidence for automatic gender stereotyping', which occurred 'regardless of subjects' awareness of the prime-target relation, and independently of explicit beliefs about gender stereotypes' (Banaji and Hardin 1996: 139).

Another implicit measure thought to measure the strength of associations is the Go/No-Go association task. Association strength is assessed by the participant's ability to discriminate items belonging to a target category and attribute from distractor items which are not members of the target category or attribute (Nosek and Banaji 2001: 627). For example, in one

[3] See Brownstein and colleagues (2019) and Gawronski and De Houwer (2014) for excellent critical overviews of implicit measures.

condition participants might be asked to simultaneously identify stimuli representing the target category *fruit* and the attribute *good*, and in a second condition to simultaneously identify stimuli representing the target category *fruit* and the attribute *bad*. The idea is that the ease or otherwise of identifying stimuli of the target category and attribute gives us insight into implicit attitudes regarding the target category (Nosek and Banaji 2001: 627).

Finally, the Affect Misattribution Procedure has participants make evaluative judgements in an ambiguous context. In one version of the task, participants are exposed to a positively or negatively valenced prime (e.g. President George W. Bush), and are then told to evaluate an ambiguously valenced target (e.g. an abstract symbol), and that they should avoid expressing any influence of the valenced prime in their evaluation of the ambiguously valenced target (Payne et al. 2005: 277). The idea is that this set up leads to subjects 'projecting their own psychological state onto an ambiguous source' when they misattribute their reactions which are caused by the prime (Payne et al. 2005: 277). The misattribution observed is taken to be a reflection of implicit attitudes.

3. Associationism

For those who want to recognize distinct mental constructs with talk of implicit and explicit attitudes, a common way of distinguishing these constructs, broadly and not just with respect to theorizing about implicit *bias*, is along associative-propositional lines. That is, those theories which endorse the implicit–explicit cognition distinction posit distinct implicit and explicit processes characteristic of implicit and explicit mental constructs: one automatic, one controlled. The canonical view of implicit biases is that they are associations whose existence can be traced to the learning history of the subject (Levy 2015: 803). Concepts are associated with valences (e.g. *women* with negative valence) or with other concepts (e.g. *women* with *weakness*) in such a way that the activation of one makes more accessible the activation of another. In the presence of appropriate stimuli, stored associations between say, *women* and *weakness* become activated.

In psychology, dual process models of bias take associations to be paradigmatic of implicit cognition, whilst explicit cognitions (e.g. beliefs) involve propositional processing. For example, according to the *Reflective-Impulsive Model (RIM)*, two systems (*reflective* and *impulsive*) guide human behaviour, and those systems operate according to distinct principles of information processing. The impulsive system, via the activation within associative networks, elicits spontaneous tendencies of approach and avoidance, whilst the reflective system, via reasoned decisions, influences behaviour (Strack and Deutsch 2004). Similarly, the *Associative-Propositional Evaluation model (APE)*, takes it that the outcomes of associative processes are implicit evaluations, whilst the outcomes of propositional processes are explicit evaluations (Gawronski and Bodenhausen 2014: 188).

As this chapter has already made clear, thinking of implicit biases in associative terms permeates the literature. Not only has the distinction between associative and propositional processing been thought to map onto implicit and explicit cognition, the indirect measures overviewed earlier are all thought

to be tracking implicit *associations* (although as Levy (2015: 804) points out, such measures might equally be tracking propositional constructs). While the orthodoxy in psychology has been to understand implicit biases associatively, in philosophy there has been a recent move away from this picture, to instead understanding implicit biases as having propositional contents and as partaking in propositional processing.

4. Propositionalism

States with propositional contents can have satisfaction conditions, whereas states with only associative contents cannot, instead associative contents are '(relations among) mental representations that lack any syntactic structure' (Mandelbaum 2013: 199, fn. 1). To see the difference between associative and propositional processing, consider the following examples. The first is what you might be primed to think (implicitly or explicitly) when you hear *salt* (*pepper*!). This is a merely associative process, the idea being that your concepts of *salt* and *pepper* are associatively linked, such that the activation of one makes more accessible the activation of the other, but there's no *relationship* thought to hold between these concepts. Now, second, imagine you get engrossed in a long running TV Series about two academics, Dr Salt and Dr Pepper, engaged in a Shakespearean love affair across University faculties traditionally at odds with one another. Now the relationship in your mind between *salt* and *pepper* is not a matter of one making more accessible the other (although perhaps that too), but rather that there's a specific relation which one bears to the other, e.g. Salt *loves* Pepper.

Return to implicit biases. On an associationist understanding of one kind of gender bias, it might be said that my concept of *women* is associatively linked to another concept, *weakness.* But this conception of the mental apparatus does not specify a particular *relationship* which holds between *women* and *weakness*, it is rather just a matter of one concept making more accessible the activation of the other. On a propositionalist understanding of this particular bias however, the story might be that I have a propositionally structured mental construct with the content *women are weak.* In this case we have a single mental item (rather than an association between two), and the specification of a particular relationship between the constituents, that of predication.

Some empirical work on the behaviour of implicit biases has motivated a move to propositionalism, the case for which has been most robustly put by Eric Mandelbaum (2016). One study he draws on by Bertram Gawronski and colleagues (2005) had participants introduced to a photo of an unfamiliar person (CS1), which was then paired with either positive or negative statements (in order to set up the association between the CS1 and a particular evaluation). Then participants were introduced to a second photo of an unfamiliar person (CS2) and told either that the CS1 liked the CS2 or that the CS1 did not like the CS2. Subjects then underwent an explicit likeability rating to gauge explicit attitudes towards the 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 previous images, to gauge implicit attitudes towards the CS1 and CS2 (Gawronski et al. 2005: 621). The results showed that, for example, if the CS1 was paired with negative statements, and participants

were told that the CS1 did not like the CS2, then participants liked the CS2. Reflecting on this Mandelbaum suggests that an associative theory would predict the opposite of this result: 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) you are activating another negative valence because you are told that the CS1 *dislikes* the CS2' (2016: 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). 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, Petty, and McCaslin 2009; for evidence that implicit biases are adjustable in light of peer judgement, see Sechrist and Stangor 2001).

Experiments such as these have led some philosophers to develop propositional models of bias. Before overviewing Mandelbaum's view, it is worth noting that the argument from the results of these studies to the propositional structure of bias is not without its critics. As Josefa Toribio points out, the move made by Mandelbaum is from biased behaviour (i.e. that measured by implicit measures) being modulated by logical and evidential considerations, to the claim that that which is responsible for the behaviour is propositional (Toribio 2018: 41; see also Brownstein et al 2019: 5). Mandelbaum moves between these claims via a supposed inference to the best explanation. However, Toribio argues that an associationist view can accommodate the results from the empirical studies Mandelbaum uses in his case for propositionalism (Toribio 2018: 44). Moreover, she argues that '[e]ven if implicit attitudes can sometimes be modulated by evidential and rational considerations, this is not their most distinctive characteristic, and any inquiry into their nature that emphasizes this aspect will be off-target as far as best explanations go' (Toribio 2018: 44).

Toribio's argument is as follows: at the general level, sensitivity to logical or evidential considerations is not necessarily best explained by the state in question being propositionally structured. She uses the example of *pain* and *disgust,* both of which can be modulated logically or evidentially, and yet we do not take that fact to be grounds on which to pronounce that they are thus propositionally structured (Toribio 2018: 44–7).

In addition, Toribio argues that the move from evidential/logical moderation to the propositional structure of implicit bias would work only if implicit biases were the sole cause of implicitly biased behaviour. However,

> no one in social psychology would deny that all sort of factors—associative, non-associative and even non-attitudinal processes—have to be taken into account when offering explanations of implicit attitudes' modulation […] It is perfectly consistent to maintain that implicit attitudes are associations and that factors other than counter-conditioning and extinction can modulate implicitly biased behaviour. (Toribio 2018: 50)

Notwithstanding criticism of Mandelbaum's argument for the propositional nature of implicit bias, I now move on to look at his model, on which implicit biases are *unconscious beliefs*.[4] Philosophical views which characterize the target phenomenon as a *belief* fall under the heading of *doxasticism,* and so Mandelbaum's view is a version of this approach. On its face, that 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* (Gilbert et al. 1993: 221). 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. If implicit biases are beliefs, these are features we might have to deny when we reflect on cases of implicitly biased egalitarians (that is, those folk for whom their implicit attitudes do not cohere with their egalitarian explicit attitudes).

However, Mandelbaum is not working with a traditional Cartesian conception of belief, but rather understands belief in a Spinozan fashion. According to this understanding, we *believe* any truth-apt proposition that we represent. So there is no gap between representing a truth-apt proposition, and believing it, that is, 'the act of understanding *is* the act of believing' (Gilbert et al. 1993: 222, my emphasis). 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* (a la Lewis 1982; Stalnaker 1984; or more recently, Egan 2008).

Mandelbaum's claim that implicit biases are *beliefs* has been subject to a host of objections.[5] However, Mandelbaum claims not to be motivated primarily by capturing implicit biases as a particular kind of mental construct, but rather as capturing them as propositionally structured. He notes that if the term *belief* offends readers should feel free to understand his hypothesis as one about *structured thoughts* (Mandelbaum 2016: 636). To this I note that if it is in the spirit of Mandelbaum's position to understand structured thoughts non-doxastically, then what we have arrived at is only the propositional nature of implicit bias (at best). The question of what kind of mental construct implicit biases are remains open.

5. Heterogeneity

Many authors have pointed out the need to recognize serious heterogeneity in our model of bias. That is, although it is common to talk of implicit bias simpliciter (without making any finer grained distinctions within the category), it is almost certainly the case that the category admits of significant

[4] It is beyond the scope of this piece to mention all of the views on the nature of implicit bias put forward by philosophers. Other candidate mental constructs have included: *alief* (Gendler 2008), *character traits* (Machery 2016), *implicit beliefs* (Frankish 2016), *in-between beliefs* (Schwitzgebel 2010), *patchy endorsements* (Levy 2015), and *mental imagery* (Nanay 2021). More radical views include Chehayeb's (2020) *mosaic view,* and Johnson's (2020) non-representational account.

[5] For example see Holroyd (2016), Levy (2015), Madva (2016). See sect. 4.1 of my (2019) for an overview of these objections.

heterogeneity. Subsuming all implicit biases under a single mental construct partaking in one kind of process risks making unwarranted generalizations about how implicit biases behave, in particular, how they might influence judgements of, and behaviour towards, members of certain social groups (Holroyd and Sweetman 2016: 84).

Firstly, there are differences with respect to implicit biases having particular features. For example, although it is often said that implicit biases are *automatic* (as opposed to *controlled*), some have argued that if not the *activation* of implicit biases (e.g. the activation of a stored association), their *expression* (i.e. the influence on behaviour) admits of control (see Chehayeb 2020: 123–6). Implicit biases are also often characterized as not being introspectively accessible, or as being unconscious (indeed, it is this feature which is central to dual process theories of implicit cognition).[6] Some studies show that under certain conditions, there is more convergence between a subject's reported explicit attitudes and their implicit attitudes as identified by indirect measures (see e.g. Nier 2005, Hahn et al 2014). This has persuaded some theorists that implicit biases may well admit of some introspective accessibility (see e.g. Brownstein et al. 2019; Chehayeb 2020: 129–133, Gawronski 2019: 575–8; Levy 2014: 30).[7]

Secondly, implicit biases differ in their contents, insofar as they are *about* different social groups. For example, we can have implicit biases about women, Black and ethnic minority people, homosexual people, members of certain religious communities, and so on.[8]

Thirdly, implicit biases *concerning the same social groups* can vary with respect to their expression. That is, even though a given individual may score highly on an IAT testing for associations between Black men and stereotypical traits, they may nevertheless not score highly on an IAT testing for negatively valenced associations involving their concept of Black men, and vice versa (see Amodio and Devine 2006). These distinct biases are also predictive of different behaviours (the former influencing judgements of competence, the latter influencing seating distance from a Black confederate). One way of understanding what the two kinds of IAT are tracking here is as two kinds of association which fall under the label of *implicit bias*: *semantic* and *affective* (see Holroyd and Sweetman 2016: 92ff for arguments against the explanatory utility of this distinction).

---

[6] Gawronski and colleagues (2006) distinguish three types of awareness: *source, content,* and *impact,* and they argue that implicit attitudes only differ from explicit attitudes with respect to *impact awareness* (cf. Gawronski 2019: 577).

[7] I think that these studies might in fact be consistent with implicit biases being mental constructs to which we do not have introspective awareness, and indeed, the psychologists running the studies suggest that the results can be interpreted in a way that retains the introspective inaccessibility of implicit biases. In short the idea is that subjects make pretty good predictions about the biases they have by inferring from their affective reactions to certain stimuli to the existence of a bias (as suggested by Hahn and colleagues (2014: 1387)). But that is not to say that they have access to the bias itself, but rather an affective reaction downstream of it. See Berger (2020) for an alternative way of understanding the distinction between implicit and explicit cognitions in terms of awareness.

[8] Of course, we also have implicit cognitions which are not about social categories, for example, associations concerning feared objects (Holroyd and Sweetman 2016: 86, fn. 7).

So how do we make good on this heterogeneity? One approach comes from Bryce Huebner, who argues for a variety of ways that implicit associations get internalized. Specifically, he has it that implicit biases reflect the combined influence of three computationally and psychologically distinctive evaluative systems (associative processing, associative Pavlovian systems, and associative model-free systems) (2016: 51).

Guillermo Del Pinal and Shannon Spaulding also speak to the heterogeneity of biases, though at the level of their encoding. Usually salient-statistical associations are thought to be the relevant ones for modeling implicit biases. But Del Pinal and Spaulding argue that some biases are encoded 'in the *dependency networks* which are part of our representations of social categories', and not all as salient-statistical associations (2018: 96). This would mean that some biases can be encoded in our concepts in ways that systematically dissociate from salient-statistical properties. Rather, concepts can encode information regarding cue-validity and saliency, but also the degree of centrality of their associated features.

We have seen then that the mental constructs falling under the label of *implicit bias* admit of heterogeneity with respect to key features (control, introspective accessibility), content, and influences on behaviour. We have also seen that those theorists keen to accommodate heterogeneity do so within the confines of associationism, that is, heterogeneity is grounded in different kinds of association (semantic/affective, salient-statistical/dependency networks) or ways in which these associations become a part of our cognitive architecture (Chehayeb 2020 is a notable exception). In the next section I argue that we should take heterogeneity more seriously than this, before briefly defending my preferred view of bias which is able to do so.

6. Biased by our imaginings

Although many theorists have wanted to recognize heterogeneity, we have seen that accounts of implicit biases fall squarely into either associationism or propositionalism, and any heterogeneity posited remains within the boundaries of these respective frameworks. On my preferred view[9] implicit biases are *constituted by unconscious imaginings,* and the heterogeneity within the category spans the associative-propositional distinction.[10] Before getting to the details, I will say something about how I am understanding *unconscious imagination*.

In place of a robust account of imagination, I appeal to three features of it upon which there is 'wide agreement' (Kind 2016: 1): it is a primitive mental state (that is, it is irreducible to other mental states, cf. Langland-Hassan 2012), it has representational content (it is *about* something), and it is not connected to truth[11] (Kind 2016: 1–3). It is this kind of state which I think can do good

[9] A full statement and defence of my view of implicit bias can be found in my (2019), a defence of my view in light of studies on mitigating bias through virtual reality can be found in my (*manuscript*).

[10] For a radically different account of the mental constructs underlying social behaviour, which accommodates more far-reaching heterogeneity than my view, see Chehayeb (2020, Ch. 6).

[11] Kind understands this third feature as specifying the absence of a *constitutive* connection to truth. Elsewhere (2017; 2018; 2020) I defend a contingent relationship between belief and truth,

explanatory work when turning to the nature of implicit bias, and is also uniquely placed to take part in both associative and propositional processes. *Unconscious imaginings* then are simply states with the three features upon which there is wide agreement, and which are tokened in a way as to be inaccessible to introspection.[12]

In pursuit of recognizing a wide-ranging heterogeneity, I also distinguish two kinds of imagining on grounds of content: propositional imaginings (imagining *that your partner's birthday is next month*) and imagistic imaginings (imagining *your partner's face*). On my view, both kinds of imaginings can be tokened unconsciously, and both kinds of imaginings have a role to play in our account of implicit bias.

If implicit biases are constituted by unconscious imaginings, and unconscious imaginings are candidate mental constructs for partaking in both associative and propositional processes, my view thus recognizes heterogeneity at the level of mental constructs and processing. That is, it can accommodate implicit biases being associatively structured (i.e associations between multiple imaginings) and it can accommodate them being propositional (i.e. single imaginings with propositional contents apt to e.g. partake in inference). This is a theoretical virtue given the mounting evidence for heterogeneity in this class of attitudes. My account can do two key things. It can say what is common among all implicit biases which justifies grouping all of these things together under a single label, and it can also admit of finer distinctions within the overall category, by appeal to the different ways in which implicit biases can be constituted by unconscious imaginings (e.g. by associatively linked imaginings, or a relationship of identity in cases of single propositional imaginings).

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. 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. of *woman* and *weakness*) (as Toribio points out, understanding implicit biases as associations is consistent with thinking of the associated mental constructs as images (2018: 42)). Alternatively implicit bias could be understood as associatively linked propositional imaginings (i.e. *there is a woman* and *there is weakness*), or as an

---

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] It might be thought that understanding imaginings as tokened unconsciously departs from a standard view of imagination, and thus that my preferred view has ontic costs to pay. However, the three features upon which there is wide agreement are neutral with respect to whether imaginings can be tokened unconsciously. So, if unconscious imagination does represent a departure from a standard view, that departure is not to be found in these three uncontroversial features (for a more through defence of the claim that allowing for imaginings to be tokened unconsciously is not to endorse a revisionary notion of the imagination see Sullivan-Bissett (2019: 631–635)).

unconscious imagistic or propositional imagining associatively linked with a negative valence.[13]

As we have seen though, there has been a recent move to modeling implicit bias as propositionally structured, and empirical work suggesting that this is required, in at least some cases. If that is right, we should understand the constituents of implicit bias as standing in determinate relations to one another. There are two ways my imagination model can capture what form implicit biases could take against a propositional 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 unconscious belief model and my unconscious imagination model look very similar and may share predictions.

Before closing, I will say something about the theoretical and predictive benefits that might be gained if we understand implicit biases in the way I suggest. The different features, structures, and behaviours characteristic of implicit biases are theoretically interesting in their own right, but are also likely to be significant when we think of the kinds of normative recommendations regarding mitigation strategies that might be suggested on the basis of the nature of implicit biases. My general point is that my account can admit of more particular carvings of the category of implicit bias along the lines of which kind of imaginings and processes are in play. Different kinds of imaginings may be predictive of different behaviour, and the more we learn about the operations of different kinds of unconscious imaginings, the more predictions we will be able to make about implicit biases understood in such terms.

In their argument for associative heterogeneity, Holroyd and Sweetman suggest that the way different associations operate may be explained by differences in content and underpinning processes (Holroyd and Sweetman 2016: 88). The way imagination can be involved in associative processing (where the components are imagistic or propositional) may be one way in which different underpinning processes are involved in associative processing. But, in addition, we should also take seriously the possibility that dissociative scores on indirect measurements may not only be down to different kinds of associations and the processes which underpin them, but may also be the result of measures tapping into different kinds of implicit processes (associative and propositional).

My model also allows for variation in the mechanisms responsible for biased behaviours. Where implicit biases are underpinned by associative processing (with unconscious imaginings as the constituents of the association), the explanation for certain judgements or behaviours can be given by appeal to one imagining activating another. But where implicit biases are

[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.

non-associative, and instead involve single propositionally structured mental constructs (i.e. single unconscious imaginings), the imagining itself has motivational credentials. For example, the idea would be that when presented with particular stimuli (e.g. a woman's face), a subject with an unconscious propositional imagining takes there to be a determinate relation between *woman* and *weakness* (i.e. the constituents of her attitude). So some implicit biases, in virtue of being propositional, posit determinate relations between the target stimuli and some stereotypical feature or valence.

It might be thought that the cost of accommodating significant heterogeneity is giving up on understanding implicit biases as a single kind of mental construct. My account has the twin benefits of accommodating both unification and heterogeneity. It is unifying insofar as it can principally group all implicit biases under a single mental category (as constituted by unconscious imaginings). But it also allows for diverse ways in which implicit biases can be constituted, as well as various processing in which implicit biases might partake.

7. Conclusion

In this chapter I have overviewed some ways of thinking about the processing involved in explicit and implicit cognition, with a focus on implicit bias. With respect to this category, evidence is mounting for the idea that we are not in the domain of a neat and tidy subset of implicit cognition, but rather that an extensionally adequate account of implicit bias and the processing involved needs to recognize significant heterogeneity. Extant accounts are unable to do this insofar as they situate themselves in *either* an associative *or* propositional framework. I suggested that my preferred model of implicit bias as constituted by unconscious imaginings is uniquely able to meet the challenge of heterogeneity, whilst also offering a unifying model of the mental constructs and processes in implicit bias.

Acknowledgements

I am grateful to Fidaa Chehayeb, Michael Rush, Robert Thompson, and an anonymous reviewer for comments on an earlier draft of this chapter.

References

Amodio, D. M. and 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*: 652–661.

Banaji, Mahzarin, R. and Hardin, Curtis, D. 1996: 'Automatic stereotyping'. *American Psychological Society, 7:* 136–141.

Berger, J. 2020: 'Implicit attitudes and awareness'. *Synthese, 197:* 1291–1312.

Brinol, P., Petty, R., and McCaslin, M. 2009: 'Changing attitudes on implicit versus explicit measures: what is the difference?' In R. Petty, R. Fazio, and P.

Brinol, eds., *Attitudes: insights from the new explicit measures.* New York: Psychology Press: 285–326.

Brownstein, M. 2018: 'Implicit bias and race'. In P.C. Taylor, L.M. Alcoff, and L. Anderson, eds., *The Routledge companion to the philosophy of race*. New York: Routledge: 261–276.

Brownstein, M. 2019: 'Implicit bias'. In E. Zalta, ed., *The Stanford encylopedia of philosophy,* Fall 2019 Edition, <https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/>

Brownstein, M., Madva, A., and Gawronski, B. 2019: 'What do implicit measures measure?' *Wires Cognitive Science,* 10: e1501.

Brownstein, M. and Saul, J. 2016. 'Introduction'. In M. Brownstein and J. Saul, eds., *Implicit bias and philosophy, vol. 1: Metaphysics and epistemology.* Oxford: Oxford University Press: 1–19.

Chehayeb, F. 2020. *Contra implicit bias.* PhD Thesis. University of Birmingham.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., and Moors, A. 2009. 'Implicit measures: a normative analysis and review'. *Psychological Bulletin, 135*: 347–368.

Del Pinal, G. and Spaulding, S. 2018. 'Conceptual centrality and implicit bias'. *Mind & Language*, *33*: 95–111.

Egan, A. 2008. 'Seeing and believing: perception, belief formation, and the divided mind'. *Philosophical Studies, 140*: 47–63.

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

Forscher, P. S., Lai, C.K., Axt, J. R., Ebersole, C. R., Herman, M.and Devine, P. G. 2019. 'A meta-analysis of procedures to change implicit measures'. *Journal of Personality and Social Psychology, 117*: 522–559.

Frankish, K. 2016. 'Playing double: implicit bias, dual levels, and self-control'. In M. Brownstein and J. Saul, eds., *Implicit bias and philosophy, vol. 1: Metaphysics and epistemology.* Oxford: Oxford University Press: 23–46.

Gawronski, B. 2019. 'Six lessons for a cogent science of implicit bias and its criticism'. *Perspectives on Psychological Science, 14*: 574–595.

Gawronski, B. and De Houwer, J. 2014. 'Implicit measures in social and personality psychology'. In H. T. Reis and C. M. Judd, eds., *Handbook of research methods in social and personality psychology.* New York: Cambridge University Press: 283–310.

Gawronski, B., Walther, E., and 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*: 618–626.

Gawronski, B. and Bodenhausen, G.V. 2006. 'Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change'. *Psychological Bulletin, 132*: 692–731.

Gawronski, B. and Bodenhausen, G.V. 2014. 'The associative-propositional evaluation model: operating principles and operating conditions of evaluation'. In J. W. Sherman, B. Gawronski, and Y. Trope, eds., *Dual-process theories of the social mind.* New York: Guilford Press: 188–203.

Gendler, T. 2008. 'Alief and belief'. *Journal of Philosophy, 105*: 634–663.

Gilbert, D.T., Tafarodi, R.W., and Malone, P. S. 1993. 'You can't not believe everything you read'. *Attitudes and Social Cognition, 65*: 221–233.

Hahn, A., Judd, C. M., Hirsh, H. K., and Blair, I. V. 2014. 'Awareness of implicit attitudes'. *Journal of Experimental Psychology: General, 143*: 1369–1392.

Holroyd, J. 2016. 'What do we want from a model of implicit cognition?' *Proceedings of the Aristotelian Society*, *166*: 153–179.

Holroyd, J. and Sweetman, J. 2016. 'The heterogeneity of implicit bias'. In M. Brownstein and J. Saul, eds., *Implicit bias and Philosophy, vol. 1" Metaphysics and epistemology.* Oxford: Oxford University Press: 80–103.

Huebner, B. 2016. 'Implicit bias, reinforcement learning, and scaffolded moral cognition'. In M. Brownstein and J. Saul, eds., *Implicit bias and Philosophy, vol.1: Metaphysics and epistemology.* Oxford: Oxford University Press: 47–79.

Johnson, G.M. 2020. 'The Structure of Bias'. *Mind, 129*: 1193-1236.

Kind, A. 2016. 'Introduction: exploring imagination'. In A. Kind, ed., *The Routledge handbook of philosophy of imagination.* London: Routledge: 1–11.

Langland-Hassan, P. 2012. 'Pretense, imagination, and belief: the single attitude theory'. *Philosophical Studies, 159*: 155–179.

Levy, N. 2014. 'Consciousness, implicit attitudes and moral responsibility'. *Noûs, 48*: 21–40.

Levy, N. 2015. 'Neither fish nor fowl: implicit attitudes as patchy endorsements'. *Noûs*, *49*: 800–823.

Lewis, D. 1982. 'Logic for equivocators'. *Noûs, 16*: 431–41.

Machery, E. 2016. 'De-Freuding implicit attitudes'. In M. Brownstein and J. Saul (Eds.), *Implicit bias and Philosophy, vol. 1: Metaphysics and epistemology.* Oxford: Oxford University Press: 104–129.

Mandelbaum, E. 2013. 'Against alief'. *Philosophical Studies*, *165*: 197–211.

Mandelbaum, E. 2016. 'Attitude, inference, association: on the propositional structure of implicit bias'. *Noûs*, *50*: 629–658.

Nanay, B. 2021. 'Implicit bias as mental imagery'. *Journal of the American Philosophical Association, 7:* 329-347.

Nier, J. 2005. 'How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach.' *Group Process & Intergroup Relations, 8*: 39–52.

Nosek, B. and Banaji, M. 2001. 'The Go/No-go association task'. *Social Cognition, 19*: 625–664.

Nosek, B., Smyth, F.L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Tucker Smith, C., Olson, K. R., Chugh, D., Greenwald, A. G., and Banaji, M., 2007. 'Pervasiveness and correlates of implicit attitudes and stereotypes'. *European Review of Social Psychology, 18*: 36–88.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. 2013. "Predicting ethnic and racial discrimination: a metaanalysis of IAT criterion studies". *Journal of Personality and Social Psychology, 105*: 171–192.

Payne, K., Cheng, C.M., Govorum, O., Stewart, B.D. 2005. 'An inkblot for attitudes: affect misattribution as implicit measurement'. *Journal of Personality and Social Psychology, 89*: 277–293.

Payne, K.B. and Gawronski, B. 2010. 'A history of implicit social cognition: where is it coming from? Where is it now? Where is it going?' In B. Gawronski and K.B. Payne, eds., *Handbook of implicit social cognition: measurement, theory, and applications*. New York: Guilford Press: 1–15.

Schwitzgebel, E. 2010. 'Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief'. *Pacific Philosophical Quarterly, 91*: 531–553.

Sechrist, G.B. and Stangor, C. 2001. 'Perceived consensus influences intergroup behavior and stereotype accessibility'. *Journal of Personality and Social Psychology, 81*: 645–654.

Stalnaker, R. 1984. *Inquiry.* Cambridge, MA: MIT Press.

Strack, F. and Deutsch, R. 2004. 'Reflective and impulsive determinants of social behavior'. *Personality and Social Psychology Review, 8*: 220–247.

Sullivan-Bissett, E. 2017. 'Biological function and epistemic normativity'. *Philosophical Explorations*, *20*: 94–110.

Sullivan-Bissett, E. 2018. 'Explaining doxastic transparency: aim, norm, or function?' *Synthese*, *195*: 3453–3476.

Sullivan-Bissett, E. 2019. 'Biased by our imaginings'. *Mind & Language, 34*: 627–647.

Sullivan-Bissett, E. 2020. 'We Are Like American Robins'. In S. Stapleford and K. McCain, eds., *Epistemic Duties: New Arguments, New Angles.* Routledge: 94–110.

Sullivan-Bissett, E. *manuscript*: 'Virtually imagining our biases'.

Toribio, J. 2018. 'Implicit bias: from social structure to representational format'. *Theoria*, *33*: 41–60.