CONTRA IMPLICIT BIAS

by

FIDAA CHEHAYEB

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy

School of Philosophy, Theology and Religion

College of Arts and Law

University of Birmingham

April 2020

Abstract

Orthodox literature on bias follows what I call the *Dualistic Alignment Hypothesis* (or the dualistic paradigm). The basis of this dualistic paradigm culminates in the assumption that there is a principled distinction between *explicit attitudes* and *implicit bias* and the behaviour guided by each. In this thesis I do two things. First, I challenge the dualistic paradigm on empirical and conceptual grounds. Further, I show that dualistic thinking faces serious challenges. Ultimately, I reject the dualistic paradigm as the best explanatory theory of bias. Second, I propose a novel explanatory hypothesis of the data, the *mosaic view*. I argue that the mosaic view is grounded in a more realistic understanding of social evaluations. What underpins social behaviour are not two unified, stable, and distinct mental kinds, but rather a complex conglomeration of interacting elements of mind within a network I call a *stance*. The principal idea underlying the mosaic view is that social behaviour is best considered as the result of complex interactions between elements of a stance, activated differently across contexts, and interacting with background beliefs, commitments, values, and other mental states.

To my late father.

Wherever you are,

I hope you are proud of me.

Acknowledgments

"At times, our own light goes out and is rekindled by a spark from another person. Each of us has cause to think with deep gratitude of those who have lighted the flame within us." – Albert Schweitzer

I am grateful for my supervisory team. I could not have wished for better supervisors. I extend my gratitude to my first supervisor, Scott Sturgeon for believing in me, for rekindling my spark whenever it went out. Without his invaluable intellectual and emotional support, my PhD experience would have been a much less pleasurable one. I thank my second supervisor Ema Sullivan-Bissett for her detailed comments on the numerous drafts. Our long discussions played a key role in improving this thesis. I hope finally to have convinced her that my view is superior. Thanks to my third supervisor, Jussi Suikkanen, for pressing me to clarify my positions on several issues, for his timely comments whenever I sent a draft, and for the conversations which greatly informed my work on this thesis. I am also deeply grateful to Maja Spener for being the first to believe in me, for introducing me to the PhD program, and for always making me feel like family at her home.

I would like to thank every family member and friend who has supported me through this journey. Without them, I could not have done a fraction of this work. My husband, Faisal, first and foremost, thank you for believing in me when I gave up on myself. Thank you for being proud of me and for loving me even in my worst days. Raya, thank you for being the light, color, and music in my life. Lynn, thank you for being my courage, my strength, and my confidence. Jad, thank you for being my sharp critical mind and my warm loving heart, and thank you for reading all seven chapters of the thesis and giving me feedback. My sister Amel, thank you for everything you do for me, not least proofreading the thesis. I love you all beyond measure. Finally, I am so grateful to Hana and Nia, without whom I could not have finished this work.

Contents

List of Figures

Chapter 1

INTRODUCTION

On April 17th, 2018, CNN reported on two men who walked into a Starbucks café in Philadelphia asking to use the bathroom but told it was for customers only. They sat peacefully at a table without purchasing anything, yet the manager came over and asked them to leave. When they declined her request, claiming to be waiting for a friend, she called the police. The two men were escorted out of the coffee shop by the police and then arrested for trespassing to be released later as Starbucks didn't press charges. These men did nothing out of the ordinary, they simply sat down at a table in a coffee shop, a common occurrence in such establishments, yet they were arrested. Protests and sit-ins followed this incident and the black community insisted that such harsh and unjust treatment of the two men was due to their skin colour (McLaughlin, 2018). Starbucks officials called for changes to their policies including training around what is commonly referred to as *implicit bias* and described as *unconscious and uncontrolled prejudice*.

Over the last few decades, the notion of *implicit bias* has emerged as a significant explanation for a wide array of inequalities and discriminations. The idea behind the concept of *implicit bias* is that many people (including those who generally consider themselves to be egalitarian and well-intentioned) are unaware of the biases which nonetheless influence their behaviour. In the dominant philosophical literature, *implicit bias* is typically equated with *unconscious* and *uncontrolled* bias. Implicit bias has not only been of great interest to psychologists, philosophers, legal officials, and policy makers, but it has also pervaded popular culture and the media where it has been touted as a significant contributor to social injustices.[1]

In our society, as we consider perpetrators of social injustice, including those who are simply complicit to injustice, we tend to judge them as unethical human beings. In all probability, if we are like most people, as we make these judgements of others and of ourselves, we are quick to acknowledge having been victims of some form of social injustice at one time or another but are reluctant to admit to having been perpetrators of injustice. As Benjamin Sherman (2015, p. 10) puts it, we "are likely to think that the vast majority of the time, [our] judgements are fair and accurate, otherwise, they wouldn't persist in being [our] judgements".

Equally, many of us may be reluctant to acknowledge being perpetrators of injustice manifested in *implicit bias,* although we are generally slightly more accepting of such a charge. Likely because what we (think we) know about implicit bias is that it is unconscious and uncontrolled, and that it is something that all of us suffer from, through no fault of our own. If we are unaware of bias and if we can't control it, then maybe we are not entirely responsible for it; in effect, questions about culpability become subject to discussion.

One reason for our reluctance may be that the image we typically have of perpetrators of injustice is often stark. For example, we generally tend to consider racists as individuals who engage in violent acts against blacks, or at least as individuals who refuse to interact with persons of colour either professionally or socially. Our notion of a homophobe may conjure images of

[1] For example, the Ohio State University has its own research institute 'The Kirwan Institute' which is dedicated to understanding (among other causes of social injustice) implicit bias and how it creates and sustains social barriers (Kirwan Institute, 2015).

conservative, religious individuals or machismo men who take pride in their aggressiveness. And the thought of a sexist man may bring about ideas of catcalling construction workers. At least for many people, such perpetrators of injustice are typically considered immoral human beings, while on the other side, non-prejudiced individuals are viewed as ‘good’ moral beings. In our daily lives, we explain some unjust social behaviours as the result of prejudices and biases, while fair and egalitarian actions, we attribute to non-biased beliefs. This is the picture that many of us, to a certain extent, fall prey to.

Yet this almost Manichaean black and white picture excludes many of the behaviours that lie in the different shades of grey.[2] Such (grey) behaviours, though appear innocuous to those executing them, they nonetheless contribute to reinforcing social injustice. By these behaviours I mean to include commonly experienced cases as the one discussed above. I also mean various empirical findings as those which show persons of colour to be given less quality medical care than their white counterparts (Green et al., 2007) and data showing the racial profiling of persons with Arabic names by immigration officers at airports (Baker, 2002). Closer to home, I also mean to include findings showing women’s underrepresentation in philosophy and their impediment from progress in this field (Saul, 2013a). Empirical research has shown that such discriminatory acts are often propagated even by persons who profess otherwise egalitarian beliefs and think of themselves as morally upstanding subjects. Such individuals are aptly referred to by leading scholar on the subject, Jules Holroyd (2016, p. 160), as *biased egalitarians*.

[2] Manichaeism is a view which considers things in the world (in a dichotomous manner) as either only good or bad, light or dark, black or white.

It may appear unclear where to situate biased egalitarians on a black and white picture of morality and equally unclear how to explain the causes of such seemingly contradictory behaviour. Are such behaviours caused in the same way and by the same psychological mechanisms in persons who perpetrate them as in stark racists, sexists, homophobic, or xenophobic people? Indeed, the lay person may be at a loss when she witnesses people she generally knows to be good, moral beings exhibiting unjust behaviours, likely because she is unable to fit this behaviour into the dichotomous black and white picture of morality she unknowingly subscribes to. It is certainly perplexing to witness otherwise good, egalitarian persons exhibit prejudiced behaviour towards certain social identity groups or members of those groups. Cases as these are also equally philosophically puzzling. Philosophers working on prejudice and bias are very interested in such findings, especially in cases where agents seem to be unaware of, unconscious of, or unable to introspect into the bias they exhibit, as when they manifest what is commonly known as *implicit bias* in their behaviour.

Scholars interested in bias are presented with a puzzling phenomenon: many individuals exhibit evaluations of members of marginalized social identity groups which, in some cases, they seem to be unaware of, and whose influence on action they seem to be unable to control. The phenomenon is considered puzzling partly because it appears to be unexplainable by avowed beliefs. Substantial empirical evidence has emerged for the claim that many people act in ways which conflict with their explicitly endorsed beliefs. Such biased egalitarians show an inconsistency between their self-reported evaluation of a target social group (or members of that group) and their subtle bias towards that social group (or its members) often manifested in microaggressions and micro-discriminations.

Biased egalitarians claim that they believe *p*, yet they behave in subtle ways which conflict with their avowed belief *p*. For example, research on prejudice finds support for a subtle yet very pernicious type of biased egalitarians labelled by Dovidio and Gaertner (1989, 2004) as *aversive racists*.[3] Aversive racists

> sympathize with victims of past injustice, support the principle of racial equality, and regard themselves as nonprejudiced, but, at the same time, possess negative feelings and beliefs about blacks, which may be unconscious. Aversive racism is…qualitatively different than the blatant, 'old-fashioned', racism, it is more indirect and subtle… (Dovidio and Gaertner, 1989, p. 3).

Aversive racists believe blacks and whites are equally intelligent, yet they choose white over equally qualified black job candidates, speak less to black colleagues, interrupt them more often, and avoid sitting close to them (Brownstein and Madva, 2012, p. 70). Such individuals are considered *implicitly* racist insofar as they are more likely to show *subtle* and *indirect* forms of behaviour which are discriminatory in nature. This is, obviously, different from outwardly bigoted racists (e.g. KKK members) who are explicitly biased and who behave in overtly discriminatory ways. According to orthodox explanations of this phenomenon, biased egalitarians act in covertly biased ways (even if they are overtly non-biased) because they harbour what is widely labelled as *implicit bias* and described as *unconscious, uncontrolled,* and/or *arational.* It is this notion of implicit bias with which I am concerned.

1.1 Objectives

This thesis aims to do two things. The first is to assess the dominant explanatory family of views which purport to explain biased behaviour, in particular that of biased egalitarians. The views I

[3] Although I speak of aversive 'racism' to illustrate the pernicious harm of microaggressions, there is similar evidence for microaggressions concerning gender, sexuality, religion, and other social categories.

have in mind hold that, over and above one's self-reported beliefs, called *explicit attitudes*, there's a distinct psychological kind called *implicit attitude* or *implicit bias*. Explicit attitude is typically tracked by certain kinds of *direct* measurement instruments (e.g. questionnaires or self-reports). Implicit attitude is captured using other measurement techniques which are referred to as *indirect* specifically because they do not require the agent to introspect, and indeed because they were conceived to avoid methodological issues concerning self-reports. According to these views, there is a fundamental distinction between explicit attitudes and implicit attitudes and the behaviour that is guided by each. This family of views endorses what I shall call *The Dualistic Alignment Hypothesis* (*dualistic hypothesis* for short).

The basis of the dualistic hypothesis – which unites all accounts falling under it – culminates in two main assumptions. The first is the idea that there's a principled distinction between explicit attitudes and implicit attitudes and the behaviour guided by each. The second is that there is an alignment in the explanatory role of explicit attitudes and that of implicit attitudes. Namely, there's an explanatory alignment between explicit attitudes and the behaviour they influence (including responses on direct measures like self-reports) on the one hand, and between implicit attitudes and the actions they guide (including responses on indirect measures) on the other hand. In effect, overt behaviour, such as prejudice, is taken to be distinct from covert micro-behaviours, and requiring different kinds of psychological explanations.

According to the dualistic hypothesis, overt behaviours and covert behaviours (as well as their underpinnings) are categorically different. On just about any normative ethical theory, the injustices caused by both overt prejudice and microaggressions are morally problematic to varying degrees. Spelling this out in moral terms, however, is not my concern in this work. Rather, my focus is to examine the dominant explanatory family of views which conceptualizes bias as distinctively either explicit or implicit and as underscored by distinct mental elements; I assess whether such views are appropriate. Specifically, I challenge the widely accepted understanding of bias as having either explicit or implicit underpinnings. I question the scientific development of our understanding of implicit bias, as well as its theoretical and empirical basis. Further, I show that philosophical accounts which endorse the dualistic hypothesis face serious challenges. Ultimately, I reject the dualistic hypothesis as the best explanatory theory of the phenomenon of bias.

The second aim of the thesis is to advance the workings of an alternative and novel explanation to the puzzling data. The hypothesis I propose is grounded in a more realistic understanding of social behaviour (including evaluations) insofar as it accommodates the fact that the psychological explainers of our behaviour are complex. I call this hypothesis the *mosaic view*. An essential contribution of the mosaic view is that it resolves the theoretical challenges faced by the dualistic hypothesis in large part because it does away with talk of *explicit* and *implicit* underpinnings of bias but also because it addresses the non-unified, non-stable, heterogeneous, and complex aspects of what psychology labels *attitudes*.

The mosaic view reconsiders the notion of attitude as it is currently used in research on bias. Psychologists generally conceive of attitude as a relatively stable evaluative tendency that can be measured directly (on a self-report) or indirectly (by inferring from a subject's performance on certain tasks). Yet, while orthodoxy treats an attitude as enduring and stable, there are conceptual and empirical reasons to view it as highly sensitive to context. The mosaic view does exactly that while offering a broader framework of the various cognitive elements responsible for our social behaviour. I argue that what underpins social behaviour, including overt and micro -level behaviours, are not two unified, stable, and distinct mental kinds, but rather a complex conglomeration of interacting elements of mind within a network I call a *stance*. The principal idea underlying the mosaic view is that social behaviour is best considered as the result of complex interactions between elements of a stance, activated differently across contexts, and interacting with background beliefs, commitments, values, and other mental states.

The mosaic view I advance is eliminative towards the notion of *implicit bias* as the orthodox literature conceives it. While I uphold the idea that there are implicit mental states and equally that there are biases, stereotypes, and discriminatory behaviours, I do not believe it is explanatorily helpful for the qualifiers *implicit* and *explicit* to apply or attach to biases as such. Rather, I defend the idea that these qualifiers may or may not attach to particular mental elements that are part of the stance or mosaic and which, depending on their interaction with various individual and contextual factors, may or may not manifest in biased behaviour.

1.2 **Overview**

I begin in chapter 2 by showing that the orthodox picture of social attitudes in its most general form endorses the dualistic alignment hypothesis. The dualistic hypothesis begins with a commitment to a distinction in psychological explainers and the types of behaviour they explain. Generally, the most commonly accepted division is between two psychological kinds of explainers, labelled as *explicit attitude* and *implicit attitude*, implicated in two types of behaviour: overt or macro-level behaviours and covert or micro-level behaviours respectively. The dualistic hypothesis assumes an alignment in the explanatory roles of each of the mental explainers and the behaviour they are thought to guide. On the one hand, there are what we tend to think of as explicit beliefs which manifest themselves in deliberative action and which may be measured directly with self-reports. On the other hand, there are also implicit attitudes with features of ‘implicitness’ (e.g. being operative in the absence of an agent’s awareness and control) which guide, and therefore explain, covert micro-aggressions, subtle behaviours, and responses captured by indirect measurement instruments. I show that the orthodox view of social attitudes in philosophy follows the commitments of the dualistic hypothesis, hence I call it the *dualistic view*. Even while orthodoxy may acknowledge that behaviour is multi-causal, i.e. that multiple factors are influential in behaviour production, it does so only marginally. When it comes to philosophical models, orthodoxy assigns distinct kinds of psychological explainers as principally operative in social judgement: a unified kind explains overt macro-level behaviour, and another psychological kind explains subtle micro-level behaviour.

In chapter 3, I question the grounds of the popularity of the dualistic view as I sketch a historical narrative of its development. My aim is to illustrate how studies of attitude built on and appropriated theories from cognitive psychology (particularly those related to selective attention and short-term memory) to explain attitudes towards socially marginalized groups. I emphasize that at this early stage there was no talk of distinct mental explainers for behaviour (i.e., *explicit* and *implicit*). The measurement instruments that were first developed to better understand, for example racist attitudes, were not intended to capture some distinct psychological kind. Rather, such measurement techniques grew out of the need to bypass subjects’ social desirability motives in order to tap into their – often morally objectionable – attitudes. Similar measurement techniques quickly developed, the most famous of which is the Implicit Association Test (IAT). In a short time, the dualistic view was bolstered by further reliance on dual process theories which took over the narrative aligning the language with a dualistic alignment framework. Talk of *explicit* versus *implicit* measures assessing distinct *explicit* and *implicit* mental states and processes had taken over the scene without much scrutiny to what exactly *explicit* and *implicit* really signified nor to what exactly was meant by an *attitude*. The narrative I suggest supports the idea that the dualistic explanation demonstrated by the orthodox accounts was arrived at by contingencies which had llittle to do with data and much to do with factors dictated by theoretical environments of the research and perhaps of the scientists themselves.

Having challenged the historical grounds of the dualistic view, I move to investigating, in chapter 4, its psychometric scaffolding. Firstly, I argue that positing a duality in the mental underpinnings (explicit an implicit attitudes) to explain the dissonance in the results of direct and indirect measurements is just one possible explanation. There are at least two alternative explanations for this dissonance. The dissonance in the results on the measurement instruments may be due to a difference in the structure and format of these instruments. Alternatively, it may be due to the psychometric weaknesses that afflict both direct and indirect measurement instruments. Secondly, I argue that current measurement instruments show low predictive validity, meaning that they are poor predictors of behaviour. If an instrument has poor predictive validity, its usefulness in predicting and explaining behaviour becomes questionable. To the extent that the dualistic hypothesis hinges (in no small part) on the results of measurement instruments, weaknesses in the psychometric properties of these instruments highlight potential drawbacks for dualistic explanatory accounts. At the very least, it leads to the conclusion that we must be wary of *post hoc* explanations of the divergence in the measurement results.

For the sake of argument, I next suppose that even with complex statistical modelling to adjust and account for the psychometric weaknesses, some variance in the results on the measures remains unexplained. In chapter 5, I ask, what explains this remaining variance? Here, I argue that explaining the dissonance as reflecting a duality in mental underpinnings is unsatisfactory. As part of my argument, I show that the dualistic view faces three important challenges. The first, explored thoroughly by Sophie Stammers (2016), I call the *unity challenge*. It points to the limitations of drawing a principled distinction between two mental kinds underpinning behaviour (an explicit and an implicit). The argument here is that there is no feature of cognitions which allows them to be carved neatly into the explicit-implicit distinction required by the dualistic view. The second challenge, partly discussed by Holroyd and Sweetman (2016), I call the *heterogeneity challenge*. It demonstrates that what the dualistic view classifies as two categories of uniformly structured mental elements (the explicit and the implicit) are not uniform. Rather, there is much heterogeneity that is not accounted for in the interpretation of dualistic theories. The last challenge which I call the *complexity challenge* is one which *I* direct at the dualistic view. This challenge comes from the complexity that is notably involved in behaviour production. The argument is that a complicated relationship between attitude and behaviour makes the mechanisms underscoring behaviour (such as responses on measurement tests) dramatically more complex than dualistic alignment interpretations presuppose. I conclude this chapter by claiming that the constructed mental categories *explicit attitudes* and *implicit attitudes* are in fact not unified, not stable, and not distinct. Rather there is much heterogeneity, a lack of uniformity, and dynamicity featuring in the mental elements responsible for social attitudes. When the empirical data is thus interpreted, the dualistic view is shown to be a weak explanatory theory.

In chapter 6 I offer the *mosaic view* as a novel alternative to the dualistic alignment hypothesis. On my view, what the dominant literature considers as ‘attitude’ is reconfigured as a mosaic (or a conglomeration) of elements of mind that constitute what I label the subject’s *stance* towards the target object. A stance is not a uniform mental entity but a conglomeration of heterogeneous mental elements responsible in part for our behaviour towards social groups. Its elements may show features of implicitness and explicitness (as current theorizing holds) yet they do not reliably cluster around explicit-implicit poles. Rather, they exhibit features typically thought to describe the implicit and the explicit (e.g. controllability, introspective accessibility, and reason-responsiveness) in varying degrees along a continuum. The intensity of activation of the elements of a stance depends on a network of interaction with each other and with other background beliefs as well as on the context. Being constituted by heterogeneous elements of mind featuring different degrees along a continuum characteristic of explicitness and implicitness, a stance resists measurement by current uni-dimensional instruments. Treating a stance as though it can be measured in such a way results not only in mistaken measurement results (at the very least an incomplete understanding of what a person’s stance is like) but it also generates a variety of needless puzzles. The notion of a *stance* is meant to replace talk of *explicit attitudes* and *implicit attitudes*. The mosaic view is well suited to adopt the complexity and multidimensionality that challenge the dualistic paradigm. It reconciles the conclusions of disunity, heterogeneity, and complexity drawn from the previous chapters insofar as it suggests a multi-dimensional conception of an agent’s stance towards its target object.

I conclude the thesis in chapter 7 by briefly sketching the ontological and practical implications of the mosaic view. Although I will have argued that there are implicit and explicit mental states, they do not cluster systematically around two dichotomous poles. As a result, the qualifying term ‘implicit’ does not attach to biases. Rather there are heterogeneous and non-unified mental elements which are part of a stance and which may manifest in discriminatory behaviour depending on various background psychological states and varying contextual factors. As I sketch an ontological picture, I suggest two ways of reading implicit bias, a light-touch and a heavy-duty reading. On the light-touch reading, implicit bias is read in the common-sense everyday notion of the term. The heavy-duty reading, however, requires theoretically filling out the notion of implicit bias along a dualistic explanatory framework. I am realist about light-touch implicit biases but eleminativist about heavy-duty ones. Orthodox literature tends to concern heavy-duty mental states, and so, I am eliminativist about its underlying psychology (namely, *implicit bias* and *explicit attitudes*). I end by briefly outlining how critical such an understanding of bias is for practical measures developed to mitigate biases and discrimination,

1.3 <u>**The value of the topic**</u>

Why consider bias a valuable topic for philosophical research? Certainly, discrimination and social injustice ought to be a matter of concern for philosophers regardless their moral theory. As philosophers, a vital question we should be asking is: what ought we do to mitigate the social injustices and harms caused by our biases? And it is this question that motivates my work. I think the full answer will require a lifetime of research (maybe more) and will need to address the role of structural causes as well as individual causes. In this thesis, however, I focus only on the role of the individual. Specifically, I examine the psychological underpinnings of biases that lead us to behave in unjust ways, especially those biases which seem to be unbeknownst to us, and which we seem to be unable to control. Before we tackle any ethical challenges related to the harms and injustices caused by bias, I strongly believe that we need to first understand what we are up against. Combatting injustices requires that we understand our enemy.

Although I believe that the effects that biases have on our behaviour are diverse and multifarious, a good place to start understanding them may be to consider their effects in matters of degrees of moral harm. That may be one way to unify what characterizes the injustices and harms that biases create and sustain. Some of these effects, like awkward body language, may be considered minor when taken in isolation; others, like treating members of a marginalized social group unfairly are evidently more pernicious; still other effects, like physically harming a person in virtue of their belonging to an outgroup, express the highest of moral deficiencies. According to many philosophers, even minor injustices, once accumulated, will substantially impact individuals, groups, and even societies. As Virginia Valian (1998, p. 54) puts it, "[I]n the long run, a molehill of bias creates a mountain of disadvantages" (see also Greenwald, Banaji, and Nosek, 2015, p. 560 and Levy, 2015, p. 803). Subtle offences, like fat shaming for example, and subtle micro-inequalities like not providing proper public bathrooms for the disabled, if iterated often and spread widely, will aggregate into macro-level social injustices. It is such systemic inequalities which make the understanding of bias an urgent issue.

1.4 <u>Methodology</u>

My approach in this thesis is multidisciplinary. My aim throughout is to consider closely what the experimental research tells us about the puzzling phenomenon of otherwise egalitarian individuals acting in subtle biased ways described as unconscious and uncontrolled. Therefore, in many respects, my thesis is grounded in empirical research. My arguments are based on a careful examination of empirical studies of implicit bias in social psychology. Some chapters, for example chapters 3 and 4, engage closely with empirical and theoretical research in social and cognitive psychology. Chapter 4 is largely devoted to examining the limitations of the psychometric research and the extent to which theorists build on it. As philosophers, we often tend to depend highly on empirical work without giving close attention to the methodological qualifications of the research. To the extent that psychometric complications have for some time plagued the empirical research on which the notion of implicit bias is constructed, and to the extent that dominant philosophical accounts of bias rely heavily on empirical work, a significant part of my research sets out to highlight and engage in the psychometric discourse.

In addition, my thesis is philosophically grounded. It critiques but also builds on philosophical accounts that claim to explain the psychological nature of implicit bias, in effect, its ontological status. I engage with the dominant philosophical accounts which have led to the current way of theorizing as I examine the dominant family of views of the cognitive nature of bias. I also examine some of the variously used fictional vignettes in these philosophical accounts in order to show that they reflect how much the literature relies on decontextualized snapshots of cases to build our philosophical intuitions upon. I believe that discussions that integrate both disciplines are vital for determining how we may proceed in our struggle against prejudice and social injustice.

Chapter 2

THE DUALISTIC ALIGNMENT HYPOTHESIS (THE DUALISTIC VIEW)

The Orthodox Framework for Understanding Implicit Bias

Introduction

The first part of the thesis (chapters 2, 3, 4, and 5) examines the dualistic alignment hypothesis. Firstly, that there is an explanatory alignment between explicit attitudes and the behaviour they guide on the one hand and between implicit [social] biases and the phenomena they manifest in action on the other hand.[4] Secondly, that there is a principled distinction between explicit attitudes and implicit attitudes. This conceptualization of implicit attitudes (and the micro-behaviours they give rise to) as distinct from explicit attitudes such as beliefs (and the actions they influence) is the hypothesis I am calling the *Dualistic Alignment Hypothesis* (*dualistic hypothesis* for short). A large part of my thesis focuses on questioning the details of, and putting pressure on, the dualistic hypothesis. In the second part of the thesis (chapters 6 and 7), I suggest an alternative picture of attitudes. I argue that what it is to be biased is complex and involves heterogeneous elements of mind (e.g. mental states and processes) which can re-configure into complex blends of psychology.

In this chapter, my aim is to discuss the dualistic hypothesis as an explanatory framework, to show that it is wedded to various underlying commitments, and to illustrate how orthodox philosophical views of social attitudes and bias endorse it. I begin in 2.1 by clarifying key concepts

[4] I draw a distinction between implicit attitudes towards *objects* like food (e.g. chocolate), clothing (e.g. miniskirts), alcohol, brands, behaviors (e.g. smoking), individuals (e.g. Donald Trump) and implicit attitudes towards *social kinds* (or members of social kinds) such as races, ethnicities, religions, people from LGBTQ community, people with disabilities, obese individuals, etc. The latter type *implicit social attitudes* or *bias* are my main focus in this thesis. Any future use of the term implicit bias refers to implicit social bias specifically (unless otherwise specified).

which I will employ. In 2.2, I develop a clearer understanding of the phenomenon in need of explanation by focusing on *microaggressions*.[5] I discuss how microaggressions are characterized in the literature and how they are distinguished from other discriminatory behaviour. In 2.3, I provide a sample of some of the empirical evidence revealing the pervasiveness of microaggressions. After which in 2.4, I discuss the dominant hypothesis that explains microaggressions, i.e. the *dualistic alignment hypothesis*. I expand on the dualistic hypothesis and describe its commitments to three types of dualisms which align in a parallel manner.[6] Finally, in 2.5 I show that the orthodox philosophical picture in its most general form endorses the dualistic hypothesis by presenting some of the most discussed accounts of philosophers explaining microaggressions by appeal to implicit bias.

2.1 Terminological considerations

I begin by disambiguating key notions that are central to the dialectic and identifying how I'll be using them. Firstly, we must be cautious that the term 'attitude' is used differently in philosophy and in social psychology. For philosophers of mind, an 'attitude' (such as a belief, hope, desire, fear) is a cognitive relation towards a proposition (or a mental representation) whose structure is required to have a truth condition. The attitude component refers to how the proposition is being taken. For example, 'Raya *believes* that it will rain on Wednesday', 'Jad *hopes* that the fridge is

[5] I narrow my focus to microaggressions (rather than the whole range of micro-behaviours which might be explainable by appeal to implicit bias) because I am concerned with the injustices that marginalized groups incur in light of these microaggressions. By other types of micro-behaviours that are not microaggressions, I mean those which advantage members of the dominant group over those in the marginalized groups. For example, giving more credibility to a white man over a black woman (see Fricker, 2007).

[6] Of course, depending on how we characterize them, there may be multiple ways to carve up the spaces of behaviour, measurement devices and social cognition. I restrict the distinction here to dual levels of behaviour (micro and macro) insofar as these two types of behaviours are predicted by responses on the two types of measurement instruments used by empirical psychology. One type of instrument captures the kind of stereotypes we are said to harbour when we respond on Implicit Association Tests for example or the kinds of mechanisms that become activated on priming tasks, and the second type of instrument captures self-reports which are then taken to be reflective of explicit attitudes.

full of food', 'Lynn *desires* the cat to sit on her lap', and so on. These are examples of what philosophers mean by *attitude* or *propositional attitude*. For psychologists and philosophers interested in implicit attitudes and implicit bias, however, the term 'attitude' does not refer to a propositional attitude in the philosophical sense. Rather, it refers to a subject's evaluative mental state directed at a certain object. For example, in social psychology, an attitude has been described as "the intensity of positive or negative affect for or against a psychological object" (e.g. 'I hate Muslims') (Thurstone, 1946, p.39 cited in Fazio and Olson, 2003, p. 140). Attitudes have also been described in psychology as the "psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavour" (e.g. 'Muslims are bad') (Eagly and Chaiken, 1993, p. 1).

Consistent with psychologists' definition, attitudes are measured using unidimensional scales (or single line scales such as the Likert scale with a measurement of liking from 1-7). On such scales, subjects are asked about their *attitude* towards various issues ranging from social phenomena (Brexit) to individuals (Donald Trump) and social groups (Muslims, obese individuals), and from behaviours (alcohol drinking) to consumer products (iPhones). Indeed, social psychology may be considered a subdiscipline of psychology that emerged to measure attitudes of various kinds (Allport 1935). Moreover, it is common practice in areas of philosophy interested in bias to use the term *implicit attitudes* and *implicit bias* somewhat interchangeably. Neil Levy (2017, p. 535), for example, uses *implicit attitudes* to mean a broad category of elements of mind that are "opaque to introspection and escape direct control" and *implicit bias* to mean a subset of this category differentiated by its "prejudicial content". It is equally common in empirical psychology to treat attitudes as cognitive states exerting some influence on behaviour (Allport, 1954).[7] I shall conform to these trends. In chapter 6, I provide a critical analysis of the historical change in the description of the notion of 'attitude' and I propose a reconsideration of this notion. Until then, however, I adhere to the literature's common usage of the term.

Secondly, while I engage in examining and critiquing the orthodox dualistic view in the first part of the thesis – chapters 2 to 5 – I use *explicit attitude* and *implicit attitude* in the same way as the literature describes. For example, when I use the term *explicit attitude,* I mean it as a kind of mental state that sits within a homogeneous range of introspectively accessible and controlled mental states which may be tracked by direct measurement devices and which influences a homogeneous range of behavioural outputs.[8] By *implicit attitude* or *implicit bias*, I mean a kind of mental state that sits within a homogeneous range of mental states which may be tracked by indirect measurement devices and which (to a large extent) drives a homogeneous range of behavioural outputs. Relatedly, I refer to overt bigoted acts (e.g. a restaurant owner who refuses to serve black customers because they are black) as *macro-level* or *overt behaviours* and to subtle biased micro-

[7] Most philosophers treat implicit bias as mental states; for example, Mandelbaum (2016) considers implicit attitudes as beliefs, Levy (2015) treats them as 'patchy endorsements', and Gendler (2008a, b) defends a *sui generis* version of implicit bias which she calls *aliefs*. An exception to this is Edouard Machery (2016) who argues that attitudes are traits.

[8] The dualistic view is certainly committed to the idea that implicit attitudes influence micro-behaviours, and it may not consider macro-behaviours to be causally isolated from the influences of implicit attitudes. In other words, there may be a crisscross in alignment from implicit attitudes to macro-behaviours, and multiple causal pathways for macro-level behaviours. For example, a person's implicit sexism may be involved in micro-acts like interrupting more women than men (a micro-behaviour) but also in explicitly deciding to vote for a male candidate rather than a more qualified female candidate because of her gender (a macro-level behaviour). Certainly, there may be several elements of equivocation that make this more nuanced than meets the eye. Although some supporters of the dualistic hypothesis may reject the idea that macro-behaviours are causally isolated from implicit attitudes, they mostly endorse the idea that implicit attitudes are tracked only by indirect measures, even if they are efficacious in the production of macro-level behaviours. I believe this is an element of dualistic models that remains a main contention as current discussions are pushing towards a less principled dichotomy.

behaviours (e.g. a shop owner who locks the till only when a black customer walks in) as *microaggressions*.

Thirdly, biased behaviour may take on two qualities, a positive quality when it results in "excess" of advantages towards people *qua* their social identity, or a negative quality when it results in a "deficit" in these advantages or in disadvantages towards people *qua* their social identity (Fricker, 2007, p. 17). The notion of bias that shall concern me in this thesis is in the latter sense because of its connection with injustice and because the underlying motivation of my work is the mitigation of social injustice (see footnote 6).

Finally, by *explanatory alignment*, I mean a type of parallel manner of explanation of the responses on the measurement devices and other types of behaviour. On the one hand, explicit attitudes are thought to explain the responses on direct measures (such as self-reports) and to explain/predict the behaviours these attitudes generate. This corresponds to the cells on the right side of figure 1. On the other hand, implicit attitudes are taken to explain the responses on indirect measurement devices (such as the Implicit Association Test (IAT) and to explain/predict the micro-behaviours produced (e.g. microaggressions). This corresponds to the cells on the left side of figure 1. As shown below, there are dual explanations that run from psychological state to behaviour.

Figure 1

Level of explanation		
Behaviour	Macro-behaviours (Overt)	Micro-behaviours (Subtle/Covert) e.g. Microaggressions
Measurement Responses	Responses on direct measures (e.g. self-reports)	Responses on indirect measures (e.g. IAT scores)
Measurement Devices	Direct measures	Indirect measures
Psychological state	Explicit attitude (or explicit bias)	Implicit attitude (or implicit bias)

Figure 1: The Dualistic Alignment Hypothesis - a schematic representation of the parallel explanation of the responses on direct and indirect measurement devices.

2.2 The phenomena in need of explanation

In this section, I elaborate on how microaggressions are described and how they differ from other types of behaviour. I begin by explaining the nature of microaggressions as the puzzling behaviour that theorists of attitude are faced with. I discuss ways of distinguishing them from knee-jerk reflexive reactions and from macro-level behaviour. I suggest that although macro-level and micro-level behaviours are considered categorically distinct by the dualistic alignment hypothesis, they may only differ *qualitatively* insofar as they lie on different areas along a spectrum of harms.

Psychologist Derald Wing Sue (2010, p.110) considers microaggressions as “subtle verbal or behavioural slights, snubs, or indignities” which communicate “hostile, derogatory, or negative messages” to individuals due to their marginalized identities. Sue and his colleagues (2007, p. 273) argue that although the influence of a single microaggression might seem “innocuous and innocent”, their accumulation across time and contexts can be harmful to the individual at the receiving end, both psychologically and physically, and can be detrimental to a just society as it leads to the perpetuation of social inequalities.[9] Sue explains that microaggressions are:

> insidious, damaging, and harmful forms of racism or sexism [that] are… everyday, unintentional, and unconscious… [and that] are perpetrated by ordinary citizens who believe they are doing right (Sue, 2005, p. 108).[10]

Sue's treatment is very much in line with McConahay's (1986) use of the term 'modern racist', Sears's (1988) 'symbolic racist', Dovidio and colleagues' (2002) 'aversive racist', and Holroyd's (2015) 'biased egalitarian' to describe outwardly well-meaning individuals who nonetheless exhibit racist behaviour not aimed to promote harm.

Microaggressions, thus, may be regarded as manifestations of modern/symbolic/aversive bias. They may be expressed even by those who explicitly avow egalitarian beliefs, i.e. biased egalitarians. Consider as an example of microaggressions again the Starbuck's manager who calls the police on the two black men who were sitting at a table without purchasing anything. Or take a store owner who follows a person of colour around the store because they seem to her to be suspicious, or a white person who sits/stands further away from a black interlocutor and claims that

[9] Micro-aggressions and micro-discriminations exert a range of subtle harms even if they constitute minor or trivial behaviours. As Lane, Kang and Banaji (2007) argue, behaviour such as body language is hardly in and of itself harmful, but its impact, once aggregated and "summed over large populations engaged in daily interactions and evaluations", may be substantial enough to constitute macro-level injustices and social inequalities (Lane, Kang and Banaji, 2007, p. 441). Additionally, Virginia Valian (2010, p. 323) worries that one might be tempted to dismiss concerns about microaggressions "as making a mountain out of a molehill". "But", she adds, "mountains *are* molehills, piled on top of one another over time". Valian gives the example of the testimonial injustice inflicted on women who, when they make a suggestion in meetings, are ignored only to hear their suggestion repeated minutes later by a male colleague and immediately taken up and welcomed as a great idea. Women who voice their complaints about such injustices are often told that such things are insignificant and that they are being overly sensitive. But these subtle acts of gender-based (or racial-based) injustices, disdain, or condescension that often lead to micro-discriminations in time can accumulate into macro-level injustices.

[10] Although Sue (2005) does not here mention other forms of prejudice and injustice such as homophobia, classism, ableism, ageism, and others, they are explicitly mentioned elsewhere in his book.

‘blacks and whites have equal opportunities for achievement in our society’. Such acts are all considered microaggressions and all have the moral character of being experienced by the target person as significantly harmful.[11]

How are microaggressions and knee-jerk reflexive reactions distinguished? Microaggressions, though may be non-visible to the discriminator, can nevertheless be detectable by the person they target, and the harm they cause the person is often potently experienced (Kraus and Park, 2017). It is much easier for a woman to notice that her male colleagues continuously interrupt her in meetings than she is to notice that her colleagues’ pupils have dilated. Reflexive, knee-jerk type behaviours like dilated pupils (in isolation from other behaviours like frequent interruptions) rarely have a direct impact on the target person. The target person might not even realize that the discriminator’s pupils have dilated or that their heart rate has increased.

Moreover, knee-jerk reflexive behaviours, such as pupillary dilation, are simple autonomic nervous system responses that cannot be reshaped or eliminated; they are what Brownstein (2017, p. 5) calls, “untrainable”. In contrast, microaggressions, at their very basis, are not simple, rather they involve a complexity of social behaviours. Autonomic nervous system responses may

[11] If you fail to see how these are examples of microaggression and how they lead to moral, physical (and often epistemological) harm, consider that the term ‘microaggressions’ was first coined by Chester Pierce, a Harvard psychologist working with Black Americans to explain the race-related offences and indignities that his clients experienced on daily basis. According to Pierce,

> …the most grievous of offensive mechanisms spewed at victims of racism and sexism are microaggressions. These are subtle, innocuous, preconscious, or unconscious degradations, and putdowns, often kinetic but capable of being verbal and/or kinetic. In and of itself a microaggression may seem harmless, but the cumulative burden of microaggressions can theoretically contribute to diminished mortality, augmented morbidity, and flattened confidence (1995, p. 281).

If you are still having difficulties seeing why these examples are microaggressions, refer to Virginia Valian (1998) *‘Sex, Schemas, and Success: What’s Keeping Women Back?’* and Christina Friedlaender (2018) *‘On Microaggressions: Cumulative Harms and Individual Responsibility’*. For more examples of microaggressions, see Appendix 1 from D. W. Sue 2010 *‘Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation’*. For a differing view, refer to Lilienfeld (2017).

accompany microaggressions, but by themselves they don’t make a microaggression, nor do they manifest in prejudiced behaviour. As Eric Mandelbaum describes it:

> microbehaviors look decidedly different than paradigmatically reflexive behaviors, such as the deep tendon reflex that controls knee jerks. One has to *decide* where to sit, and even if such decisions are unconscious, they are subject to decision-theoretic processes in a way knee-jerk reflexes aren’t (Mandelbaum, 2016, p. 649).

Consider a biased egalitarian white woman who encounters a black man alone in an elevator. This minimal social interaction may lead her to engage in complex behaviours including clutching her purse, perhaps smiling anxiously into space, or maybe shuffling or tapping her feet. Simultaneously she may start experiencing an increased heart rate or dilated pupils. Crucial are the complexities involved in this woman’s behaviour: it is not simply that she experiences an increased heart rate and dilated pupils. These things do not happen in isolation. Rather, the woman’s experience is complex and multifaceted, and along with simple autonomous nervous system reactions, it also involves the clutching of the purse and other physical acts. If Mandelbaum and I are right in our analysis, then what concerns any discussion of morally relevant attitudes like those related to persons from marginalized social identity groups has little to do with reflexive knee jerk behaviours and much to do with more complex type behaviours like microaggressions and overt macro-level behaviours.

How are microaggressions distinguished from overt discrimination? Although microaggressions and overt discrimination are both harmful and implicated in social injustice, the literature considers them as somehow categorically different insofar as they require different explanations. I discuss two ways in which the literature may consider them different. Then I argue that there is no reason to think microaggressions are a phenomenon *categorically* distinct from overt bias. They may just lie on a spectrum of harms.

Firstly, there is what might be called an *distinction with respect to epistemic access* between microaggressions and overt-type behaviours.[12] Whereas microaggressions *seem* to many (namely to the perpetrators) to be ambiguous and non-visible, overt or macro-level discriminations are displayed for the world to see; they *announce* an individual's antipathy or displeasure towards someone usually from a marginalized identity group. A typical example of overt racism is a hate crime where an overtly biased person intentionally causes harm to another individual because she belongs to a different race; for example, refusing to rent an apartment to a family for the sole reason that they're black.[13]

[12] It is in the sense that macro-level behaviours are easily recognized whereas microaggressions may be more difficult to know (at least from the perspective of the perpetrator and perhaps even the spectator). It is from this angle (the epistemic accessibility) that I am referring to the distinction as epistemic.

[13] Haidt (2017) notes that a shortcoming of microaggression research is its insufficient emphasis on subjective appraisal of harm. But his worry is that the onus of responsibility of harm rests squarely on the perpetrators of harm while little concern is given to the role of the 'victim's' personality traits in influencing the [victim's] responses to perhaps ambiguous situations. The worry is that some people may, due to their personality, evaluate *ambiguous* situations as microaggressions. Although this sounds like a plausible consideration, it doesn't undermine the claim that there's plenty of micro-aggressive behaviour around given the expanding evidence of the injustices which plague otherwise purportedly egalitarian societies. Although it may be the case that overly sensitive personalities interpret ambiguous situations as microaggressions, it is unlikely that this accounts for much. Racism, sexism, homophobia, Islamophobia, anti-Semitism and other forms of prejudices against traditionally marginalized groups (whether manifested overtly or covertly) are imbued with a continuing history of violence and injustice that to view them as subjective interpretations of the person at the receiving end downplays the experiences of the target in understanding the discrimination perpetrated against them. It also ignores the significant power dynamics at play when microaggressions play out. To decontextualize and de-historicize microaggressions is to misconceive them as localized aggressions that are divorced from their context and history. Microaggressions are best seen as a part of the large context of continuing injustices against traditionally stigmatized groups.

Secondly, there may be a distinction in *intention to harm* between the two types of behaviour. Microaggressions are thought to differ from overt discriminations because they are not behaviours *aiming* to promote harm, although they do in fact harm.

Thinking about microaggressions and overt behaviour as distinct categories is nowhere as clear as it is in cases of harmony or what Holroyd (206, p. 159) refers to as "non-conflict" cases. In non-conflict cases, microaggressions are thought to be distinct from, but align with, blatant discriminations as would be in cases of bigoted racists who express their racism overtly as well as covertly through microaggressions. Holroyd (2016) notes that such cases are not given much attention in the literature. I too shall give them less focus, (but I mention them here to demonstrate how the literature affirms a duality between overt and covert discrimination). Instead, I focus more on disharmonious cases because they are the cases that draw the most attention, and because their disharmony according to Holroyd, Scaife, and Stafford (2017, p. 6) "is perhaps the most striking feature of paradigm cases of implicit bias". It is simply not as interesting, puzzling, or thought provoking if a bigot were to manifest her racism covertly, but it is surprising to think that someone who endorses egalitarianism will still act in covert racist ways manifested in microaggressions.

What I want to suggest from the discussion so far is that microaggressions may not be *categorically* different from overt bias and thus need not require categorically distinct psychological explanations. Of course, it can't be stressed enough that the difference between overt bias and microaggressions is a *morally* significant difference. My point, however, is simply to question whether we ought to consider the difference to be a matter of *qualitative degree* rather than a matter of *distinct categories* of behaviour in need of distinct explanations. The suggestion that overt prejudice and microaggressions fit better on a continuum with various degrees of harmful acts in between (rather than being considered as qualitatively and categorically distinct types of behaviours in need of distinct explanations) will substantially inform my investigation in this chapter. As shall become clear shortly, microaggressions (and overt bias) may best be considered on a prejudice spectrum guided by a complex array of mental states and processes.

2.3 Empirical evidence for microaggressions

I review in this section some of the empirical evidence that reveals the pervasiveness of micro-behaviours, specifically microaggressions, in different aspects of life. First, I briefly introduce the idea behind the measurement devices used in the empirical research that I review (a thorough explanation of the measurement instruments will follow in chapters 3 and 4). I then review some of the more commonly discussed studies in the implicit attitude research.

As discussed above, psychologists looking for the underpinnings of microaggressions dismissed the possibility that these might be caused by subjects' explicitly reported attitudes. They reasoned that subjects harbour biases which they are *unwilling* to report (because they wanted to conform to social norms of equality) or biases which they are *unable* to report (because these biases are unconscious) (Gawronski, *personal correspondence*; see also Carruthers, 2018, p. 2; Payne and Gawronski, 2010, p. 6; Levy, 2016, p. 3). Psychologists looked into developing advanced measurement techniques to tap into these biases and their underpinnings.

Some of these measurement techniques, typically called *indirect measures*, involve subliminal priming that is not consciously detected within the experimental setting (e.g. sequential priming tests, Fazio et al., 1995). Others use reaction time differences between tasks thought to be consistent and tasks thought to be inconsistent with a known stereotype (e.g. Implicit Association Test (IAT), Greenwald and Banaji, 1995). These instruments measure the subject's responses (which, using statistical techniques and certain interpretations by the researchers, are transcribed into numerical scores on these measurement devices) and by proxy are considered to track the psychological explanations of the micro-aggressive behaviour. Below, I offer some examples from empirical research which have contributed to the rise of the notion of implicit bias as it is understood by the dualistic hypothesis.

Research on "shooter bias" computer simulation shows the impact of race on weapon identification and the decision to shoot (Payne, 2001; Correll et al., 2002; Correll et al., 2014; Plant and Peruche, 2005; Payne, 2005, 2006). In such experiments, participants are shown photographs of men holding an object (e.g. a gun or a wallet) with different city landscapes in the background. The participants are asked to shoot only if the object shown in the man's hand is a weapon, the time they're given to respond is limited to 500ms (Payne, 2006). Results show that participants are faster and more likely to shoot an unarmed black person than an unarmed white person, and more likely *not* to shoot an armed white person than an armed black person. This is taken to suggest that a prime of a black person enhances the tendency to shoot, consequently implying an association between BLACKS and DANGER or THREAT, or at least some kind of mental element which is likely implicated in the shooting behaviour.

It is worth noting that the above studies are based on computer simulations rather than on real world situations. Real world incidents of shooter bias often involve police having to act in ambiguous situations. And while researchers of the shooter bias paradigm accept that "laboratory studies involve impoverished simulations of complex behaviours", they also acknowledge that "it is difficult to demonstrate a causal relationship between suspect race and police use of force based solely on correlational data" (Correll et al., 2014, p. 202). They discuss mediating factors that need to be considered when interpreting results. Such factors include the need to respond quickly (e.g. response time around 500ms), the ambiguity of the situation (e.g. the object in the man's hand is not clearly visible), the presence of environmental cues (e.g. the background being a dark alley or an office) and personal factors which may affect police decisions (e.g. fatigue, fear, and cognitive load) (ibid). Be that as it may, the studies are taken to show a certain degree of systematicity in the behaviour of these individuals, which we ought to expect to see exhibited in police officers (since there's no pre-theoretical reason to think that they would react differently).

Research in court hearings shows that stereotypic priming influences jurors such that they are more prone to convict a black person than to convict a white person even with identical incriminating evidence, and that they are more likely to give harsher sentences to black defendants than white defendants (Faigman et al., 2012; Kang and Lane, 2010). In a simulated jury study, Levinson, Cai, and Young (2010) found that participant (undergraduate students) jurors who held strong associations between the concepts BLACKS and GUILTY on IATs were more likely to judge ambiguous evidence as incriminating of guilt for a black defendant more so than a white one, and this was the case even with participants who reported feeling warmer towards blacks. In another study, participants were shown photographs of a crime scene, one of which featured a masked gunman with a visible arm and hand (Levinson and Young, 2010). The race of the gunman was manipulated for two randomized groups, the skin colour of the hand and arm was either dark or light, and the trial evidence presented was designed to be ambiguous as to whether the defendant was indeed the gunman. The jurors were asked to evaluate the evidence and decide if the defendant was guilty or not. Findings showed that participants were more likely to convict a black man than a white man suggesting that some type of bias influences how jurors assess the evidence in the case. Evidence for this type of bias was shown even in participants who exhibited low explicit racism.

Similarly, data from research on hiring decisions shows that participants are found to give more favourable evaluations to the resumes of applicants with white-sounding names than to identical resumes of applicants with black-sounding names. Crucial to note is that the findings are not explained by the participants' avowed beliefs. In a field experiment, researchers sent resumes to 'help wanted' adverts in newspapers and measured callbacks for interviews for each sent resume (Bertrand and Mullainathan, 2004).[14] The credentials of the applicants on the resumes were varied and their race was manipulated via fictitious names (e.g. either white-sounding such as *Emily*, *Meredith*, *Geoffrey*, and *Greg* or black-sounding names such as *Aisha*, *Kenya*, *Rasheed*, and *Tyrone*). The results indicated that employers who were considered to be restricted by affirmative action laws, as well as those explicitly stating they are "Equal Opportunity Employers" on their adverts, showed preferential treatment for resumes with white-sounding names over ones with black-sounding names. The racial differences in callback rates was statistically significant in that applicants with white names needed to send only 10 resumes to get a callback while black applicants needed to send 15 resumes before getting a callback. Black applicants needed to send more resumes to get a reply *irrespective* of their higher credentials and other variables. Similar discrimination in callback rates was found in favour of Swedish names (where the research is done in Sweden) over Arab-Muslim names (Rooth, 2010).

As with racism, so is the case with sexism. Empirical evidence shows micro-discrimination based on gender – often in the absence of explicit sexism – in various experiments. An interesting

[14] Field experiments refer to real world conditions as opposed to testing under artificial laboratory conditions.

line of research on sexism talks about “constructed criteria” or “shifting standards”. This kind of research emphasizes the “malleability of merit” or judgement standards (Uhlmann and Cohen, 2005; Faigman et al., 2012, p. 1156). For example, in one experiment by Uhlmann and Cohen (2005), participants were asked to evaluate two final candidates (a male and a female) for the position of chief of police. The profile of the two candidates suggested that they were either “educated” or “streetwise” and the trait was attached either to the woman or the man. Results showed that (on average) male participants chose the male candidate over the female candidate (regardless of the candidate’s profile), while both male and female participants defined the hiring criteria according to the profile attached to the male candidate. For instance, the “educated” profile was considered to be a more important criterion when the man had it and the “streetwise” criterion was deemed the more important when the man had it. Subjects who described their decision as objective were more likely to engage in “biased hiring criteria” (Uhlmann and Cohen, 2005, p. 3).

In sum, clear and wide-ranging empirical evidence seems to suggest that many individuals engage in certain racist or sexist behaviour that may be in tension with their explicit commitment to racial or gender equality and their sincere disavowal of racial or gender-based discrimination. Note here that the studies clearly implicate what has above been described as microaggressions. This is only a sample of the wide and growing literature *interpreted* by theorists who endorse the dualistic hypothesis as highlighting and implicating certain psychological influences. Specifically, the sample is interpreted as showing that some psychological kind distinct from explicit attitudes such as beliefs is explanatory of the findings of the micro behaviours and judgments. Obviously, the significance of such empirical research underscores the importance of recognizing the various complexities involved in judgments and decision-making in real world situations. As we shall see shortly, such complexities highlight a tension in our understanding of social judgements, namely, in how they are measured, how they manifest in behaviour, and in how they are explained.

2.4 The dualistic alignment hypothesis: Underlying assumptions and commitments

In this section I expand on the dualistic alignment hypothesis and illustrate how it commits to various assumptions. First it assumes two distinct types of behaviours: macro and micro behaviours (discussed in section 2.2). Second, it assumes two distinct types of measurement instruments (direct and indirect measures) each of which captures two distinct kinds of behavioural responses. Finally, it commits to a corresponding dualism in the psychological explainers of the responses on these measurement instruments.

The dualistic alignment hypothesis may be said to explain the phenomenon of microaggressions as displayed by biased egalitarians and aversive racists. (It also explains interesting cases of dissonance in measurement results on direct and indirect measures including those discussed in section 2.3). For example, it explains how some males may profess egalitarianism (their response on self-reports is egalitarian and they consciously and sincerely endorse egalitarian beliefs), yet still harbour bias towards women. This negativity is captured by their responses on indirect measures and manifested in acts of microaggression such as hiring decisions where candidates have identical CVs. They simply hold explicit egalitarian attitudes and implicit sexism, each of which is responsible for a different kind of behaviour.

The dualistic hypothesis seems also to explain why black and white (or female and male) interlocutors often disagree as to the existence of microaggressions. Whereas a man typically views himself as non-sexist based on *what* he says in the interaction with a woman, the woman is likely to judge the man to be sexist on the basis of *how* he covertly or subtly acts (i.e. based on his micro-behaviour or his subtle non-verbal cues such as seating distance, body postures, etc.) (Dovidio et al., 1997). Thus, the dualistic hypothesis makes various commitments about differing behaviours (including different results on the two measurement instruments) and about their psychological underpinnings.

Firstly, from the discussion in section 2.2, we may claim that the dualistic hypothesis accepts a distinction in two types of behaviours in need of explanation. Namely, it commits to:

(Commitment 1)

A dualism in the modes of behaviours in need of explanation:

- *Macro-level, blatant, overt behaviours such as outright bigotry*.
- *Micro-level, subtle behaviours such as microaggressions.*

That blatant discriminatory acts are the result of outright prejudice (and the kind of attitudes we generally attribute to persons such as beliefs) is an undisputed matter: racist/sexist/homophobic beliefs produce racist/sexist/homophobic behaviour. Nonetheless, when it comes to subtle behaviour or microaggressions, the picture is significantly more puzzling. The dualistic hypothesis takes micro-behaviours to be *un*explainable by the same psychological underpinnings of our overt actions and consequently seeks other psychological drivers. Indeed, cases of disharmony between a person's measured reported attitudes and their subtle behaviour are taken to attest to the un-explainability of microaggressions with attitudes such as beliefs. Furthermore, cases of disharmony may be the motivating force behind the search for underlying psychological explanations of microaggressions. If not the attitudes that we attribute to persons (such as their professed beliefs) that are responsible for microaggressions, then what are the psychological underpinnings of this covert behaviour? For the dualistic hypothesis, what underwrites microaggressions is a psychological kind *distinct* from that which underwrites overt prejudiced behaviour such as self-reports.

Before I discuss the second commitment related to the methods in which psychologists traditionally track people's attitudes, let me first examine these methods briefly. To track subjects' beliefs, psychologists examine personal reports (self-reports) and overt actions which are taken to reflect these beliefs. For example, they assess a subject's attitude towards women *directly* by asking the subject whether she claims egalitarianism towards women, professes anti-sexist beliefs, or evaluates women positively. The subject's response is taken to indicate her *explicit attitude* towards the target object (in this case women). Explicit attitudes, then, are assessed using *direct* instruments such as self-reports. These instruments are called *direct* because they are self-assessments of the respondent's attitude that can be conveyed/reported directly and overtly in a verbal or written manner (Gawronski and De Houwer, 2014).

Since empirical evidence shows microaggressions to sometimes dissociate from explicitly reported beliefs, some psychologists (and later philosophers) hypothesized that microaggressions must be the behavioural manifestations of psychological mechanisms that are distinct from the mechanisms which generate beliefs: they call these psychological underpinnings *implicit attitude* or *implicit bias*.[15] Their reasoning is that if a subject explicitly endorses a certain attitude but behaves in ways that diverge from it, their behaviour must be caused and guided by some distinct

[15] Carruthers (2018) elaborates on this type of reasoning in psychology. He writes, "In much of the social psychology literature on (explicit) attitude formation and attitude change, an attitude is thought to be a disposition to make a certain type of evaluative statement….and on a similar dispositional reading of implicit attitudes, such an attitude would be a disposition to engage in certain sorts of non-verbal evaluative behaviour. On this construal, it follows immediately that explicit and implicit attitudes are distinct…because they are distinct types of behavioural dispositions" (p. 2).

bias that is *implicit* (typically 'implicit' is fleshed out as unconscious and unintentional).[16] In order to assess these *implicit attitudes*, psychologists use methods which are taken to *in*directly tap into them. The measures are often called *indirect* because they do not require the subject to introspect her attitudes and report them. They may also be called *indirect* because the responses on these measurement instruments are used as proxies to *infer* the existence of psychological underpinnings that give rise to the response.[17] In any case, implicit attitudes are said to be tracked using *indirect* measurement devices that don't require the subject to report on her attitude, e.g. the Implicit Association Test (IAT) (Payne and Gawronski, 2010; Carruthers, 2018; Nosek, 2007; Fazio and Olson, 2003; Brownstein, 2017; Gendler, 2008; Shoda, McConnell, and Rydell, 2014; Calanchani and Sherman, 2013; Gawronski and Bodenhausen, 2011; 2014a, b; Gawronski, Brannon, and Bodenhausen, 2017; Neslon, 2009).[18]

[16] Keep in mind that explicit and implicit attitudes may be in harmony such as in Holroyd's (2015) non-conflict cases.

[17] De Houwer and colleagues (2009, p. 8) distinguish between the response observed on a measurement procedure (e.g. the speed, latency, and/or accuracy of response on an IAT – i.e. the raw data) and the outcome measure that is derived/inferred from the responses (e.g. the IAT score, or the self-report). Deriving the latter from raw data involves several statistical techniques, but also some theoretical interpretations. According to this interpretation, for direct measurement devices such as self-assessment procedures, the outcome measure is based *directly* on the response of the participant, hence the name *direct* measures, but for indirect procedures, the outcome involves additional interpretation of the response/raw data by the researcher (e.g. the interpretation of reaction time performance on the IAT), hence the name *indirect* measures.

[18] The literature on implicit attitude often uses the terms *direct measures* and *explicit measures* and the terms *indirect measures* and *implicit measures* interchangeably. However, Jan De Houwer (2005) draws an important distinction between direct (indirect) measures and explicit (implicit) measures that complicates matters. He explains that the terms *direct* and *indirect* refer to the "objective properties of the measurement procedure"; direct measures involve the participants' self-assessment of their attitude but in indirect measures, the attitude is inferred from another behaviour (p. 18). To determine whether a measurement tool is direct or indirect, one simply needs to look at the procedure. However, whereas direct/indirect refers to the objective properties of the procedure, the terms *explicit* and *implicit* refer to the "functional properties of the measurement outcome" or the response on the measurement test. The features of a measurement outcome or response include features of intentional control and awareness. He further notes that "not all indirect measures produce outcomes that have functional features typical of implicit measures…. Likewise, direct measures do not by definition provide explicit outcomes" (ibid, p. 19).

I will have more to say on direct and indirect measures of attitude in chapters 3 and 4. At this stage I mean only to highlight the second underlying commitment of the dualistic hypothesis.

(Commitment 2)

Two distinct types of measurement procedures track two distinct types of responses:

- *Direct measures such as self-reports track explicit responses which in turn are thought to predict explicit overt behaviour.*
- *Indirect measures such as the IAT track implicit responses that are thought to predict covert behaviour such as microaggressions.*

The function of these two types of measurement procedures is to track the responses, to interpret these responses in a meaningful manner (e.g. to interpret quantitative scores of latency and accuracy so as to have a meaningful measurement from which to infer an attitude), and to *infer* from these responses the underlying psychological explainers.

Thus, a possible (and perhaps practical) step from a duality in behaviour and a duality in measurement outcome or response is to assume a corresponding duality at the level of psychological explainers (see footnote 16). In effect, this is the step which the research typically follows; the literature uses the measurement responses as proxies for psychological items (we will see in chapter 6 that such a step does not come for free and that researchers advocating the dualistic hypothesis owe us an explanation for how to justify this step). This will serve as an important consideration as we continue: namely that a main motivation behind hypothesizing a duality in the underlying psychological explanations (explicit attitudes and implicit attitudes) are cases of dissonance between a person's professed beliefs and their micro-behaviours or between their self-reported beliefs and their responses on indirect measurement instruments. Otherwise, why would one not consider microaggressions to be another behavioural manifestation of prejudice, especially given the changing realities of prejudice and how it evolves.

To reiterate, it is the two types of (sometimes conflicting) behaviours and the two types of measurement responses which demand an explanation. As far as psychology is involved, direct responses are guided by a person's explicit attitudes, and so the search is aimed at finding out what guides the responses on indirect measures and what is implicated in microaggressions. With – to my knowledge – few exceptions, the accepted assumption throughout is that the psychological constituents of implicit bias are a unified homogeneous kind irrespective of what that might be (be it *sui generis* state, or some other mental state which we are familiar with like beliefs or associations).[19] It is a determinately unified homogeneous psychological kind that underpins responses on indirect measures and micro-level behaviours.[20] It is similarly a distinctive (even if related) unified homogeneous psychological kind which grounds explicit attitudes. Indeed, this is the third underlying assumption of the dualistic hypothesis.

> <u>(Commitment 3)</u>
>
> The mental kind which explains overt behaviours and responses on direct measures is distinct from the mental kind that explains microaggressions and responses on indirect measures.
>
> - *Explicit attitudes (such as introspectively accessible beliefs).*
> - *Implicit attitudes/biases (a homogeneous mental kind unlike conscious beliefs).*

[19] Certain philosophical models are exceptions here (e.g. Machery, 2016; Sullivan-Bissett, 2019).

[20] By homogeneous, I mean to describe one set of mental kinds: either associative states, propositional states, patchy endorsements, gut-feelings, imaginings, or any unified type of mental state, but not a conglomeration of these various kinds of states. Mandelbaum (2016) for example, describes this mental kind as propositionally structured unconscious belief, while Gendler (2008a, b) describes it as alief.

At its very core, then, the dualistic hypothesis posits, at the level of the underlying psychology, distinct and determinately unified homogeneous explanatory states that are responsible for results on indirect measurement instruments and for subtle modes of behaviour such as microaggressions.[21]

The dualistic hypothesis also partakes in aligning explicit attitudes, the responses they cause on direct measures, and the action they influence, and in a parallel manner in aligning implicit bias, the responses they generate on indirect measures and the micro-behaviours they guide. Said differently, explicit attitudes, such as beliefs, influence macro-level behaviour and the behavioural responses on direct measurement procedures while implicit attitudes underpin micro level behaviour and the behavioural responses on indirect measures. (See figure 1 in section 2.1, p. 21).

As these commitments have largely gone unquestioned, research has been keen on answering a main question related to the nature of the psychological underpinning(s) involved in implicit bias. If it is taken as a given that the mental underpinnings constituting implicit attitudes are distinct from those constituting explicit attitudes, then the question is how might such a distinction be characterized? Keep in mind that only a few philosophers have questioned whether implicit attitudes should be treated as a homogeneous kind of mental item (for example Holroyd and Sweetman, 2016; Machery, 2016; Sullivan-Bissett, 2019). A few others have investigated whether the mental kind underlying implicit bias is principally distinct from that which underwrites beliefs (e.g. Carruthers, 2018; Stammers, 2016; Brownstein, 2017). However, none have questioned both concerns at once: whether implicit bias is homogeneous and unified *and* whether

[21] Again, with the notable exception of Sullivan-Bissett (2019) who acknowledges heterogeneity in implicit bias but retains the other assumptions of the dualistic alignment hypothesis.

there is a principled distinction between implicit attitudes (or implicit bias) and explicit attitudes. It is these two questions which motivate my thesis.

To sum up this section, the dualistic hypothesis commits to the notion that implicit attitudes form a unified and homogeneous psychological kind that explains micro-behaviours such as microaggressions and responses on indirect measures of attitude. It also commits to positing implicit attitudes as distinct in substantial ways from explicit attitudes (such as beliefs) which guide overt behaviour and responses on direct measures.

2.5 The dualistic alignment hypothesis: Philosophical accounts

This section is devoted to a brief survey of prominent philosophical views on implicit bias in order to show that they all endorse the commitments of the dualistic hypothesis. First, I briefly review some philosophical works which, in one way or another, endorse an associative view of implicit bias, namely Brownstein's (2016; 2017), Holroyd's (2012; 2016), and Saul's (2012). Then I discuss Eric Mandelbaum's (2013; 2014; 2016) doxastic model according to which the psychological reality underpinning implicit bias is unconscious structured beliefs. I continue by explaining Tamar Szabo Gendler's (2008a, b) and Neil Levy's (2015; 2016; 2017) accounts of implicit bias as grounded in *sui generis* mental states. Finally, I overview Ema Sullivan-Bissett's (2019) account of implicit bias as constituted by unconscious imaginings.[22] At this stage, the aim is not to undertake a thorough critique of these accounts; instead, I simply establish how they endorse the dualistic alignment hypothesis.

[22] This is of course not a comprehensive review of the explanatory accounts of the structure of implicit bias. I choose these philosophical accounts either because their work has contributed (and still does) to much discussion in the field and/or because their work attempts novel characterizations of implicit bias.

To the extent that these accounts distinguish between two mental kinds underpinning two distinct modes of behaviour and tracked by different measurement instruments (i.e. to the extent that they involve a dualistic alignment as that indicated by figure 1), I present them as endorsing the dualistic hypothesis. Whether the dualistic hypothesis holds true is one of the guiding questions of this thesis. In later chapters I challenge this principled alignment. The aim of this section, however, is to briefly review how dominant philosophical models belong within the dualistic family of views.

Associative models

Standard philosophical accounts (in line with most cognitive and psychological models – as will become apparent in the next chapter) take implicit bias to be grounded in associative structures which are the result of the learning history of the subject (as discussed in Levy 2015, p. 803). An association is the pairing of a concept (e.g. MUSLIM MAN) and a valence (e.g. negative affect) or two concepts (e.g. MUSLIM MAN and TERRORISM) which occurs frequently enough in the subject's environment that the activation of the concept comes to automatically activate the valence (or the other concept) or make its activation more readily accessible. Associations, or associatively structured attitudes, are thought to differ starkly from propositional attitudes or beliefs. In effect, according to associative accounts, associatively structured implicit biases are (a unified psychological kind) distinct and separate from propositionally structured beliefs or explicit attitudes. While implicit biases affect responses on indirect measures and certain behaviours that may be described as microaggressions, explicit attitudes differ as they guide agential beliefs and overt behaviours. In that regard, associative accounts take on the commitments of dualism at the levels of behaviour, measurement, and psychological underpinnings of behaviour as well as the commitment of alignment in the explanatory roles of the psychological underpinnings. In other words, associative models endorse the dualistic hypothesis.

Michael Brownstein (2017, p. 4) adopts psychology's understanding of attitudes conceptualized as *likings* or *dislikings*, specifically "as associations between a concept and an evaluation". He characterizes implicit attitudes as "preferences that need not enter into focal awareness and are relatively difficult to control" (Brownstein, 2017, p. 8). As these attitudes are *implicit*, it follows that they are to be contrasted with attitudes that are *explicit* – where explicit is described as conscious, agential, and "what people report on questionnaires or other 'direct' measures" (ibid, p. 4). There's a distinction between on the one hand *explicit attitudes*, how they are measured, how they are characterized, and the actions they produce, and on the other hand *implicit bias*es, how they are tracked, how they are characterized, and the behaviour they generate.

Jules Holroyd (2016) similarly adopts psychology's description of attitudes as evaluations. She discusses implicit biases as associations; indeed, she uses bias and association interchangeably. Holroyd (2012, p. 275) describes automatic associations, as "unintentionally" guiding certain "micro-behaviours" whose aspects the subject is unaware of. For instance, she writes,

> An individual harbours an implicit bias against some stigmatized group (G), when she has automatic cognitive or affective associations between (a concept of) G and some negative property (P) or stereotypic trait (T), which are accessible and can be operative in influencing judgement and behaviour without the conscious awareness of the agent (Holroyd, 2012, p. 275).

Along (crudely) similar lines, Jennifer Saul adopts psychology's portrayal of implicit bias as associations. She describes a conception of implicit bias as

> unconscious tendencies to automatically associate concepts with one another…. These are unconscious, automatic tendencies to associate certain traits with members of particular

> social groups, in ways that lead to some very disturbing errors: we tend to judge members of stigmatized groups more negatively, in a whole host of ways (Saul, 2013b, p. 244).

Saul contrasts unconscious implicit bias with "bias as traditionally understood (e.g. the conscious belief that women are bad at philosophy)" (Saul, 2013a, p. 39). She does not explicitly discuss an alignment between implicit bias against women in philosophy and their underrepresentation or an alignment between belief and blatant sexism. Nonetheless, her endorsement of the dualistic hypothesis rests on the notion that unconscious attitudes (or influences) distort our epistemic practices such as judgements and evaluations while we have more control over our conscious attitudes.

Eric Mandelbaum (2013; 2014; 2016): implicit bias as 'unconscious beliefs'

Although the canonical picture of implicit attitudes is that they are grounded in associations, Eric Mandelbaum (2013; 2014; 2016) moves away from the associative picture and proposes a revisionist understanding of implicit bias as grounded in unconscious beliefs. Mandelbaum (2016, p. 649-650) argues that individuals can have, simultaneously, an explicit and an implicit attitude, both of which he describes as grounded in propositionally structured beliefs, yet each residing in "isolated" memory stores. He writes

> …I'm inclined to think that central cognition is fragmented and contains redundant representations. By 'fragmented' I mean that some of our beliefs are causally isolated from other beliefs. (Mandelbaum, 2016, p. 650).

According to Brownstein (2015, section 2.2), Mandelbaum draws upon "theories of the fragmentation of the mind" such as "the Contradictory Belief Hypothesis" to expound his explanation of implicit bias. These theories hold that an agent's implicit and explicit attitudes although contradictory, “both reflect what she believes, and that these different sets of beliefs may be causally responsible for different behaviours in different contexts” (ibid).

Mandelbaum, thus, considers implicit biases not as associations but as

> underwritten in unconscious beliefs …. [that] are honest-to-god propositionally structured mental representations that we bear the belief relation to (Mandelbaum, 2016, p. 635).

He argues that although implicit biases are beyond the reach of consciousness, they are propositionally structured mental representations (or structured beliefs for short). They are also “inferentially promiscuous”, reason-responsive, and validity apt. What this means is that like beliefs, implicit biases have propositional content and this content “can play a role as a premise in valid inferences” and thus can have satisfaction conditions. In contrast, associations whose content is merely relational, enter into “associative chains of thought” and lack the syntactic structure needed for satisfaction conditions (Mandelbaum, 2013, p. 199).

To arrive at his view that the structure of implicit bias is not associative, Mandelbaum suggests that we investigate whether implicit bias allows for inferences or whether it can be modified through extinction or counterconditioning.[23] If it can be modified, then it is associative.[24]

[23] Extinction and counterconditioning are behavioural interventions aimed at breaking certain kinds of undesirable behaviours. They are methods used in behavioural therapy which aim to break the conditioned pairing of a stimulus S with the behavioural response R. For example, to extinguish the conditioned pairing (association) of buying chocolates (R) every time one goes on the train (S), the pairing is broken gradually by disallowing the behaviour (the shop owner doesn’t sell you chocolate) from taking place. In counterconditioning, a behaviour (buying a novel) that is incompatible with the undesirable behaviour (buying chocolate) is induced.

[24] For example, if Tima learns, through repeated pairings, to associate the sound of a school bell with snack time, she will expect to eat every time she hears the bell ring. If every time Tima hears the bell, she is forbidden from eating, i.e. if the association between bell and food is extinguished, eventually, Tima will understand the lack of connection. Similarly, if one substitutes water for food every time the bell rings, then we have applied counterconditioning, and Tima no longer expects food at the sound of the bell (of course now she will expect water). However, being told not to associate the sound of the bell with food or receiving rational arguments about her excessive eating habits just won’t be effective at breaking the association. In fact, no amount of good reasoning to get Tima to stop expecting to eat at the sound of the bell will affect her expectations and perhaps her hunger pangs.

However, if it is amenable to rational argumentation, then something else must be at play. He cites evidence that shows that modification of implicit attitudes can follow a logic that is not associative. In fact, he marshals extensive empirical evidence from studies to support his unconscious belief view (I only briefly discuss his work here but for a detailed account, see his 2016 article). He reviews various research findings from social and cognitive psychology to show that implicit attitudes are sensitive to inferences, responsive to reason, and to the strength of arguments.

For example, he notes the findings of Gawronski, Walther and Blank (2005) which reveal that if subject *S* is conditioned to associate person A with a negative implicit attitude, and if *S* is then told that person *A* hates person *B*, *S* is thereafter more likely to have a positive attitude towards person *B* (as measured by the IAT). This shows that implicit biases operate along the lines of "the enemy of my enemy is my friend", which Mandelbaum argues can't be explained associatively, rather it suggests that some form of inference is at play (Mandelbaum, 2016, pp. 638-639). Similarly, he cites other findings (e.g. Brinol, Petty and McCaslin, 2009) which show that implicit attitudes can be modified more effectively using strong arguments than weak ones. Mandelbaum concludes that if implicit attitudes function as beliefs, i.e. if they are "inferentially promiscuous, interact with motivational states to cause behaviour, and have the capacity to be sensitive to evidential consideration", then they are beliefs, albeit unconscious ones (Mandelbaum, 2016, p. 649).[25]

[25] The argument that implicit biases are inferentially promiscuous is a point of much controversy between Mandelbaum and proponents of associative models of implicit bias. Madva (2016) for example, argues that implicit biases are probably not beliefs since in many cases, empirical evidence suggests that they are insensitive to logical form, and thus flout a minimum requirement for something to be a belief. Similarly, Levy (2015) argues that even if implicit biases are propositionally structured, they are insufficiently responsive to evidence to be described as beliefs. Holroyd (2016) also objects to the model on the grounds that, if Mandelbaum were right, we'd have to give up the idea of belief as

Mandelbaum's doxastic view commits itself to the dualistic hypothesis. First, it understands two modes of actions: responses on direct measures such as self-reports or overt behaviours and responses on indirect measures or micro-behaviours as tracked by different direct and indirect measurement procedures. Moreover, it posits a homogeneous and unified mental kind (*determinately* unconscious belief) responsible for implicit bias, and although structured in the same way as "the conscious judgments that are often also called beliefs" which guide explicit attitudes, the two kinds of beliefs (the conscious belief and the unconscious belief) do "not belong to the same natural kind" (Mandelbaum, 2016, p. 636).

Tamar Szabo Gendler (2008a, b): implicit bias as 'aliefs'

Gendler (2008a, b) posits a *sui generis* kind of mental state she calls 'aliefs' as being responsible for certain actions which subjects might not fully endorse, such as micro-aggressions. She distinguishes our everyday agential *beliefs* from *aliefs*. A belief has a representational component that figures in practical reasoning in virtue of its propositional content and guides action in accordance with desire. An alief, however, is

> a mental state with associatively linked content that is representational, affective, and behavioral (R-A-B) and that is activated – *consciously or unconsciously* – by features of the subject's internal or ambient environment. (Gendler, 2008a, p. 642)

To illustrate, Gendler (2008a) cites Paul Rozin and colleagues' (1986; 1990) poison experiments where participants are shown empty bottles. In one version of the experiment, in plain sight of the participants, the experimenter fills each bottle with sugar and gives the participant two labels saying

delineated along the lines of propositional content. (For detailed arguments against Mandelbaum's doxastic model, see Josepha Torbio, 2018).

either ‘sucrose’ or ‘sodium cyanide’ and asking them to stick the labels on either of the bottles they choose. The contents of each bottle are then filled into two separate glasses of water and dissolved, and the participant is asked to drink from each glass. Participants generally hesitate to drink from the glass filled from the bottle with the ‘sodium cyanide’ sticker, despite the fact that they themselves stuck the label and know it was actually sugar inside. Gendler explains that the participant both believed that the bottles contained sugar and *alieved* that one of the bottles contains sodium cyanide. The R-A-B alief is CYANIDE-DANGER-AVOID.

Gendler (2008a, b) articulates the distinction between belief and alief by identifying a belief with what one would explicitly reply if directly pressed on a question about what one endorses. If one finds evidence that disproves their beliefs, they are likely to alter this belief. Stressing the distinction, she notes that beliefs are sensitive to evidence and are subject to revision based on rational arguments, i.e. they are truth evaluable. In contrast, aliefs by their nature are “insensitive to how appearances misrepresent reality”, they are reality insensitive (Gendler, 2008b, p. 570), or as Brownstein (2017, p. 272) calls them “arational”. Gendler writes,

> If I believe that *P*, and subsequently learn that not-*P*, I will revise my belief… Learning that not-*P* may not cause me to cease alieving that *P*… alief just is not reality-sensitive in the way belief is. Its content does not track (one’s considered impression) of the world (Gendler, 2008a, p. 651).

So, whereas beliefs are “reality-sensitive”, aliefs are not.

Gendler does not give a detailed explanation of how implicit bias is an example of aliefs. Nonetheless, she does offer some useful suggestions on how to consider implicit biases and the actions they generate. She explains that given the nature of beliefs and aliefs, there will be cases when the two are discordant in a subject as when they “activate contrary behavioural repertoires” (Gendler, 2008a, p. 570). Aliefs, for example, bring about implicitly biased behaviour of “an avowed anti-racist [who] exhibits a differential startle response when Caucasian and African faces are flashed before her eyes” (ibid, p. 553).

Gendler’s account of implicit bias as aliefs suggests a commitment to the dualistic hypothesis. Firstly, there’s a commitment to the claim that overt behaviour is caused by beliefs (and desires) while another type of (automatic) behaviour is caused by aliefs. This is just another way of expounding the commitments of the dualistic hypothesis. Gendler’s account supports the understanding that two different mental elements (beliefs and aliefs) underpin two different modes of behaviour.

Neil Levy (2014; 2015; 2016; 2017): implicit bias as ‘patchy endorsements’

Neil Levy (2014; 2015; 2016; 2017) has written considerably about implicit biases, specifically whether they are agential or not and whether we ought to be held morally responsible for them. As I am interested in showing how the dominant accounts of implicit attitudes endorse the dualistic alignment hypothesis, I will focus mainly on this feature of Levy’s account.

Generally, Levy (2015; 2017) (inspired in part by Mandelbaum’s (2014; 2016) propositional understanding of implicit bias) claims that what explains microaggressions are *sui generis* mental states. Levy calls these states “patchy endorsements”. They are *endorsements* because they “have sufficient propositional structure to have truth conditions” but they are nonetheless *patchy* because they feature only some of the inferential relations to the agent’s other mental states and therefore do not always integrate with them (Levy, 2015; 2017, p. 13).[26] Implicit attitudes, then, have patchy propositional structure, but not the *bona fide* propositional structure to allow them to underpin the "continuous, broad and systematic responsiveness" rightly associated with beliefs (Levy, 2016, p. 10). Insofar as implicit attitudes do not interact with other representational states in ways that are sensitive to content, they are a mental kind distinct from explicit attitudes such as beliefs.

What drives micro-behaviours such as microaggressions (at least "partially") is the "unitary phenomena" of implicit attitude, i.e. patchy endorsements (Levy, 2016, p. 539). This is because, Levy claims, implicit bias exhibits a "failure of systematicity" insofar as the agent or the mechanisms underlying the agent's behaviour fail to respond to (and to be sensitive to) reason in the way beliefs do (Levy, 2017, p. 13). Conversely, he treats behaviour that is driven by explicit attitudes alone to have a contrasting character "because these attitudes are not insensitive to the reason-giving force of the relevant considerations" (Ibid).

Levy also argues that explicit attitudes "have contents that are introspectable, and which control personal-level cognition" (Levy, 2017, p. 535). Beliefs and the actions they guide are agential and involve personal-level control. Personal-level control is control that is deliberative and explicitly intentional. Levy explains how this level of "control over the moral character of our actions requires that that moral character features (though not necessarily under that description)

[26] The upshot is that if implicit attitudes are patchy endorsements, they are not integrated into the agent, and the actions they guide are thus not agential (Holroyd, Scaife, and Stafford, 2017). Said differently, since patchy endorsements are not attributable to the agent, then the agent cannot be held responsible for the actions guided by them. The matter of moral responsibility is the focus of much lively debate, but it is not an issue that concerns me at this time (see Brownstein and Saul 2016, Volume 2 for accounts of moral responsibility of implicit bias).

in our explicit intentions" (Levy, 2016, p. 9). In contrast, implicit attitudes and the actions they influence escape personal-level control. They are "not able to be deployed at the personal level; rather, they influence cognition in ways that escape conscious control" (Levy, 2017, p. 535). Explicit attitudes, then, are a unitary phenomenon distinct from implicit attitudes (another unitary and homogeneous phenomenon). The former are responsive to reason, feature personal-level control and conscious deliberation, and are measurable with explicit measures, whereas the latter (implicit attitudes) lack systematic responsiveness to reason, escape personal-level and conscious control, and are measurable with implicit measurement instruments.

Ema Sullivan-Bissett (2019): implicit bias as 'unconscious imaginings'

Finally, Sullivan-Bissett (2019) considers the constituents of implicit bias to be *unconscious imaginings*. According to her "imagination model", the constituents of implicit bias are unconscious imaginings which may be structured associatively or non-associatively. When implicit bias involves associative processing, one of three things take place. For example, as the subject encounters a stimulus (e.g. a bearded brown person), an association between two "unconscious imagistic imaginings" may be activated (e.g. the unconscious imagistic imagining of the concept *bearded brown man* associatively linked to the unconscious imagistic imagining of the concept of *terrorism*. It may also be that an association between an unconscious imagining and a valence is activated (e.g. the unconscious imagining *bearded brown man* and *bad*). Alternatively, an association between two propositional imaginings may be activated (e.g. an unconscious propositional imagining with the content *there is a bearded brown man* associatively linked to an unconscious propositional imagining with the content *there is terrorism* (Sullivan-Bissett, 2019, p. 12).

However, when implicit bias is structured non-associatively, the subject may simply experience a single instance of an unconscious imagistic imagining (e.g. the imagining of *bearded brown terrorist*) with no association involved. Alternatively, the subject may unconsciously propositionally imagine that *bearded brown men are terrorists* (ibid, p. 13).

In any case, the imagination model considers implicit bias and the behaviour it guides (even if not uniformly structured) to be a psychological kind distinct from explicit beliefs and the behaviour they generate. Thus, insofar as her model endorses a duality and an alignment between what indirect and direct measurement devices track, the type of behaviour they influence, and the mental elements explaining them, her view places her within the dualistic alignment camp. Although her understanding of implicit bias as constituting heterogeneously structured unconscious imaginings differentiates her from other dualistic theorists, Sullivan-Bissett remains a dualist and retains her place in the dualist camp of theorists.

2.6 Three illustrative cases

Before I end this chapter, and to make our later discussion more concrete, I present three cases of individuals and their actions based on typical vignettes discussed in the literature of implicit bias. These cases shall prove useful to illustrate the challenges faced by the dualistic view in chapter 5 and will help in demonstrating my alternative framework in chapter 6. The literature presents such cases as readily comprehensible and recognizable and as common examples of real-life scenarios. Considering these cases, however, helps us to see how advocates of the dualistic view present the puzzling story in too simple a manner that fails to account for the complexity of human behaviour and cognition.

Case 1: Rachel the Receptive Racist

This case is largely based on Eric Schwitzgebel’s (2010) case of “Juliet the implicit racist”. It is echoed in Keith Frankish’s (2016) discussion of implicit bias and parallels Jules Holroyd’s (2016) “Biased Egalitarian” test case. Rachel is a Caucasian professor of psychology who, as an essential part of her research, studies racial differences in intelligence. Her research convinces her that racial differences in intelligence are not supported by any reputable scientific evidence. Her egalitarianism is in line with her overall liberal stance on diverse sensitive topics such as race, sex, sexual orientation, etc. She avows explicitly and argues, as does Juliet, “coherently, sincerely, and vehemently” for the equality of intelligence between races (Schwitzgebel, 2010, p. 532). Yet Rachel’s subtle behaviour demonstrates systematic racial bias. At the beginning of each term, she can’t help but think that some of her students have more potential than others, and those students never happen to be black. When a black student makes an insightful remark, or writes an exceptional essay, she feels surprised more so than if a white student does. When she is on a committee to choose student applicants, she is more likely to be convinced of the credentials of white applicants and often – without awareness – requires more evidence than if the applicant is white. As with “Juliet the implicit racist”, Rachel “deliberatively strives to overcome” any potential bias (ibid). At times, she tries to counteract any possible bias, for example, by being especially generous in her evaluation of her black students. Yet, as Schwitzgebel observes, such “patronizing condescension” could itself be an indirect reflection of bias (Schwitzgebel, 2010, p. 532). As Rachel remains constantly alert to any possible discrimination, she realizes that such self-conscious vigilance is impossible to maintain at all time.

Case 2: Sam the Self-Justifying Sexist

This case is an embellishment of Holroyd's (2016) "Protocol Observing Racist". Sam, also a university professor, is strongly motivated as a professor of ethics to adhere to liberal norms of equality of the sexes. At his department, females constitute a larger proportion of the student population. However, as it stands, the highest-ranking professors are males and the junior level professors are females. This gendered 'hierarchical' inequality in ranking has persisted despite Sam's *purported* efforts to encourage gender diversity and ratify the situation. He forms the intention to support the promotion of all professors in his department, but data shows that more men than women faculty get promoted every year. He talks about the lack of diversity in the field and continues to take part in hiring more males than females in his department. He often complains that the lack of competent philosophers and the failure of his department to come in good rankings for publications is largely due to affirmative action policies.

Case 3: Barb the Brute Bigot

This last case is based on Jules Holroyd's (2016) "Racist" test case which she later calls *non-conflict* or *harmony cases*. It tells the story of another university professor, Barb, who holds prejudiced beliefs and judgments towards non-white individuals and is known to share them openly. Barb might be predicted to harbour implicit bias towards persons of colour as measured on Implicit Association Tests. When on a hiring committee, she makes what might be considered racist decisions because they are not informed by the objective facts. As she believes a person's race is relevant to their suitability as an academic, she refuses to accept a person of colour as a colleague. Similarly, she displays micro-behaviours in interaction involving individuals of colour;

she shows discomfort and irritability in interracial interactions, she doesn't engage in discussions with and often sits very far away from black and brown-skinned individuals.

It is interesting to note that test cases such as those reviewed here are often presented in the research in a manner displaying dualistic alignment assumptions. For example, Holroyd explicitly discusses the distinct underlying causes of the behaviours of the individuals. She writes about individuals having explicitly prejudiced or egalitarian beliefs and harbouring "implicitly negative associations", and she attributes actions to these underlying causes (Holroyd, 2016, p. 160). For example, she writes:

> [Individual] B's (the biased egalitarian's) evaluations of the applicants are guided by his explicit intention not to discriminate. However, his behaviour is also influenced by his implicit bias…he displays unintentional 'micro-behaviours' in interracial interactions: he manifests discomfort, reacts with more irritability, sits marginally further away from the black interviewees (Holroyd, 2016, p. 160).

We will see in coming chapters that apart from underlying commitments to dualistic alignment, the dominant accounts present the puzzling (and the non-puzzling) cases of prejudice as though the data shows stark systematicity in individuals' behaviours across time and context. These cases, however, are presented as decontextualized and with little or no reference to situation, context, or other possible factors which affect human behaviour.

2.7 <u>Conclusion</u>

In this chapter, I provided an introduction to the phenomenon of implicit bias and to its psychological explanatory framework. I considered what I called the Dualistic Alignment Hypothesis (dualistic hypothesis for short) as the psychological explanation for this phenomenon; that it is because we harbour implicit biases *distinct* from our explicit beliefs that our social behaviour across a broad range of everyday situations is negatively impacted. I reviewed some empirical research that is often used to support this dualistic explanation. I showed that the dualistic hypothesis commits to a number of dualisms that culminate in a hypothesis of alignment between the attitudes we think of as agential (such as beliefs) and the actions they guide on the one hand, and implicit bias and the actions they govern (such as microaggressions) on the other hand.

Further, I illustrated how several dominant philosophical frameworks endorse the dualistic hypothesis. The orthodox view among philosophers in the field of implicit attitudes defends an account along the lines of the dualistic alignment hypothesis. Orthodox philosophical accounts commit themselves to dualistic alignment assumptions. Namely, they are committed to a distinction in the mental explainers of the responses on direct and indirect measurement instruments and a distinction in the behaviours they guide. The details of what exactly constitutes implicit bias may differ between different philosophical accounts (it may be associations, unconscious beliefs or some *sui generis* mental state like aliefs, patchy endorsements, unconscious imaginings, and so on), but the general framework is the same. Two distinct, uniform, and homogeneous psychological kinds are responsible for responses on direct and indirect measures and for macro-level and micro-level behaviours.

Chapter 3

HOW WE GOT FROM THERE TO HERE

A Historical Narrative of the Hegemony of the Dualistic view

Introduction

This chapter offers a history of the empirical research into implicit bias and the accompanied development of the measurement techniques used to track it. To that end, I describe dominant cognitive models in philosophical psychology and explain how they emerged. I trace the emergence of the notion of implicit bias (as a unified homogeneous mental kind) guiding responses on indirect measures as *distinct* from that which explains other forms of behaviours (such as outright overt discriminatory actions). As I explicate the historical events leading to the state of current orthodoxy, the narrative I thread supports the idea that this orthodox framework was arrived at in no small part because of certain contingent factors, or as what Helen Longino (1990) would call *non-epistemic* factors, influencing scientific decisions.

I begin in 3.1 by elaborating on the description of direct measurement instruments and discussing how they have evolved through the years. I do not assess these devices here as I aim merely to chart a historical path of their development. (A discussion of the appropriateness of attitude measurement instruments will follow in chapter 4). In 3.2, I trace the emergence of the dominant cognitive models of implicit bias to what psychologists Payne and Gawronski (2010, p. 2) consider as “two roots of research”. Michael Brownstein (2017), inspired by Payne and Gawronski (2010), reframes the discussion and calls these two parallel roots the *True Attitude* stream (3.3) and the *Driven Underground* stream (3.4). I discuss each of these in turn and I argue that the Driven Underground stream has hegemonized the research, inspired philosophers to fixate on the explicit-implicit dichotomy and established the psychological underpinnings of implicit bias as unified and distinct phenomena responsible for microaggressions. In 3.5, the final section, I sketch a historical narrative of how, in less than three decades, we got to the situation we are in now. I explain how the research started using implicit measures as *bona fide* instruments of attitude and as a replacement for explicit measures but ended up using both instruments to track two different yet genuine attitudes. I end by advancing some cautionary notes. Firstly, that the development of the research uncritically appropriates theories from cognitive psychology to explain social phenomena. Secondly, that my historical analysis depicts a shift – even a reconstruction – of the meaning and purpose of direct and indirect measurement instruments which in turn translates into a shift in hypothesizing about the psychological mechanisms underpinning behaviour.

3.1 The predicament that started it all: Direct measures of attitude

Before I sketch the historical narrative of the research in the next section, here I delve further into explaining direct measures (often called explicit measures) because, as we shall soon see, they were an essential motivator for the development of the research on implicit bias.

As I explained in section 2.4, measures of explicit attitudes are often called *direct measures* because they only require the participant to overtly (verbally or non-verbally) report their attitude towards a social identity group or a member of that group (henceforth referred to as *target object*). Typically, direct measures instruct respondents to indicate their liking or disliking of the target object on a scale, their degree of agreement or disagreement with a particular statement on a Likert type scale (ranging from *strongly agree* to *strongly disagree*), or their response to a statement on a temperature scale (typically ranging from -5 to 5 for their level of warmth towards the target object). Scales may consist of single-item or multiple-item inventories or questionnaires, and they can be administered and scored easily and quickly (Carifio and Perla, 2007; Schwatrz, 1999). The dominant understanding is that the reported attitude is “an evaluative summary” of information about the target object (Fazio 2007, p. 606). Direct measures have many common underlying assumptions, for example, that the participant is knowledgeable of her attitudes and willing to express them accurately (more on this in the next chapter).

The most common direct measure of attitudes typically involves a self-report of temperature on a 1-10 scale. On some types of Feeling Thermometer scales (e.g. Krysan, 2000), subjects may be asked to rate their feelings towards a particular social group (say Muslims) on a scale from 0 to 100 (0 being *very cold* or *unfavourable* and 100 being *very warm* or *favourable*). A typical question may be ‘How would you rate this group (Muslims)?’. On another type, the Semantic Differential Scale, respondents are asked to rate a social identity group (say blacks or women) on a bipolar scale with contrasting adjectives at each end (Heise, 1970). The scale is composed of increments from -3 to 3 where 0 denotes a neutral attitude. For example, participants may be asked to rate black people on the trait of friendliness on a scale from -3 (*unfriendly*) to 3 (*friendly*).

Likert-type measures are thought to be slightly more sophisticated than simple explicit statements of evaluation or feeling thermometers (Carifio and Perla, 2007). They were developed to assess explicit attitudes by asking respondents to evaluate a range of statements to gauge their beliefs. Respondents are asked to rate how much they agree with the statements on a scale from 1 to 7 (1 being *highly disagree*, and 7 *highly agree*). For example, the Attitude Towards Blacks Scale (ATB), which enjoys some use in current research, features an array of items portraying different aspects of racial bias such as interracial contact, policy issues, and interracial marriage (Brigham, 1993; Payne, Burkley, and Stokes, 2008; Nelson, 2009). The following items are representative of the kind of items included on such measures:

- I worry that in the next few years I may be denied my application for a job or a promotion because of preferential treatment given to minority group members.
- Some blacks are so touchy about race that it is difficult to get along with them.
- I would rather not have blacks live in the same apartment building I live in.
- Racial integration (of schools, businesses, residences, etc.) has benefited both blacks and whites.
- Interracial marriage should be discouraged to avoid the ‘who-am-I’ confusion which the children feel.

Similarly, the Attitudes Towards Women Scale (ATW) (Spence, Helmreich and Stapp 1973) features items about gender roles, authority, and dependency such as:

- Women should worry less about their rights and more about becoming good wives and mothers.
- Economic and social freedom is worth far more to women than acceptance of the ideal of femininity which has been set up by men.
- There should be a strict merit system in job appointment and promotion without regards to sex.
- The intellectual leadership of a community should be largely in the hands of men.
- In general, the father should have more authority than the mother in bringing up the children.

The ATB and ATW have been shown to correlate well with other measures of explicit attitude, and thus to be effective at what they do. However, as we will see in the next chapter, results on direct measures are significantly shaped by many variables, including contextual factors and motivational factors not to appear prejudiced (Schwartz, 1999; Nelson 2009). Although these instruments are said to explain overt prejudice, it is controversial that they do. They suffer from the obvious shortcoming that respondents who want to appear in a positive social light and those who fear negative social judgements for their prejudice tend to distort their responses (Nelson, 2009; Alwin and Krosnick, 1991). In theory, they are thought to do well in explaining and predicting overt, blunt prejudice, but they are ineffective at explaining subtler, yet arguably explicit prejudices.

As alternatives to the more blatant measures of racism and sexism discussed above, newer scales such as the Modern Racism Scale (MRS) and the Modern Sexism Scale (MSS) have been developed (McConahay, Hardee and Batts, 1981; Swim, Aikin and Hall, 1995). They are taken to capture what Todd Nelson calls more "subtle forms" of racist/sexist attitudes (leaving unclear the distinction between explicit forms, more subtle forms, and implicit forms of racism/sexism) (Nelson, 2009, p. 371). These measures are said to target evolved types of racism and sexism insofar as they employ "non-reactive means of tapping more subtle prejudice" (Swim, Aikin and Hall, 1995, p. 104). Rather than items which express 'old-fashioned' racism, MRS and MSS items express more evolved forms of bias that are sometimes believed to be more socially acceptable. The subject is instructed to respond on a Likert type scale how much they agree or disagree with each item. Items on the MRS include the following:

- Discrimination against blacks in no longer a problem in the US.
- Blacks are getting too demanding in their push for equal rights.

While the MSS taps into both levels of hostile and benevolent sexism and includes items like:

- Most women interpret innocent remarks as acts of being sexist.
- Women should be cherished and protected by men.
- Once a woman gets a man to commit to her, she usually tries to put him on a tight leash.[27]

[27] Modern sexism is theorized to be ambivalent in the sense that sexist persons hold beliefs which appear, on the surface, to be positive. This is called *benevolent sexism* and it involves beliefs about the complementarity of gender differences, the "endorsement of paternalistic behavior", and "heterosexual intimacy" (Swim and Hyers, 2009, p. 415). Although these appear to be positive, they carry harmful implications towards women because of the assumptions associated with them. For example, endorsing the complementary nature of gender differences translates into believing

Scales such as the MRS and the MSS are generally considered direct measures of attitudes, but I think there is reason to be sceptical about this categorization. Modern – or evolved – racist/sexist attitudes are associated with less endorsement of egalitarian beliefs and are less likely to be identified as racist/sexist than are the traditional explicit biased beliefs (Swim and Hyers, 2009). Yet there is overlap as well as *uniqueness* regarding modern racist/sexist attitudes and other more 'traditional' forms of race/gender related explicit biases (Nelson, 2009). Methodologically, it isn't clear how this overlap translates; are these scales measuring similar or distinct constructs as other direct measures like self-reports? Granted, trying to understand how to categorize these measures leads to more questions than answers. Specifically, where would modern racism/sexism scales fit within the distinction in the measurement instruments and in the responses generated by the instruments? Are modern racism/sexism scales considered *direct* and measuring explicit attitudes in the sense discussed above? And if not, would they classify as *indirect* and measuring implicit bias? Or would they classify into some third category?

As long as theorists continue to rely on measurement instruments to support their theories, I think it is vital to address these issues. The literature on implicit bias does not concern itself with this worry, and I don't propose to do this here. But I think one way to move forward is to define and agree on what is meant by *explicit* bias, especially given that current models of implicit bias do not clearly identify *exactly* what that is. If there's anything obvious about how prejudice works in modern society, it is the idea that it may manifest itself in subtle ways which may (or may not)

that women are less competent than men in male dominated fields, the endorsement of paternalistic behaviour translates into viewing women as child-like, and the focus on heterosexual intimacy renders the belief that women control men through their sexuality. There is evidence for a positive correlation with these beliefs and hostile sexism (Swim, Aikin and Hall, 1995; Swim and Hyers, 2009).

be tracked directly. This is a worry that will prove problematic in chapter 4, but for now suffice it to note that there are conceptual difficulties surrounding direct measures.

Bracketing these worries for the time being, it remains the case that social psychologists at the turn of the nineties had these limited tools at their disposal for measuring attitudes. But these tools suffered from serious weaknesses including the worry that they were subject to self-presentation motives (more on this in chapter 4). Respondents were likely not to report their attitudes honestly for fear of presenting themselves in a negative light. As we shall see in the next chapter, a fundamental underlying assumption of direct measures is that the subject is honest, self-reflective, and clearly understands what is being asked of them. These are stipulations which clearly the experimenter is not able to guarantee. And even when the newer measures like the ATB and ATW or the MRS and MSS were developed, the worry remained because the issue of reporting attitudes spreads equally to the questions on these measures. As a result, theorists began looking elsewhere for measures that would offer more representative assessment of subjects' attitudes. Methods from other fields of psychology, namely cognitive psychology, offered the potential to assess respondents' veritable social judgements. I examine these methods next.

3.2 <u>Origins of implicit bias research</u>

In this section, I pursue the areas into which psychologists looked for more effective measures of tracking people's attitudes. I discuss "two roots of research" into indirect measures (and implicit bias) as reviewed in Payne and Gawronski (2010, p. 2) and labelled by Brownstein (2017) as the *True Attitude* stream and the *Driven Underground* stream.

Probing the origins of implicit attitude research is an essential step to uncovering any misunderstanding in our conceptions of prejudice and discrimination and how they function. As I shall argue towards the end of this chapter, the appropriation of methods and theories from cognitive psychology may have resulted in the puzzling situation in which philosophers find themselves. Specifically, it may have led to a misleading construal of social judgement as bifurcated, and of social behaviour as ahistorical and decontextualized.

Theorists trace the beginnings of research into implicit social cognition to the early and mid-nineties when social psychologists appropriated ideas and methods from cognitive psychology to address social phenomena (racism, sexism, etc.) related to attitudes (Payne and Gawronski, 2010; Hughes, Barnes-Holmes and De Houwer, 2011). Close investigation into these origins led Payne and Gawronski (2010) to two different historical streams of research (Hahn and Gawronski, 2016; Brownstein, 2017).

The first stream, inspired by cognitive research on selective attention, emphasized the automatic functioning of underlying cognitive mechanisms, where *automaticity* is characterized as *spontaneity, inescapability,* or *uncontrollability* (Fazio et al., 1995). Michael Brownstein (2017) calls this the *True Attitude* stream of research. It posits that, insofar as indirect measures are not susceptible to self-presentation demands as are explicit self-reports, they are genuine measures of a person's attitude about stigmatized social groups. As these measures track automatically activated associations, they reflect one's 'true' attitude towards stigmatized social groups. In most situations, people are able to control the behavioural manifestations of their biases and align their behaviour with prevailing social norms. But in certain situations when their cognitive resources are sparse such as when the subject needs to act spontaneously, the activation of the 'true' attitude is inescapable, and the attitude is tracked by the indirect measure. Whereas responses on direct measures or explicit self-reports may be more amenable to revision and control, indirect measures such as priming tasks (more on these below) constrain participants' ability to control their responses, or so argues the True Attitude stream.

Whereas the first stream of research grew out of research on selective attention, the second stream grew out of research on implicit memory.[28] Specifically, it focused on awareness or introspective accessibility of attitudes, and as such, it gave way to a distinction between *explicit* understood as *conscious*, and *implicit* understood as *unconscious* (Greenwald and Banaji, 1995; Wilson, Lindsey and Schooler, 2000). Brownstein (2017) calls this the *Driven Underground* stream as it has it that people have two dissociated but genuine attitudes towards a target object (e.g. a stigmatized social group): one attitude is explicit and reflects an individual's introspectable beliefs, and the other is implicit and reflects un-introspectable attitudes driven into the unconscious (Payne and Gawronski, 2010; Gawronski and Brannon, 2017; see also Machery, 2016). The broader notion derived from this steam of research is that biased egalitarians persist in holding contradictory attitudes, consciously endorsing a non-prejudiced attitude which guides their beliefs and being unaware of their bias as it influences their micro-behaviours and responses on indirect measures. This line of research has been described by Edouard Machery as "the Freudian picture of attitudes" pointing to its Freudian "pedigree" (Machery, 2016, p. 110). Machery draws a parallel between Freud's conscious and unconscious desires and explicit and implicit attitudes.

3.3 <u>The True Attitude stream of research</u>

In this and the next section (3.4), I trace in more detail the rise of the two streams of research. Here, I first briefly discuss the relevant work of the originators of the True Attitude stream of research. I then elaborate on priming tasks as the indirect measures which contributed to its development and

[28] Implicit memory is thought to be implicated in the positive effects of prior study on later performance, despite the person not being consciously able to remember what she has learned.

popularization (3.3.1). I end by reviewing the Motivation and Opportunity as Determinants of Evaluation (MODE) model as the account that exemplified this stream of research (3.3.2).

The True Attitude stream developed out of cognitive theories of learning and selective attention where the processing of information was divided into two modes: automatic and controlled (Payne and Gawronski, 2010). Automatic information processing was understood as being unlimited in capacity, difficult to voluntarily alter or stop, and as needing little attention. Conversely, controlled information processing (involved in selective attention) was thought to be limited in capacity, voluntarily initiated and changed, and as requiring attention. Well-learned items were retrieved automatically, while poorly learned items required much selective attention and cognitive effort to be retrieved. Much of this research employed the classic priming paradigm concerned with automatic activation processes (more on this shortly). This classic work focused on the speed of responses to whether a string of letters is a word or a non-word and the extent to which the responses are facilitated by the prior presentation of a prime (Fazio et al., 1995). For example, the presentation of the word 'doctor' as a prime was found to facilitate the identification of the word 'nurse' as a word. Such research indicated that concepts which are associated with the prime are *automatically* activated upon the presentation of that prime and hence the responding to semantically related target words is facilitated by the prime (Fazio and Hilden, 2001).

Patricia Divine (1989) and Russell Fazio and his colleagues (1986; 1995; 2000) adopted the priming paradigm. They wanted to show that certain attitudes may also be activated automatically at the presentation of an associated prime. To this end, they conceptualized attitudes as "object-evaluation associations in memory" (Fazio, 2001, p. 116). First they drew a distinction between well-learned strong attitudes understood as automatically activated and poorly learned weak attitudes which required attention and intent.[29] The rationale for this distinction was that strong attitudes will be automatically activated and thus difficult to control whereas weak attitudes will require intention and effort to be activated. Connecting attitude strength with its automatic activation allowed them to use the priming paradigm to track attitudes.

Sequential priming tasks which were newly developed and showed promise in cognitive psychological research on attention were adopted as unobtrusive and indirect measures of attitudes. A central characteristic of priming tasks (and indirect measures generally) is that a participant's automatically activated attitude towards a target object may be *inferred* from certain "objective performance indicators" (Gawronski and Brannon, 2017, p. 1). Performance indicators are described as measures of accuracy or speed of response to the target object, usually represented by a stimulus involving an image or a word (ibid). Faster and more accurate responses reflect uncontrolled or inescapable attitudes (i.e. strong associations/attitudes), and slower and less accurate responses mean the attitude has been mediated and controlled (i.e. weaker associations/attitudes). I describe priming tasks in more detail in what follows.

3.3.1 Priming Tasks

Priming tasks are indirect measures designed to track subjects' automatically activated attitudes without having to ask the subjects to report on them (Fazio et al., 1986; 1995). Fazio and colleagues (1986) developed a priming paradigm to investigate the automatic activation of attitudes upon presentation of the target object. I give a brief discussion of this measuring device below.

[29] A well learned or a strong attitude corresponds to the strength of the association between an object and its evaluation. It measures how easily an evaluation 'comes to mind' when an object is encountered (for example, if the evaluation of 'tasty' is strongly associated with the object 'chocolate', then the evaluation is activated automatically upon encountering the object).

In a standard priming task, a participant is presented with a series of trials, each one consisting of a valenced priming stimulus followed by a target stimulus, and then, depending on the nature of the task, the participant is instructed to classify the target word evaluatively or semantically. The speed and/or accuracy of the response is thought to be facilitated for target words which are conceptually congruent with the associations activated by the priming stimulus. For example, if the nature of the task is evaluative (i.e. measuring affect), the participant is primed with a valenced word (HAPPY, SAD) followed by a target word (CANCER or FLOWER). The relationship between the prime and the target stimulus is manipulated: on some trials they're congruent (HAPPY-FLOWER or SAD-CANCER), on others they're not (HAPPY-CANCER or SAD-FLOWER). The participant is then asked to classify the target words (CANCER or FLOWER) as either *good* or *bad*. Alternatively, if the nature of the task is semantic (i.e. measuring stereotypes), the participant is primed for example with masculine or feminine personality traits (e.g. ARROGANT or CARING) and then asked to classify the target stimulus which are common female and male names (e.g. JANE or JOHN) in terms of their semantic property as either *female* or *male*. Variant tasks with slight modifications in the procedure such as the Affect Misattribution Procedure (AMP) and others have been developed but the general idea behind these is the same (Payne et al., 2005; Gawronski, 2009; Gawronski and De Houwer, 2014; Gawronski, Brannon and Bodenhausen, 2017).[30]

[30] In the AMP, participants are presented with a (positive or negative) prime stimulus followed briefly with a Chinese ideograph. The ideograph is then masked by a black and white pattern and the participants task is to rate whether they liked or disliked the (neutral) Chinese ideograph. Typical findings have shown priming effects in that when primed with a positive stimulus, participants are likely to rate that they liked the neutral ideograph and disliked it when primed with a negative stimulus. The tendency to rate a liking of the neutral ideograph when primed with a white face more so than when primed with a black face is taken to indicate a preference for white over black individuals. This is shown to be the case even when participants are instructed not to let the prime stimulus influence their evaluation and even when they were informed, in details, about how prime stimuli may influence their

Classifications are found to be faster and more accurate for target words which are preceded by a prime with the same valence (e.g. the classification of FLOWER as *good* is faster if it's preceded by the prime word HAPPY) or if the target word is preceded by a prime word which is semantically congruent (e.g. the classification of JANE as *female* is faster if it's preceded by the priming word CARING) (Spruyt, Gast and Moors, 2011; De Houwer, 2014). The idea here is that since affectively or semantically congruent prime-target pairs/associations yield faster and more accurate results than incongruent prime-target pairs/associations, the underlying mechanism must involve automatic activation of the associations (De Houwer, 2014). As Spruyt and colleagues (2011, p. 4) conclude, the underpinnings are due to "subliminal stimulus processing".[31]

Priming tasks were developed further and used by social psychologists Divine (1989) and Fazio and colleagues (1995) to explore the phenomenon of aversive racism. Divine (1989, p. 5) argued that automatic activation of cultural stereotypes (for example, the association of WOMEN and FEEBLE) was universal to everyone in society, but she distinguished between "knowledge" of a cultural stereotype and its "endorsement". Implicated in her model is the idea that while both high and low prejudiced individuals are subject to automatic activation of cultural stereotypes with

evaluation (Payne et al. 2005). The assumption here is that participants misattribute feelings evoked by the prime stimulus to the presentation of the neutral Chinese ideograph. The presentation of the prime gives rise to a certain 'feeling' which the participant mistakenly believes to be a feeling elicited by the ideograph itself. The participant then judges the ideograph in line with the valence of the prime. A challenge for AMP presented by Bar Anan and Nosek (2012) is the idea that participants may be basing their responses on intentional use of the prime stimuli in evaluation of the neutral ideographs, a challenge that potentially undermines the implicit nature of the task, but it isn't clear whether the correlation between AMP scores and self-reports of intentional use of the prime reflects causal effects of intentional processes during the evaluation of the ideograph or retrospective confabulations of intentionality (i.e. whether participants really used the affect elicited to evaluate the ideograph or they confabulated that the reason for their responses was because they had such an intention) (Payne et al., 2013).

[31] Subliminal priming in social psychology research means that the prime stimulus is presented for such a short time so as to be below the threshold of awareness; that is, the participant cannot detect it. Subliminal priming is contrasted with supraliminal priming where the prime stimulus is in plain view but in very subtle way; it is detected at the conscious level.

the relevant prime, those with high prejudice have the stereotypes and their personal beliefs overlapping, while those with low prejudice have decided the stereotype is inappropriate, and so engage in an effortful and controlled inhibition of the activated stereotype and activation of their personal egalitarian beliefs. Hence, for Divine, the distinction between automatic and controlled processing relates to a distinction between knowledge of cultural stereotypes and endorsement of those stereotypes (i.e. through personal beliefs), and the variability in priming scores reflects a measurement error (Payne and Gawronski, 2010).

Unlike Divine, Fazio and colleagues (1995) argued that variability in scores reflect meaningful individual differences in the strength of the attitude or the object-evaluation association. Their reasoning is that subjects who have weak or neutral racial attitudes are not affected by primes on subsequent tasks, while participants who have strong positive or strong negative racial attitudes show corresponding priming effects. Fazio and his colleagues (1995, p. 1014) reported that priming tasks are a "bona-fide pipeline" into participants' attitudes, meaning that such tasks overcome the worry of self-presentation inherent in direct measures and tap the subjects' strong attitudes.[32] The distinction between automatic processing (as uncontrolled or spontaneous) and controlled processing thus reflects a distinction between the subject's automatically activated 'true' attitudes and their contaminated or 'bogus' self-reports which are moderated by image management and social desirability. The broader notion is that priming tasks

[32] In a later article, Fazio and Olson (2003) point to a confusion which had arisen from the (1995) reference to the priming techniques as "bona fide pipeline" into one's attitudes. They clarify that "'bona fide' is simply intended to indicate that any automatic attitude activation occurs farther upstream than the overt response to an explicit measure" (p. 304). These researchers stress the notion that responses on explicit measures are expressions of an activated attitude which can be affected and moderated by motivation and opportunity as well as whatever the activated association happens to be. In that sense, the activated attitude (what the implicit measure tracks) happens upstream while the overtly expressed judgement (what the explicit measure tracks) is the activated attitude which may have been influenced and moderated by motivational factors (see the section on the MODE model for more details).

offer a window into people’s true attitudes without the effects of other factors such as impression management, social desirability, or cognitive depletion. Mitchell, Nosek and Banaji (2003, p. 468) claim that this approach was “offered up as a chameleon’s mirror for social cognition” because to use indirect measures is to see past obscuring influences of social desirability to the true underlying attitude.

Resonating with such theories, experiments such as the *bogus pipeline* revealed that participants hold more prejudiced attitudes than they were willing to report.[33] As a result, racial attitude research was to gain most from priming tasks because studies of racial attitudes had been particularly challenging to researchers. Priming tasks, along with their variants, became the empirical driving force behind the True Attitude stream of research which Fazio and his colleagues describe as providing “a bona fide pipeline for attitude measurement” (Fazio, 1995, p. 1025).

Crucial for our purposes is to note that prior to Fazio and Hilden (2001), the term ‘implicit’ was never once used by either Divine (1989) or Fazio and his colleagues (1995) to denote the automatically activated psychological elements. Moreover, the notion of a unified psychological phenomenon – distinct from the subject’s attitude towards a given target – as underlying the responses on priming tasks was also not considered. What was at issue was only the participant’s *attitude* indicated by the strength of the object-evaluation association. Throughout the early development of the True Attitude stream of research nowhere was the term ‘implicit’ used to denote an automatically activated attitude (or subject-evaluation association) nor was the term ‘explicit’ used to describe self-reports. It wasn’t until at least six years later (in Fazio, 2001; Fazio and Olson,

[33] The bogus pipeline is a technique used by social psychologists to reduce false responses on self-reports. It is like a fake lie-detector or polygraph used to get subjects to respond truthfully to sensitive questions.

2003; 2014; Olson, Kendrick and Fazio, 2009) that the term ‘implicit’ gained ground. And even so, it was only the measures which were referred to as either implicit or explicit (not the attitude). For example, indirect measures of attitude such as the *Implicit* Association Test (IAT) started to be referred to as *implicit measures* and direct measures of attitude such as self-reports as *explicit measures*. The reason for this, as Fazio and Olson explain is that:

> …it is more appropriate to view the *measure* as implicit or explicit, *not* the attitude (or whatever other construct). What makes priming or the IAT implicit is that these techniques provide estimates of individuals’ attitudes without our having to directly ask them for such information (Fazio and Olson, 2003, p. 303).

In line with their interpretation of implicit and explicit as denoting the characteristics of the measurement instruments, Fazio and colleagues (1995) developed the Motivation and Opportunity as Determinants of Evaluation (MODE) cognitive model which has become influential in attitude research, and which I discuss next.

3.3.2 The MODE model (Fazio et al., 1995)

Fazio (1990) and Fazio and his colleagues (1995) developed the Motivation and Opportunity as Determinants of Evaluation or the MODE model as a theory of social attitudes. At its core is the assumption that attitudes can be activated automatically/spontaneously. The MODE model considers the human mind to hold a unified psychological state towards a target object, but two classes of attitude-to-behaviour processes: spontaneous and deliberative (Fazio, 1990; Fazio et al., 1995; Fazio and Olson, 2014). When attitudes are activated automatically, they may influence behaviour in the absence of a goal to evaluate the target object, this is known as the *spontaneous* process. A priming task, for example, involves spontaneous processing such as when a prime (e.g. the face of a black man) automatically activates the association (BLACKS and a negative valence BAD). This automatically activated association (BLACKS and BAD) influences the response on the priming task, and so the subject evaluates the target picture negatively (see section 3.3.1 on priming tasks). However, when individuals have the motivation and the opportunity to deliberate on the information, as in the case with direct measures like questionnaires, the influence of the automatically activated association can be overcome and its impact on behaviour may be reduced. This is known as the *deliberative* process.

The main prediction of the MODE model rests on the idea that automatically activated associations influence behaviour depending on two factors: (1) the subject's cognitive resources and (2) the subject's motivation and opportunity to use these resources. On direct measures, mental processing and strategic control are permitted with ample time for deliberation. This allows the automatically activated associations to be mitigated. It also gives sufficient motivation and opportunity to enhance self-presentation and diminish any harm by moderating one's immediate impression of the target object. The deliberation reduces the influence of the automatically activated associations. In such cases, the agent's behaviour (her self-report) will be controlled and perhaps censored. In contrast, when the motivation and the opportunity to engage in mental processing is low and when strategic control is limited as is the case with procedures on indirect (or implicit) measures such as priming tasks, then the response reflects the subject's automatically activated associations (Fazio et al., 1995; Fazio and Olson, 2014; Hughes, Barnes-Holmes and De Houwer, 2011).

To explain the difference between responses on direct/explicit and indirect/implicit measures, the MODE model introduces contextual and motivational variables that moderate how the activated associations are expressed. Said differently, the difference in the results of these measures reflects a difference in the cognitive processing. On direct/explicit measures, cognitive processing is possible and so the responses are moderated by contextual factors, the motivation to reflect and control, as well as ample resources of time to produce the self-report. On indirect/implicit measures, cognitive processing is limited due to time and cognitive resource inhibition and thus the response reflects the subject's unmoderated 'automatically activated' associations (Fazio, 1990; 2000; 2001; 2007; Fazio and Olson, 2003; 2007; 2014).

Supporters of the MODE and the True Attitude stream in general make no 'mentalistic' commitment to a distinction in the psychological underpinnings of the two types of responses on direct and indirect measures. Their focus is rather on the distinction between automatic and controlled mental processing *imposed* by the measurement procedures. For theorists following this stream, the primary motivation behind developing indirect procedures to assess any hidden bias (especially challenging ones involving race and gender) was to overcome the limitations of direct procedures such as self-reports. They did not consider the responses of these direct procedures as an indication for a distinct underlying psychological kind. Further, these psychologists considered that the automatically activated attitudes (i.e. the object-evaluation associations) are implicated in responses on indirect as well as on direct measures. Effectively, they did not claim a distinction between an associative mental state underpinning the response on indirect measures and a distinct mental state underlying the response on direct measures. Measures of attitude were taken to track a complex "summary evaluation" of the target object, and depending on the conditions imposed by the instrument, the expression of the mental state was either automatic or moderated (Fazio, 2000, p. 2). It was the instruments, in fact, that may be referred to as either ex*plicit* or *implicit* (Fazio, 2000; Fazio and Olson 2003; 2014; Olson, Kendrick and Fazio, 2009).

Although the MODE model has received good empirical support, it has also been recently criticized with reference to the literature on control of implicit bias which is both extensive and complex (refer to Brownstein, 2017 for a brief review). Empirical research has shown that implicit biases are not automatic in the sense of being inescapable akin to reflexive (recall the discussion in the previous chapter where I argued that microaggressions require some decision making and the processes involved are much more complex than knee-jerk type actions). I will have more to say about controllability of attitude in chapter 5. Next, I discuss the second stream of research.

3.4 The Driven Underground stream of research

While the True Attitude stream has its followers, it is the second stream, the Driven Underground stream, championed by Greenwald and Banaji (1995), that comes to have a firm hold of the literature. This may be due to the popularity of the Implicit Association Test which developed out of this stream of research. The Driven Underground stream draws a distinction in the underlying psychological underpinnings of behaviour to explain the dualism in the responses on the different measurement instruments, thereby bolstering the dualistic alignment hypothesis. In what follows, I discuss the emergence and the rise of this stream.

At around the same time as Fazio and his colleagues (1995) were publishing their findings on the effects of priming tasks on the automatic activation of attitudes, Anthony Greenwald and Mahzarin Banaji (1995) were publishing their own research findings on attitudes which drew on theories of implicit memory (Gawronski and Bodenhausen, 2011; Hahn and Gawronski, 2016). Implicit memory is described as the influence of past experience (e.g. previous exposure to a stimulus) on later performance (e.g. on a memory task related to the stimulus) in the absence of any conscious recollection of that past experience (Payne and Gawronski, 2010). Memory for recent events may be expressed *explicitly* as a conscious recollection of the events (e.g. after reading a novel, the subject explicitly recounts the plot) and it may be expressed *implicitly* as the facilitation of memory test performance without the conscious recollection of the event (e.g. responding to a word completion task) (Greenwald and Banaji, 2017, p. 861).[34]

Greenwald and Banaji (1995) modelled their understanding of implicit bias along the same lines as cognitive psychology's understanding of implicit memory. Just as implicit *memory* had been described as consequences of past experience on later performance in the absence of conscious awareness of that experience, so was implicit *attitude* characterized as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (Greenwald and Banaji, 1995, p. 8). Brownstein (2017) calls this line of research the *Driven Underground* stream, the idea being that in our modern world, people hold non-prejudiced attitudes towards stigmatized groups yet they persist in holding – simultaneously – past prejudices which have been driven into the unconscious.

Greenwald and Banaji's (1995) paper was the first published paper to employ the terms *implicit*, *explicit*, *conscious*, and *unconscious* to describe attitudes, and it set the stage for the subsequent dichotomization of our understanding of social attitudes. Indeed, this paper may have been seminal in establishing the now dominant explicit-implicit distinction in the literature and in

[34] In a recent article, Greenwald and Banaji (2017) clarified that explicit-implicit distinction is equated with direct-indirect distinction but not with conscious-unconscious distinction. They stressed that the use of direct measures doesn't imply the study of conscious phenomena, so it doesn't exclude unconscious phenomena, nor does the use of indirect measures imply the study of unconscious phenomena, because it could include conscious phenomena. Meanwhile, they acknowledge that they "find themselves occasionally lapsing to use implicit and explicit as if they had conceptual meaning (Greenwald and Banaji, 2017, p. 862). I would add that much of the research on implicit bias lapses into this distinction, and as we shall see in chapter 5, this is unfortunate because it suggests that the use of the different instruments is informative about the mental elements driving the responses (and the behaviour).

the identification of *explicit* with *conscious* and *implicit* with *unconscious*. In the first footnote of this now famous paper, Greenwald and Banaji write:

> The terms *implicit-explicit* capture a set of overlapping distinctions that are sometimes labelled as *unaware-aware*, *unconscious-conscious*, *intuitive-analytic*, *indirect-direct*, *procedural-declarative*, and *automatic-controlled*. These dichotomies vary in the amount and nature of implied theoretical interpretation. This article uses the *implicit-explicit* pair because of that dichotomy's prominence in recent memory research, coupled with the present intention to connect research on attitudes, self-esteem, and stereotypes to memory research (Greenwald and Banaji, 1995, p. 4).

These authors' use of the terms *explicit* and *implicit* to describe attitudes. Their presentation of the explicit-implicit distinction as parallel and overlapping with other distinctions that characterize memory research may be linked to their intention to model a theory of implicit *bias* along the same lines as theories of implicit *memory*.

As certain dichotomies such as implicit-explicit cognition, indirect-direct measures, reflexive-reflective, unconscious-conscious were useful in memory research, so they became uncritically adopted and interchangeably employed in social attitude research. Responses on indirect measures became characterized as implicit, unconscious, introspectively unavailable, reflexive, and automatic while responses on direct measures were described as explicit, conscious, introspectively available, reflective, and controlled without supporting evidence for such usage (Payne and Gawronski, 2010). As we will see in section 3.5, this will come to highly resonate with the duality of dual-process or dual-system theories such that the properties describing one side of the dual-systems will be later taken to feature in the description of implicit attitudes and the properties describing the other side will be taken to feature in the description of explicit attitudes.

Greenwald and Banaji's (1995) use of the terms *implicit* and *explicit* to denote attitudes greatly facilitated terminological slippage as theorists in the field later adopted the terms to describe psychological elements underpinning social behaviour. Although the original justification for indirect measures was motivated by the concern to minimize self-presentation motives, and although this need meant that indirect measures were theoretically essential in studying social cognition, the justification for their usage somehow became (for Driven Underground theorists) about capturing a distinct psychological *element* that lies beyond introspective awareness and that guides certain subtle kinds of behaviour.

Wilson, Lindsey and Schooler's (2000) dual-attitude model, for example, may be characterized as squarely within the Driven Underground stream. According to this model, aversive racists have "dual attitudes which are different evaluations" of the same race (Wilson, Lindsey and Schooler, 2000, p. 102).[35] Implicit attitudes have their roots in childhood experiences and operate in an unconscious mode. Thus, a response on an indirect measure is taken as the automatic, "habitual", and unconscious negative evaluation of the target object. Conversely, explicit attitudes are "more recently constructed positive evaluations" of that target object adopted by the individual later in life and expressed in conscious mode (ibid, pp. 102-104).[36] To illustrate, consider a man who was raised in a sexist family environment and learned to be prejudiced against women. As an adult, this man learns and adopts egalitarian ideals and so rejects sexism. His recently constructed egalitarian attitudes do not eradicate the earlier (childhood) sexist ones but only replace them as

[35] The reader would be right to draw a parallel with Eric Mandelbaum's implicit bias as unconscious structured beliefs differing from explicit attitudes as beliefs (refer to section 2.5).

[36] Psychologists use the term 'psychological construct' instead of 'psychological item' to describe the mental underpinnings of the responses on the measurement tasks.

the two operate alongside one another, each guiding a different mode of behaviour. Note that the response on indirect measures is taken as proxy for the mental item *implicit attitude* while the response on an explicit measure is used as a proxy for *explicit attitudes*. This man would, according to Wilson and his colleagues (2000) score as sexist on indirect measures such as priming tasks and Implicit Association Test IAT.

3.4.1 The Implicit Association Test (IAT)

In support of the theories of social cognition, Greenwald and Banaji (1995) developed The Implicit Association Test, IAT, as an indirect measure of attitudes. The IAT has since become the most well-known of implicit attitude measurement procedures. In this sub-section, I elaborate on the IAT and argue that it is the driving force behind the Driven Underground stream of research.

The IAT consists of binary categorization tasks which are arranged in such a way as to be either compatible or incompatible with common stereotypes. In a standard IAT, for example, the race IAT, a participant is requested to respond to 5 categorization blocks or tasks. In block 1, she is presented with a picture of either a black or a white face at the centre of a screen and instructed to respond (as quickly and as accurately as possible) by pressing a key at either side of the screen that corresponds with words related to a social group (AFRICAN AMERICAN/EUROPEAN AMERICAN). In block 2, she is presented with valenced words at the centre of the screen (AGONY, SWEET) and again instructed to categorize the word as either *pleasant* or *unpleasant* by pressing a left or a right key. Block 3 reflects prejudice congruent pairings. In this block, the participant is presented with either a word (CRASH/SINCERE) or a picture of a white or black face at the centre of the screen and instructed to hit a left key if the word/picture is categorized as *African American* OR *unpleasant* and a right key if the word/picture is *European American* OR *pleasant*. In block 4 the orientation is switched so now the participant is instructed to press a left key to categorize white faces as *European American* and a right key to categorize black faces as *African Americans*. In the final block, the pairings are switched to reflect prejudice incongruent categories (if *African American* OR *pleasant* then hit the right key, if *European American* OR *unpleasant* then it's the left key). Blocks 1, 2, and 4 are considered practice blocks while pairing blocks 3 and 5 are considered critical blocks (Greenwald and Banaji, 1995; Greenwald, McGhee and Schwartz 1998).[37]

The idea underlying the IAT is that responses are facilitated when the pairings (i.e. associated words) are arranged in such a way as to be compatible with one's preferential evaluations (black/unpleasant or white/pleasant). Most subjects are found to be faster and to make fewer mistakes on prejudice congruent vs prejudice incongruent trials (Nosek, Greenwald and Banaji, 2007).

The dichotomies discussed in Greenwald and Banaji's (1995) article (namely the explicit-implicit and the conscious-unconscious dichotomies) were reflected in the theorizing about the IAT. This instrument was presented by its developers as a measurement tool that captures *implicit biases* described as the psychological underpinnings of the measured response. In large part as a result of the IAT's ease of administration as compared to other indirect measures and its "popularity", researchers - without critical analysis - "simply adopted the 'implicit attitude' versus 'explicit attitude' duality in their work" (Gawronski, *personal correspondence*).

[37] To rule out any results being due to order effects in widely collected data, the order of blocks 3 and 5 is different across participants.

However, in their overview of the research, Fazio and Olson (2003) express misgivings about the usage of the terms *implicit* and *explicit* (imported from cognitive psychology and used to refer to attitudes) in large part because such a usage smuggles within it conscious/unconscious connotations. Referring to the response outcomes of indirect measures as *implicit attitudes* brings with it the unwarranted conceptualizing of these responses as unconscious. This in turn exerts, according to Fazio and Olson (2003, p. 302), “subtle, and not so subtle, influences” on the way we theorize about the nature of these attitudes and about the behaviour in which they are implicated.

Certainly, using this kind of terminology in attitude research suggests that implicit attitudes are ones of which individuals lack introspective awareness. But this, according to Fazio and Olson (2003), is far from the truth. There is nothing about the current measurement instruments aiming to track implicit attitudes, such as priming tasks and the IAT, that warrants such a suggestion. These authors dispute the Driven Underground stream’s usage of the explicit-implicit distinction because, they add, it “implies pre-existing dual attitudes” (Fazio and Olson, 2003, pp. 302-303). In other words, the use of the terms suggests (without empirical endorsement) that two distinct attitudes exert their influence on behaviour, when in fact there is no evidence from current measurement techniques that “guarantees” the independent existence of explicit and implicit attitudes in memory (ibid).

The Driven Underground stream of research has long dominated theory in areas of philosophy interested in attitudes, and in many ways it still does. Its influence has greatly surpassed any other theoretical paradigm. One can see this clearly in the work of the prominent philosophers reviewed at the end of the last chapter, the accounts of which describe implicit attitudes as unconscious (see section 2.6). As Gawronski and his colleagues (2017) explain, it is one of the "most common interpretations" of implicit bias. Indeed, lay understandings of the notion of implicit bias mainly rely on the conscious-unconscious duality (Kirwan Institute, 2015; Devlin, 2018). The most commonly given piece of empirical evidence referenced to support the Driven Underground interpretation is that, since indirect measures show a very low correlation with direct measures, indirect measures must be tracking some unconscious mental element. This line of reasoning continues with the idea that what makes it impossible for people to report their attitudes on direct measures is that they lack introspective access to them. But a discordance in the results of direct measures and those of indirect measures should not, by itself, be taken as evidence that what is being measured by indirect measures is unconscious. (We shall soon see, in chapters 4 and 5, that there are some other plausible reasons for why people don't report their attitudes on self-report measures and none of these other reasons have to do with introspective access).

To be fair, it may be likely that persons harbour associations of which they are unaware, and which differ from (or even conflict with) their conscious evaluations. However, to make any claim about a systematicity in relation between consciousness and measurement instruments relies on empirical studies which have shown reason for scepticism about such a claim (Fazio and Olson 2003; Gawronski, Brannon and Bodenhausen, 2017). The available evidence suggests that people may be aware of what it is that indirect instruments are trying to capture, and still they respond in a different (or opposite) manner on direct self-reports for a variety of reasons (Hahn et al., 2014; also see Gawronski, Hofmann and Wilbur, 2006 for a review).

3.5 How we got from there to here

I am now in a better position to sketch the outlines of a historical narrative of the research literature on social attitudes: a narrative that describes how we got from there to here. The claim that I advance is that the dualistic view has come to hegemonize the literature’s understanding of attitudes in no small part because of contingencies. I draw a path which links three key stages leading from the original conception of indirect measures as true measures of attitude to the dualistic alignment hypothesis which currently dominates the theoretical thinking for philosophers interested in bias.

My account begins with the period when direct measures, instead of being used as collaborative techniques became used as separate measurement procedures independent of indirect measures. The story continues with the second phase signalling Greenwald and Banaji’s (and the many theorists who came after them) usage of the explicit-implicit dichotomy as parallel to the conscious-unconscious dichotomy. The narrative culminates in the literature’s dominant adoption of dual-system theories as the theoretical framework with which to understand social attitudes. My argument suggests that the appropriation of the explicit-implicit, conscious-unconscious terminology that started with the Driven Underground stream framed how we subsequently came to think about bias.

I begin with the first phase. The advent of indirect measurement techniques was hailed as revolutionary for cognitive psychology research for the main reason that these measures provided a technology that tracks cognitive states without the conscious mediation of the subject. With instruments such as priming tasks and IAT, social psychologists for the first time held the capacity to observe the deep layers of psychological reality where psychological elements are said to be *automatically* activated without the need to depend on the subject’s self-reports. Indirect measures, thus, were developed with the idea that they were procedures to *bypass* subjects’ social desirability motives, self-enhancement, limits of introspection, and the other limitations of direct measures. These newly developed measures were characterized, not as *supplementary* to the inadequate direct measures, but as a completely *distinct* category of measures. As it was, direct measures which were rendered somewhat ineffective by the development of indirect measures, maintained their foothold in the research and were established, not as supplemental to indirect measures, but as measures of distinct cognitions. This set a significant turning point for the research as it served to change the face of theorizing about attitudes. Essentially, indirect measures lost their status as collaborative of direct measures and took on an independent standing as measures which track independent mental elements.

It is questionable how, somewhere along the way, research into bias which started out searching for alternate methods to overcome the limitations of direct measures morphed into affirming these measures and continued to employ them to track distinct mental elements. Direct measures, as already mentioned, were considered plagued with limitations especially when it comes to topics related to social judgements such as racism, sexism, homophobia, and the like; they were criticized for resulting in bogus or not genuine responses. It is similarly questionable how both types of measures (the problematic direct method and the alternative indirect method) became crucial for tracking *two* distinct psychological categories responsible for the dissonance in the responses on these methods. This treatment, as I see it, dismisses the suspicion in the respondent's endorsements of explicit attitudes, dismisses the problems of introspection, and legitimizes the effectiveness of problematic measurement instruments (i.e. direct measures) which had given rise to this suspicion in the first place.[38]

[38] The suspicion in direct measures arose as a result of the serious problems of introspection and the powerful effects of conformity to norms, social desirability, and self-enhancement motives, and the complex ways in which sexism and racism operate in modern society.

Using direct and indirect measures to track distinctive mental elements paved the way for the second phase in the narrative; this came with the widespread acceptance and use of the explicit-implicit dichotomy and its correspondent conscious-unconscious dichotomy. That was set into place with Greenwald and Banaji's (1995) article on implicit bias. It may be that the use of direct and indirect instruments to capture distinct mental elements (i.e. the first phase) coincided with and developed in support of the use of the explicit-implicit dichotomy. Conversely, it may be that the use of the explicit-implicit dichotomy supported the employment of both measures as capturing distinct kinds of attitudes. A detailed sequence of events is inessential. What is crucial, however, is the treatment of the results of the two measurement techniques as indications of distinct psychological elements and as firm evidence for an explicit attitude-implicit attitude dichotomy. One can see the self-supporting cycle.

Certainly, the research within the Driven Underground stream *required* that direct measures be independent of indirect measures and capture distinct elements. For the explicit-implicit dichotomy to reflect a conscious-unconscious dichotomy, the instruments used needed to be shown as tracking either the *explicit-conscious* mental elements or the *implicit-unconscious* elements, thus solidifying the dichotomy. That was the move taken by the Driven Underground researchers, but it was an unwarranted move because it was tangential to the available evidence. To see why, consider Fazio and Olson's argument which I quote at length:

> …. Participants may be unaware that their attitudes are being assessed, but that does not mean they are unaware that they possess those attitudes. We would encourage researchers not to equate an implicitly measured construct with an unconscious one. Although an implicit-explicit dissociation may occur because the implicit measure reflects associations to which the individual lacks introspective access, such a dissociation also may occur because people are reluctant to admit (on the explicit measure) to the tendency that is revealed by the implicit measure. This ambiguity alone is reason to be wary of the

> connotations that the term “implicit” carries regarding unawareness. Logically, reference to an implicit attitude…should require evidence of unawareness and not solely the use of an implicit measurement technique (Fazio and Olson, 2003, p. 303).

Fazio and his colleagues repeatedly warned against using the term *implicit* to denote an attitude or the underpinnings of behaviour, because using this term to denote an attitude (as opposed to a measurement procedure) carries with it the problematic notion of *unawareness*.

Let me briefly recap the first two phases of the narrative. Fazio’s (1995) discussion originally focused on the conditions under which the responses on the different measures were produced. Implicit measures were called *implicit* because they involved the restriction of time and cognitive resources, and the responses on these measures were considered automatic (i.e. spontaneous). Explicit measures were called *explicit* because they allowed for ample time and cognitive resources, and the results were considered non-automatic. This was a significant contrast to what Greenwald and Banaji (1995) argued for. These authors’ focus was on the introspective accessibility of the underlying causes of the responses on the two measures; what was captured by indirect measures they considered as *implicit* because it was introspectively inaccessible while what was tracked by direct measures they described as *explicit* because the subject has introspective access to it.

Although the True Attitude and the Driven Underground streams were two very different ways of thinking about social cognitions and about the processes involved in social evaluations, the literature was radically drawn to the Driven Underground stream way of theorizing. For curious reasons that have little epistemic import, the work of Fazio and his colleagues from the True Attitude stream received far less attention and was far less engaged with than the work of Greenwald and Banaji and their colleagues from the Driven Underground stream. An illustration of this disparity comes from the vast difference in citations of these theorists' seminal articles. Whereas Fazio's most cited article (his 1995 publication) received a total of 1308 citations on google scholar citations search, Greenwald and Banaji's (1995) article shows to be dramatically more influential with 6856 citations.[39] It remains unclear why the research was drawn to this kind of interpretation of the empirical results. But it certainly was not because the interpretation had more evidence in its favour.

In the decades since the True Attitude and Driven Underground streams took hold, research using indirect measures has undergone an explosion in the development of indirect instruments inspired by priming tasks and the IAT. The increase in research interest and the development of newer indirect measurement tools helped expand the field in various ways, none of which, however, involved the re-examination of the field's core assumptions about the use of the explicit-implicit dichotomy to refer to the attitudes that are being measured. The data was vastly growing, yet theory struggled behind.

According to Payne and Gawronski (2010, p. 6), a "theoretical framework…that could specify how the two kinds of representations may differentially influence judgements and behaviour" was missing. (Note how the language used here was already being hijacked by the notion of dualities in mental states). It may have seemed to researchers psychologically more reassuring to consider the sorts of microaggressions people engage in as unconsciously driven. Irrespective of the reasons, the explicit-implicit as conscious-unconscious dichotomy was more

[39] Interestingly, this was their second most cited publication, while Greenwald and colleagues (1998) 'Measuring individual differences in implicit cognition: The Implicit Association Test' unsurprisingly surpassed it with a total of 11 249 citations on google scholar. True, the higher citation need not indicate positive engagement, but in this case, it does. These results were taken from a google search on September 19, 2019.

appealing to theorists at the turn of the century. What they needed was a theoretical framework to interpret this explicit-implicit as conscious-unconscious dichotomy.

The needed framework came with the advancements of dual-process modelling theories. Research on dual-process theories is extensive and highly complex, but for my purposes I take it to link a broad range of cognitive features along two systems underlying intuitive and reflective thought processing. The main assumption behind these theories is that it takes multiple cognitive dualities to be systematically correlated to form two distinct clusters and thereby two distinct functional systems (Payne and Gawronski, 2010; Gawronski, Sherman and Trope, 2014). Although there are minor differences between different dual-process theories regarding the core psychological attributes belonging within each clustered system, the similarities are greater than the differences, so I consider dual-process theories somewhat crudely. Kahneman's (2003) dual-system theory is a good example. According to Kahneman, *System 1* is "*associative, automatic, unconscious, slow-learning, effortless, experiential, affective, parallel, and holistic",* and *System 2* is "*rule-based, non-automatic, conscious, fast-learning, effortful, rational, cognitive, sequential,* and *analytic"* (cited in Gawronski, Sherman and Trope, 2014, p. 8).

Payne and Gawronski (2010, p. 9) describe dual-process theories, which originated in the psychological research of the 1970 and 1980s, as "domain specific" in that such theories were concerned with "particular phenomena in social psychology" like "persuasion", "impression formation", and "dispositional attribution". At the turn of the millennium, however, all that changed. According to these authors, an influential review article by Smith and DeCoster (2000) reconceptualized dual-process theorizing "in terms of a *general* set of processes underlying a *variety* of phenomena" (Payne and Gawronski, 2010, p. 9, *italics mine*). Perhaps social cognition could also be a phenomenon explained by these theories. And thus, the theory that needed to frame the methodology of attitude research and the data that the methods produced was realized.

In the field of social cognition, dual-process theories became of fundamental importance for explanatory accounts of the discrepancies between social behaviour and reported attitudes (Chaiken and Trope, 1999; Hahn and Gawronski, 2016; Gawronski and Bodenhausen, 2006; 2007; Hu, Gawronski and Balas, 2017; Carruthers, 2018; Fazio and Olson, 2014; Frankish, 2016; Levy, 2016). Simply put, dual-process theories came to explain the relation between attitude and behaviour by postulating two distinct psychological pathways through which ‘the mind’ translates attitude to behaviour – so to speak.[40]

[40] For example, the well-establish dual process theory of attitude, the Associative-Propositional Evaluative model (APE), conceives of associative processes as providing the basis for implicit bias while propositional processing implicates explicit evaluations or beliefs (Gawronski and Bodenhausen 2006; 2007; 2011; 2014; Gawronski, Brannon and Bodenhausen, 2017; Gawronski et al., 2005; Gawronski et al., 2014; Gawronski and Sritharan, 2010). According to APE, information is stored in associative networks which “provide the basis for propositions” about objects (Gawronski and Bodenhausen, 2011, p. 63). As one encounters a relevant stimulus (e.g. a black job applicant), associative processing is engaged, and an association with a certain “gut feeling” is automatically activated (e.g. the association of BLACKS and INTELLECTULAL INFERIORITY) (Gawronski and Bodenhausen, 2006, p. 693). If encountering the stimulus elicits negative features, then a gut feeling with similar negative valence is activated. This “gut-feeling [is] translated into a corresponding propositional evaluation:

(1) blacks are intellectually inferior to whites

resulting in an implicit evaluation which may manifest in an act of implicit bias such as a biased response towards black people on indirect measures. The essential feature of implicit bias is that they are independent of truth-value assignment, i.e., they are activated whether the subject personally endorses or considers them accurate or not. But the encounter with the stimulus (a black job candidate) also activates beliefs considered to be relevant for the evaluative judgment such as

(2) My research shows that blacks can be equally intelligent as whites, and
(3) I trust my research to be accurate (Gawronski and Bodenhausen, 2011).

Propositional processing is engaged in validating whether the activated association translated into the propositional evaluation ‘blacks are intellectually inferior to whites’ is consistent with the other activated beliefs or not. To the extent that the content of the proposition ‘blacks are intellectually inferior to whites’ – elicited by the negative valenced gut reaction – is inconsistent with the other propositions in the network, the inconsistency must be resolved by rejecting any one of the activated propositions including the gut reaction. Evidently, these three propositions are inconsistent with each other and so cannot all be endorsed at once without violating the norms of cognitive consistency. Therefore, one of them must be rejected by the propositional process. If either (2) or (3) are rejected, then the propositional process can change the activated associations into the evaluation:

Based on dual-process modelling, psychologists adopted dual attitude models to frame the questionable dichotomization of the mental underpinnings of behaviour (e.g. Wilson, Lindsey and Schooler, 2000; Bosson, Swann and Pennebaker, 2000; Rudman et al., 1999; Greenwald and Nosek, 2008; Gawronski and Bodenhausen, 2011). Such a move irrevocably divided the realm of social attitudes into the conscious, controlled, and rational versus the unconscious, uncontrolled, and irrational. Many philosophers likewise adopted this dualistic line of thinking about attitudes (see for example Gendler, 2008a, b; Frankish, 2016; Brownstein, 2017; Levy, 2015; 2016; Madva, 2016). Dual-process theorizing gave legitimacy to the explicit attitude-implicit attitude distinction by framing the problematic explicit-implicit dichotomy. It was this shift that marked the final phase of the narrative where the dualistic alignment hypothesis took hold of the literature on attitudes.

The reader, at this point, might argue that there is nothing illegitimate about the third step where dual-process theories were adopted to interpret the data. Dual-process theories are well-established, and they fit the evidence from the research. I fully agree. There is nothing unjustified about the third phase. However, I urge the reader to re-examine, closely, the second phase in the narrative because it is foundational for adopting the third phase. Using, without evidence, the explicit-implicit dichotomy to reflect a conscious-unconscious dichotomy in the mental

(~1) or ‘blacks are *not* intellectually inferior’

and endorse it in her explicit evaluation. In this case, she explicitly avows her non-racist views while she implicitly retains her negative gut-feeling or implicit bias. Importantly, “reversing the subjective truth of the value of (1) does not necessarily deactivate the associations that gave rise to the affective gut response” and so its rejection leads to a dissociation between explicit and implicit evaluations” (Gawronski and Bodenhausen, 2011, p. 64). Thus, propositional processing involves syllogistic reasoning and inferences derived from any relevant propositional information. The important feature for propositional processing is that reasoning and inference assess the validity of the propositions.

underpinnings of behaviour is an unjustified step and it clearly makes any subsequent steps suspect and unstable.

3.6 Conclusion

In this chapter, I built a historical narrative of the state of the research while offering an interpretation of where things might have gone astray. My story supports the idea that the dualistic alignment hypothesis was arrived at by contingencies. Such contingencies I suggested had little to do with the actual data and more to do with factors dictated by the theoretical environment of the research.

In effect, the dualistic hypothesis ratified the assumption of dualism in the measurement instruments: one to measure explicit attitudes and the other to measure implicit bias, and it ratified the dual corresponding explanatory mental underpinnings: explicit attitude and implicit attitude. As we have just seen, the dualism in measures was grounded in a dubious distinction between two methodological categories, one category was meant to supplement the other disputed category but instead it ratified it as distinct. Still, if we are to assume a justifiable distinction between two *bona fide* methodological procedures, a direct methodological procedure and an indirect one, it remains unclear whether each of these measures is considered effective at capturing a unified category of mental underpinnings as the dualistic view would have us believe.

As I see it, an essential question that any researcher in the field of social attitudes must ask herself is, to what extend do measurement devices (on which the whole enterprise of implicit bias research stands and falls) measure what they set out to measure? As we will see in the next chapter, there are unresolved methodological issues related to measurement procedures. My aim for the next chapter will be to show that the research on implicit bias depends fundamentally on the psychometric properties of the measurement instruments, and if these devices are shown to have problems, serious doubt is placed on the theories which animate the whole project.

Chapter 4

PROBING THE EVIDENCE

Alternative Explanations of the Puzzling Data

Introduction

in this chapter my aim is to raise scepticism around the methodological basis of the dualistic hypothesis. To this end, I discuss various methodological shortcomings that afflict the research on social attitudes. Sections 4.1 and 4.2 present two alternative explanations for the divergence in the measurement responses, none of which involve a difference in kind in the underlying psychological underpinnings. The first of these (4.1) takes it that the divergence may be due to the lack of *structural fit* between direct and indirect instruments. Independent of the differences in the underlying cognitive process, if the measurement tools are dissimilar in structure, then the results they generate will mistakenly suggest a dissonance. The dissonance might be reflective of the different structures of the measurement devices rather than of a dissociation in the mental underpinnings purportedly tracked by these devices. In 4.2, I expand on the possibility that the divergence in the results of the instruments might be due to error related to methodological drawbacks of the instruments. The strength of measurement procedures hinges on the strength of their psychometric properties (i.e. how faithfully they track what they are intended to track). However, if measures are faulty and produce wrong results, then any conclusion drawn from the research is not as meaningful, and more crucially, any puzzling dissonance in the results may best be read as a result of measurement error. My focus is to call into question the divergence in the responses on measurement instruments *interpreted* as evidence for distinct underlying explicit and implicit mental elements. Section 4.3 challenges the predictive relationship between implicit attitude and micro-aggressive behaviour. Here, I examine closely the psychometric issue related to what psychologists call *predictive validity*. I demonstrate that current research methodologies of implicit bias are weak predictors of behaviour, i.e. they show low predictive validity. I conclude that any hypothesis drawn from the empirical research ought to be very cautious when making claims about dualistic psychological explanations of behaviour.

4.1 Alternative explanation 1: Lack of structural fit

I begin with the first alternative explanation. I clarify what is meant by the *lack of structural fit* and illustrate this with an example from research on memory. I end the section by discussing how the lack of structural fit between attitude measurement devices gives reason to be sceptical about the orthodox interpretations of the responses on those devices.

Given that empirical work shows direct and indirect measures of attitude to produce divergent responses, the question is why do these responses diverge? The orthodox explanation, as I have previously noted (2.5), is that the two kinds of measures reflect two independent mental elements (an explicit and an implicit). The assumption is that the distinction between 'explicitness' and 'implicitness' underlies the divergence in the two responses. The True Attitude stream understands direct/explicit measures as tracking *deliberately* edited responses and indirect/implicit measures as tracking *automatic* responses. The Driven Underground stream understands direct/explicit measures as tapping *conscious* mental elements and indirect/implicit measures as tapping *unconscious* ones. While it may seem obvious that the main difference between explicit and implicit measures is that one is *explicit* (it involves conscious and controlled processes) and the other is *implicit* (it involves unconscious and automatic processes), it helps to change perspectives and question whether these tests differ in other ways.

Payne, Burkley and Stokes (2008) suggest one way in which the two measurement instruments differ beyond explicitness and implicitness, and this involves what they call *lack of structural fit*. By *structure* of an instrument, these authors mean “the parts” that make it up and how these parts “work together to measure an attitude” (Payne, Burkley and Stokes, 2008, p. 17). By *structural fit* they mean the degree of “methodological similarity” or the extent of commonality or difference between the key parts of the measurement instruments (Payne, Burkley and Stokes, 2008, p. 18; Cameron, Brown-Iannuzzi and Payne, 2012; Hall and Payne, 2010).

Thus, assessing the structural fit between measurement devices involves assessing the parts that make up their structure. This is important for interpreting any relationship (or lack of) between the results that are obtained by these devices. The more similar the tests are with regards to their parts/structures (i.e. the more the structural fit between the devices), the less the divergence between the obtained results. The less the structural fit between the devices, the more the divergence in the results (Payne, Burkley and Stokes, 2008). Keep in mind that the relationship between the results/responses on direct and indirect measurement devices is highly relevant for dualistic interpretations of attitudes (where a divergence in the results is taken to reflect independent underlying mental elements).

Payne and his colleagues (2008) (citing Jacoby and Dallas, 1981; Tulving, Schacter and Starck, 1982; Warrington and Weiskrantz, 1974) provide a helpful example from memory research to illustrate what they mean by lack of structural fit. They consider early implicit memory research which compared tasks measuring explicit and implicit memory. Explicit memory is typically measured using *recall* and *recognition* tasks (think of school history exams: *recall* tasks are open-ended questions while *recognition* tasks are multiple choice questions). Implicit memory, however, is captured using tasks such as word completion, word identification, and lexical decisions (deciding whether a string of letters is a word or not). With explicit memory tasks, the subject explicitly intends to recall, while with implicit memory tasks, recall happens automatically without intention. Differing results on these measures are taken to show enough difference to establish two forms of memory (explicit and implicit).

Yet these tasks differ in ways beyond explicit (intention to recall) and implicit (automatic/unintentional recall) forms of memory. They also differ in structure (e.g. recall tasks have a different structure from lexical decision tasks), and thus some extraneous irrelevant variables may influence one task and not the other. For example, the frequency of encounter with the target word influences word identification tasks (implicit memory) but not recall tasks (explicit memory) (Schacter, Bowers and Booker, 1989; Payne, Burkley and Stokes, 2008). The concern is that it might be the influence of these extraneous variables which causes the divergence in the results and not a difference in the to-be-measured variable (explicit and implicit memory). When the structures of memory tasks are set up in more similar ways, the results show more similarities (Schater, Bowers and Booker, 1989). When extraneous variables are not controlled for, the obtained results become questionable. It becomes unclear whether these results reflect explicit and implicit forms of memory or other (irrelevant) differences in the operations that each task requires. To ensure that what is being measured are the intended variables (explicit and implicit memory), other variables need to be held constant (Payne, Burkley and Stokes, 2008, p. 18).[41]

[41] To illustrate the point further, but in slightly crude terms, consider the following case of lack of structural fit, or what Ajzen and Fishbein (1977) call *conceptual correspondence*. Post-secondary school performance may be assessed using the following measures: SAT tests, school tests, and organizational skills tests. The first two methods are similarly structured, SAT tests and school tests both work on similar levels of specificity, and thus show good structural fit and a good level of conceptual correspondence. Any dissociation in their results may indicate a

Thus, from a methodological perspective, it is important that all possible confounding factors are held constant and the to-be-measured variable (i.e. explicit or implicit memory) is isolated, otherwise, the results will suffer from error. Turning to attitude research, a lack of structural fit between direct and indirect measurement instruments poses a risk to a dualistic interpretation. As Payne and colleagues (2008) argue, independent of any difference in the underlying causal mechanisms (explicit and implicit attitudes), when direct and indirect measurement devices have differing structures (when they lack structural fit), the responses they generate will typically diverge. Contrarily, when there is good structural fit between the measures, i.e. when they are equated on their structural features, their results tend to converge (Hall and Payne, 2010).

Consider, first, the structure of direct measures. On the Attitudes Towards Women Scale (ATW) subjects read statements like 'The intellectual leadership of a community should be largely in the hands of men' and rate them on a Likert-type scale (Spence, Helmreich and Stapp, 1973). On a sematic differential scale subjects are asked to rate a social group (e.g. women) on certain traits (e.g. responsible/irresponsible, nurturing/neglectful, and faithful/promiscuous), and on a feeling thermometer scale, they are asked to rate their feelings towards women from *very cold* to *very warm*. Direct measures, then, require the subject to read a verbal statement, retrieve their relevant attitude (or formulate one on the spot), and decide how to respond on a numerical scale.

Now consider the structure of indirect measures which, although differs from one measure to the other, there are clear commonalities. Complex statements (typical of direct measures) are in

dissociation in what is being tracked. However, SAT tests and organizational skills tests have little in common in terms of their conceptual correspondence. Therefore, any divergence in their results may be more reflective of a lack of structural fit, in this case conceptual correspondence, than a genuine difference in what is being measured.

indirect measures replaced with images or simple words. The IAT, for instance, instructs subjects to categorize the presented stimuli (namely images or single words) using single word categories mapped onto two response buttons. Subjects may be asked to categorize a photo (e.g. of a black or white person's face) or a word (e.g. AGONY or SWEET) as either one of four categories PLEASANT/UNPLEASANT or AFRICAN AMERICAN/EUROPEAN AMERICAN mapped onto left and right keys (see section 3.4.1). On a priming task, the stimuli (primes) are also images or words and they are presented for a fraction of a second before the presentation of the target image or word. Subjects are then instructed to categorize the target item as either a *word* or *non-word* or to evaluate it as *good* or *bad*. Similar structures are implicated in other indirect measures (see De Houwer, 2003; Payne, Burkley and Stokes, 2008). Indirect measures, then, require the subject to simply categorize the stimulus or evaluate it. Additionally, the metric scale used, as Payne, Burkely and Stokes (2008, p. 17) report, is response-latency for the IAT (meaning the time it takes between stimulus presentation and response) and accuracy of categorization (e.g. of word or non-word) for the priming tasks.

Thus, we can note some key parts to the structure of the direct and indirect measurement instruments which are different. These include:

a. The "stimuli presented" (e.g. complex statements for direct measures and simple words or images for indirect measures).
b. The “metric” scale (e.g. numerical scale for direct tasks and response latencies for indirect ones).
c. The “abstractness” of the response on the measures (“broad social opinions” for direct measures and simple categorization for indirect measures) (Payne, Burkley and Stokes, 2008, p.17).

Noteworthy is that none of these key parts of the structure has anything to do with earmarks of implicitness or explicitness such as automaticity, control, consciousness, and so on. These parts of the structure of the instruments are rather "incidental properties [of these instruments] that are confounded with the implicit-explicit distinction as it has been instantiated in the popular methods" (Payne, Burkley and Stokes, 2008, p. 17). Certain extraneous variables may influence one task but not the other (e.g. a bias for visual cues may influence the rating of a subject's feelings towards a photo on an indirect measure but not influence her responses towards social policies on a direct measure). Hence, as confounds, such extraneous variables bring systematic error to the results. Given the structural differences between the two types of measurement instruments, it becomes unclear whether a divergence in the results reflects a difference in explicitness and implicitness or a difference in the parts that make up each instrument procedure.

To sum up, the dualistic alignment interpretation of the divergence in the results on explicit and implicit measurement instruments is not the only interpretation. That there are two independent underlying psychological mechanisms (explicit and implicit attitudes) is only one way to interpret the results, but in order to be sure of such interpretation, it is crucial to have robust methodologies. When explicit and implicit measures of attitude have differing parts/structures, the responses they produce ought to be compared with caution. Any difference might be due to variables beyond the explicitness/implicitness involved in the measurement procedures. When tasks have more parts/structures in common, they are said to be more structurally fit, and when tasks are more structurally fit, their results tend to converge. In cases where there is a lack of structural fit, the results diverge, and conclusions drawn from this divergence become subject to scrutiny.

4.2 Alternative explanation 2: Faulty measures

Here I discuss the second alternative explanation for the divergence in the results of the instruments measuring attitudes. It is related to weaknesses in the psychometric properties of these instruments.

If the measurement tools used in the research are flawed, then the results they produce will be flawed. The divergence in the responses on these tools may not reflect a difference in what is being measured, rather it may be due to the faulty instruments used (Blanton and Jaccard, 2006; 2017; Mitchell 2017). First, in 4.2.1, I explain what I mean by psychometric properties of measurement instruments, namely what is known as *reliability* and *validity.* Then in 4.2.2 and 4.2.3, I describe the psychometric properties of direct and indirect measurement instruments respectively. I discuss how weaknesses in these properties affect measures of attitude, making the responses they generate questionable.

A caveat before I continue. The purpose of reviewing the problems of validity and reliability inherent in direct and indirect measurement instruments is not intended to discredit the well-established empirical research for sub-personal influences on behaviour. Instead, my discussion is meant to challenge the *interpretation* of the results of the measurement devices on which the literature of implicit bias rests.[42] Specifically, it is meant to call into question whether these measurement instruments can be thought of as having empirically established a homogeneous domain of ‘implicit’ mental states which influence a homogeneous domain of behaviours.

4.2.1 Psychometric properties

The concepts of reliability and validity differ in empirical psychology from those in philosophy. In psychology, *reliability* is a psychometric property of measurement instruments that refers to the extent to which a measurement device yields consistent results across multiple trials. In other

[42] Keep in mind that raw data such as speed and accuracy of categorizing pairs of words is *interpreted* by the researcher. For example, the measurement response on an IAT which is ‘incorrect’ (meaning for example when a black face is inaccurately categorized as BLACK/BAD instead of BLACK/GOOD), this response is further interpreted as symbolizing ‘I dislike’ black persons. The measurement response is thus derived from the meaning assigned to the accuracy and latency of the responses, not from the accuracy and latency themselves (for more on this critique see De Houwer and Moors, 2010).

words, it is the extent to which an individual’s responses correlate for different attempts at measuring the same thing (independent of individual change). For example, consider a quantity of interest, your height in metres, as being measured by a tape measure. If upon repeated measurements of your height, you obtain consistent results (under the assumption that neither the quantity being measured, nor the measurement scale has changed), then the measure/scale (the tape measure) is said to be reliable. Social psychologists seek estimates or degrees of reliability as it is not an all or none concept.

Validity, another property of measurement instruments, refers to the extent to which an instrument measures the theoretical measure/construct of interest. For example, the Standard Assessment Test (SAT) subject exam is said to be a measure of high-school students’ knowledge and understanding of a given subject (e.g. history, literature, science, and so on). The SAT has good validity if, in fact, it is a good measure of the students’ knowledge and understanding of that subject.

Measuring reliability is often more straightforward than measuring validity. Whereas reliability is tracked using repeated measures and assessing their convergence, validity is ideally measured based on how much the measurement result corresponds with the actual attribute/item/measure of interest (all the while taking the reliability of the measurement tool into account). This is referred to as *construct validity*. In all cases, construct validity hinges on a proper definition of the ‘measure’ of interest; the better defined it is, the better the likelihood that construct validity may be tracked. When the measure of interest is well defined (e.g. the distance in centimetres between two points), the construct validity of the relevant measurement tool (e.g. of a measuring stick) is better assessed. Assessing construct validity also often depends on assessing *convergent* and *discriminant* validity. *Convergent* validity refers to how closely a given test instrument is related to other measures of the same construct. If the results of a given instrument converge with previously established measures of the same construct, then the given measure has good convergent validity. For example, when a self-report of the importance of physical exercise correlates positively with a performance-based measure of physical exercise, this shows good convergent validity for the self-report measure. Not only should the given measure correlate with similar instruments, it should also *not* correlate with dissimilar and unrelated ones. This is known as *discriminant* validity. A self-report of the importance of physical exercise should not correlate with a report on calorie intake, this shows it to have good discriminant validity.[43]

Another important form of validity, considered especially for its predictive utility, is *predictive validity* and it refers to the extent to which scores on a given instrument predict behaviour. For example, the predictive validity of a test designed to show how well a student will perform at university will be high if, in fact, students who score high on the test do well at university, and students who score low on the test do badly. In the field of psychometrics, a reliable measure is necessary but not sufficient for a valid measure. A measurement tool, for example, can be reliable but not valid. A digital thermometer that is off by 2 degrees because of some internal fault (indicating low validity) may still give the same result on every measurement trial (indicating high reliability).

Issues of reliability and validity raise considerable worry about the resulting data on the one hand, and about the interpretation of this data on the other. They compel us to consider the

[43] I will not consider these two types of validity in detail here. For a detailed discussion of the various measures of validity of the IAT, however, refer to Schimmack (2019).

following: if measurement devices do not meet the psychometric criteria, why should we trust the data they generate? Problems of reliability and validity ought to lead theorists to be at the very least sceptical about taking measurement responses at face value. If the instruments turn out to show low reliability and low validity, then the results they produce are questionable, and a divergence in the results of the two instruments may be due to error rather than to a difference in what is being measured. Next, I focus on issues of validity and reliability of direct (4.2.2) and indirect (4.2.3) measures in general, after which in section 4.3, I focus specifically on the predictive validity of measures of attitude.

4.2.2 Psychometric properties of direct measures

I begin with direct measurement instruments. In the previous chapter (section 3.1), I explained how direct instruments capture responses that are expressed as overt self-reports and responses on feeling thermometers or Likert type scales. As Banaji, Nosek and Greenwald (2004, p. 280) recognize, there has been a historical reliance on self-reports and other direct measures "more from convenience and a lack of alternative measures". With the advent of alternative indirect methods, there was a clear shift in focus towards indirect measures of bias and their psychometric properties. Direct measures were thus given far less critical attention. And although direct measures suffered from significant limitations, they were not taken to compliment indirect measures. Instead, as I argued in section 3.5, they were used as indicators of a completely distinct mental phenomenon. Thus, to the extent that research conceived the measurement tools as tracking distinct mental elements, it becomes, as Greenwald and Banaji note, "increasingly clear that self-report measures need similar validation" as indirect measures (Greenwald and Banaji, 2017, p. 868). Indeed, there is much to critique about the workings of direct measurement instruments (Nelson 2009; Simeoni

2005; also see Greenwald and Banaji, 2017 for a review). That is what I aim to do in the following: I start off with discussing the reliability and then the validity of direct measurement instruments.

1) Reliability

Direct measurement methods such as surveys and questionnaires rely on several factors which greatly affect their reliability. These include the social situation in which the measure is administered, the design of the measure (wording and context), and the nature of the topic being assessed (Nelson, 2009). In other words, participants' responses on such measures are affected by the social situation in which they find themselves, but also by how the questionnaires are worded, and how sensitive and relevant the topic is to the respondent (Schwartz, 1999).

Features of the measurement tools, then, shape the answers provided by the respondents and influence the conclusions of the researchers. Schwartz (1999) claims that "questionnaires are a source of information that respondents draw on in order to determine their task and to arrive at a useful and informative answer" (p. 94). In other words, the way the questions on these direct measurement procedures are asked may determine the answers given by the respondents. Minor changes in question format or wording may result in dramatic changes in the answers. For example, when asked to consider the most important thing one can teach a child, 61.5% of respondents answered 'to think for themselves' when this was given as an item on the list of things to choose from. But only 4.6% volunteered this answer when it was an open question (Schuman and Presser, 1981 cited in Schwartz, 1999).

Consider Alwin and Krosnick's (1991) suggestion that several measurement errors may result from the way questionnaires/surveys are structured and implemented and from expectations from respondents that may not be met. Factors which may lead to higher measurement error and lower reliability include:

- When the questions being asked are uncomfortable or cause discomfort for the subject – as sometimes they might be (especially for sensitive topics related to sex, race, religion, etc).
- When the questions are posed in an ambiguous way – such that the respondent may misunderstand them.
- When the categories provided by the questionnaires don't allow the respondent to communicate her attitude properly (i.e. by limiting the categories of response on a quantitative measure).
- When the respondent does not have a ready attitude available to report.
- When the respondent is not motivated to report her attitude accurately because of the nature of the questions being asked – the respondent is socially inhibited from expressing her attitude (perhaps to appear progressive and socially desirable).

Insofar as these factors influence the results of the measurement tools, they jeopardize the results on repeated measures, i.e. the reliability of the measures. Direct measures, as such, suffer from shortcomings in their reliability making them subject to measurement error.

2) Validity

As for the validity of direct measures, or the extent to which they track what they are intended to track, this shows to be a rather precarious psychometric property. By definition, direct measures require the respondent to report their attitude explicitly, and by implication they assume a *willingness*, an *awareness,* and an *ability* on the part of the respondents to accurately report their attitudes (Schwartz, 1999). Such measures expect that the respondent is:

(a) Honest

  - She is willing to share her attitude, i.e. she is not concerned to present herself in a particular light, but rather she is motivated to offer her honest accurate response.

(b) Self-reflective

- She has an already formed attitude or at least can form one on the spot.
- She is aware of, has access to, and can accurately report this attitude.

(c) Clear

- She understands the scope of the question being asked and responds in a clear and relevant manner.

In explicit attitude research, it can't be substantiated that expectations related to honesty, willingness to be sincere, and to self-reflect are met. Inaccurate responses can lead to low validity and by implication to faulty results in at least three ways: (a') when the respondent lies about her explicit attitude; (b') when the respondent is unreflective, enhances her self-evaluation, or for various reasons does not reflect on her attitude, and (c') when the respondent is uncertain how to respond because she has mixed cognitions/emotions about the attitude object (Schwartz, 1999).

Respondents often have motivations not to appear prejudiced for personal, self-derived reasons, and concerns to conform to anti-prejudiced social norms. Not surprisingly, individuals who are motivated by these concerns tend to report (on direct measures) being less prejudiced than they actually are (Paulhus and Reid, 1991). Moreover, respondents may be unclear on what they mean when they report on direct measures being 'egalitarian', 'non-egalitarian', 'respecting women', or 'thinking women's natural place is in the home'. It is not clear what exactly such responses indicate, nor is it clear that a respondent can judge herself objectively as egalitarian, even if we considered that there is a common and nuanced understanding of the concept of egalitarianism (I'm not sure that there is).

Explicit attitude research relies heavily on the above three assumptions, and this significantly affects the validity (as well as the reliability) of explicit attitude measures. In fact, it suggests that no explicit measure is free from the influence of a subject's manipulation. The validity of explicit attitude measures is thus always in question. One cannot rule out the possibility that a negative bias is being concealed by a verbally expressed positive attitude (Hall and Payne, 2010; Brunel, Tietje and Greenwald, 2004).

In addition to the unaddressed assumptions discussed above, direct measures, as Yao and Reis-Dennis (no date, p. 10) put it, are generally "crude" and "unsophisticated". On the Harvard Project Implicit website, for example, testing explicit attitude/bias (to be used for comparison with implicit attitude/bias) involves rather rudimentary methods of assessment (Project Implicit, 2019). On the gender-career version of the IAT measure of explicit attitudes, participants are instructed to answer the question "How strongly do you associate family with males and females?" and on a Likert type scale the participant is asked to choose from a set of answers ranging from "strongly/moderately/slightly male to neither male nor female to slightly/moderately/strongly female". Similarly, on the race version of the measure, the participants are presented with a feeling thermometer scale to measure their explicit attitudes. They are asked "How warm or cold do you feel toward Black people?" and "How warm or cold do you feel toward White people?" and instructed to choose a response from a list ranging from "strongly/moderately/slightly prefer White people to Black people" or "strongly/moderately/slightly prefer Black to White people" (Project Implicit, 2019). This is the extent of the measure of explicit bias for Harvard's Project Implicit. The responses on these direct measures are then taken at face value as reflective of the agent's explicit attitude and compared with the responses found on the IAT, often to show a divergence in the results.

So far, my review is meant to raise scepticism around the psychometric properties of the direct measurement tools which the research employs (often uncritically), but as importantly, it is meant to put pressure on the hypotheses drawn from the results of these tools. As I have demonstrated so far in this sub-section, direct measures suffer from issues of reliability and validity which make the results they generate suspicious. I next turn to discussing the reliability and validity of indirect measures.

4.2.3 Psychometric properties of indirect measures

While direct measures are thought to track people's explicitly expressed attitudes including *explicit* biases, indirect measures capture responses that are thought to be reflective of the respondent's automatically activated, or *implicit*, attitudes (Payne and Gawronski, 2010). As discussed in the last chapter (3.3.1 and 3.4.1) indirect measures involve computerized instruments which researchers use to infer a respondent's underlying implicit attitude based on her response on certain tasks. The subject's performance (or response) on the tasks (which involve psychological and/or physical responses) is taken as reflective of underlying object-evaluative associative (or non-associative) psychological elements responsible for the response. A central characteristic of these measures is that the subject's implicit attitude is *inferred* from her performance on the task (e.g. her speed and accuracy in responding to an attitudinal object or to a stimulus). Because I examined priming tasks and the IAT in detail in the previous chapter, and because they are two of the most used indirect measures, I am mostly concerned with their methodological limitations which I review in what follows.

1) Reliability and validity of priming tasks

I recall that *reliability* in psychometrics is the degree to which the task produces stable and repeatable results. In the case of priming tasks, repeated tests don't give systematically similar results. The reliability of priming tasks is reported to be very low (Gawronski and De Houwer, 2014; De Houwer et al., 2009; Soruyt et al., 2011). For example, Spruyt and colleagues (2011) report an "unsatisfactorily low" reliability measure for priming tasks. They cite findings of Bosson, Swann, and Pennebaker (2000) who estimate low to moderate test retest reliabilities for priming measures of self-esteem. Similarly, they cite work by Banse (2001) and Stolz, Besner, and Carr (2005, pp. 26-27) who report very low reliability estimates for measures of automatically activated attitudes towards liked and disliked objects. De Houwer and his colleagues (2009) suggest that the low reliability of priming tasks may be due to the fact that the prime stimulus is not made explicit for the participant. Because of that, the experimenter has "no control over whether or how the subject processes and categorizes the prime" (De Houwer, 2009, p. 359). It may turn out that the participant processes the prime during some measurements but not others. Thus, it becomes unclear whether the prime has had any effects on the participants' responses; it may, or may not, have been registered in the respondent's system. This lack of control over the variables in the study, as De Houwer and his colleagues (2009) suggest, may result in a large amount of error variance and low reliability scores.

As for validity of priming tasks, the reasoning behind priming effects is that exposure to a prime stimulus (with certain evaluative features) for a fraction of a second influences a respondent's reaction towards a subsequently presented target object. For example, when examining racial attitudes, photographs of black and white people may be used as prime stimuli and the subject asked to categorize various positively and negatively valenced target objects as *good* or *bad* (refer to 3.3.1 for more on the procedure of priming tasks). The speed and accuracy of the subject's responses are measured. Since the influence of the prime (e.g. the photograph of a black person) is thought to occur without conscious intention or guidance, the evaluative response on the priming task (e.g. categorizing the target object as *bad* faster than categorizing it as *good*) is said to reflect automatically activated associations relevant to the prime stimulus (e.g. BLACKS and BAD). Yet it remains uncertain whether the prime had an effect on the subject or not.

De Houwer and colleagues, moreover, argue that "[p]riming effects can be based not only on evaluative features of the stimuli," (whether the photograph has a positive or negative valence) "but on a range of other features that may be confounded with the evaluative features" (De Houwer et al., 2009, p. 359). An example of confounding features which they present is salience (defined as "the degree to which a stimulus pops out within a background of other stimuli" (ibid, p. 353).[44] Black faces and negative words are, at least for some individuals, more salient than white faces and positive words, and thus are more easily primed. What this means is that some stimuli (black faces and negative words) are more similar, not only with regards to their valence, but also with regards to their salience. Hence, it is possible that the priming effect (i.e. responding faster to negative valenced words that are preceded by a photo of a black face) is not reflective of a negative attitude towards blacks but rather to the fact that black persons (and negative words) are more salient to the subject.

This possibility poses a risk to the validity of priming tasks, but as De Houwer and colleagues argue, it is "rarely acknowledged" and is given little attention in validity research (De Houwer et al., 2009, p. 359). What follows from these difficulties is that it is also uncertain that

[44] There is experimental evidence to support the hypothesis that black faces and negative words are more salient for white subjects than are white faces and positive words and so white participants respond faster to the category black face and negative words (as well as white faces and positive words) than they would to black faces and positive words (see De Houwer and colleagues, 2009 for more).

priming tasks measure what they are supposed to measure (the psychological underpinnings of implicit bias/attitude), ultimately questioning the strength of their validity.

2) *Reliability and Validity of the IAT*

The Implicit Association Test (IAT) is much easier to carry out than priming tasks (Greenwald, McGhee and Schwartz, 1998; Payne and Gawronski, 2010; Hahn and Gawronski, 2016). However, although it is often depicted as resulting in empirically more reliable and more consistent results, this is only *relative to* priming tasks (Gawronski et al., 2017). In fact, multiple longitudinal studies (Cooley and Payne, 2016; Cunningham, Preacher and Banaji, 2001; Devine et al., 2012; Gawronski et al., 2017) demonstrate that the IAT has low test-retest reliability. The scores on the IAT of a single respondent at different times, are significantly unstable (Payne, Vuletich and Lundberg, 2017a, b). For instance, if (all things considered), I am tested on the race IAT today and then tested again one month from today, I am unlikely to show similar levels of bias. This suggests, as Edouard Machery points out, that the IAT is “an extremely labile” test (2016, p. 118). Although it is the reliability of the IAT for a single individual which concerns me here, it is worth noting that the reliability of the IAT for average scores of a large group is higher than the reliability of the IAT for the scores of an individual respondent. The mean score of a large group who complete an internet IAT measure will be similar to their mean score one month later, but a single individual participant is unlikely to have converging scores one month later.[45]

[45] This puzzling finding has recently been the subject of much attention. To address it, Payne, Vuletich and Lundberg (2017a, b) reconceive the notion of implicit bias as the product of situational factors that make certain associations more accessible than others – or as “the bias of crowds” – rather than as the product of individual agents. For more on ‘the bias of crowds’ hypothesis, refer to Payne, Vuletich and Lundberg (2017a, b). I briefly outline ‘the bias of crowds’ hypothesis in section 7.1 as it relates to the ontological implications of my proposed alternative view.

With regards to the validity of the IAT, researchers agree that empirical evidence is still pending on whether the IAT measures the psychological attribute it is supposed to measure (i.e. implicit attitude) or whether it reflects some other attributes. De Houwer and colleagues (2009) suggest that the results on the IAT may not always be indicative of associations between concepts (or concepts and valences) but in many cases, results may also indicate the influence of confounding variables. By confounding variables these authors include:

1. Extra-personal knowledge such as knowledge of societal values, cultural associations, and common stereotypes (e.g. mothers are nurturing, ISLAM and TERRORISM, men are better at driving cars than women) (Gawronski, Peters and LeBel, 2008).
2. Perceptual similarity or similarity in the perceptual features of the objects (performance on the IAT is faster when categorizing objects with similar perceptual features than when categorizing objects with dissimilar features). For instance, to the extent that pizzas and coins (or rivers and snakes) share similar shapes, they are categorized faster than pizzas and snakes (or coins with rivers) which have differing shapes (De Houwer, Geldof, and De Bruycker, 2006).[46]
3. Salience asymmetries (as mentioned above, this refers to the idea that performance on the IAT is faster when categories assigned to the same key are similar with respect to salience (De Houwer et al., 2009).

The IAT effects, then, may be caused by various confounding variables besides implicit attitudes. Moreover, it isn’t clear whether and how the effects of these confounding variables have been isolated in the experiment. This makes it difficult to interpret the results of the IAT measures.

The IAT has been also criticized for being inherently relative in the sense that it provides relative assessments involving two target objects (Hahn and Gawronski, 2016; Payne and Gawronski, 2010). For example, in the race IAT, if a person shows a preference for white faces

[46] Stimuli can be similar or dissimilar with respect to their perceptual features as well as their affective meaning and this means that IAT effects may be attributable to multiple variables and not just the psychological attribute in question (De Houwer et al., 2009).

over black faces, it isn’t clear whether she has a preference for white faces, an aversion towards black faces, or a combination of both (Payne and Gawronski, 2010).[47] This makes the results of such tests considerably difficult to analyse. What exactly is being measured? It simply isn’t clear if what is being measured is an agent’s prejudice towards black people or their preference for white faces as compared to black faces, or their familiarity with white faces over black faces (which has nothing to do with having a preference of whites over blacks).[48]

Moreover, the processes involved in responding on the IAT measures are intricate and complicated. There is evidence to show that the processes are influenced not only by the (in)compatibility of stimuli and valence (such that they are either sped or slowed down), but also by other confounding processes such as task switching. For example, on a typical IAT task, the subject is instructed to pay attention to two stimulus dimensions (faces as well as words) and to categorize them (on the basis of racial categories as well as valence). Faces and words are presented in alternating order, so the subject has to be continually switching tasks (from paying attention to faces to paying attention to words). Performance is better when stimuli and categories are compatible (either both faces or both words). Thus, task switching may be responsible for the declining performance rates, and not the incompatibility of the category (black face and positive word).

[47] What is being measured by the IAT is comparative in the sense that what is being captured is not directed towards a single target (blacks) but to a target compared with another target (blacks compared with whites). The IAT assesses relative preferences for whites over blacks (or blacks over whites) but not the respondent’s ‘attitude’ towards whites or towards blacks. The critique is that one can have a preference for white faces over black faces without having a negative bias (aversion) towards black faces. Alternatively, one can have a negative bias towards black people (aversion towards black faces) while having a positive preference for white faces over blacks (for more on the validity of IAT see De Houwer et al., 2009).

[48] *Mere-exposure effect* or the *familiarity principle* is a known social psychological phenomenon (Zajonc, 1968). The idea here is that people tend to develop a preference for things merely from exposure to them.

The weak psychometric properties of the priming tasks and the IAT are similarly reflective of the weak psychometric properties of other indirect measures. Other indirect measures often in use in the research include the Go/No Go Association Test (GNAT) (Nosek and Banaji, 2001), the affect misattribution procedure (AMP, see section 3.3.1, footnote 31) (Payne et al., 2005), and many others.[49] These also pose their own challenges to the research.[50] According to a review of the literature by Payne and his colleagues (2017a), indirect measures, in general, suffer from serious complications including (1) low test-retest reliability, (2) low predictive validity (more on this in the next section), and (3) substantial variability in the correlation between different indirect measures (possibly indicative of low validity of one or more of the tests) (more on the variability in the correlation between different indirect measures in section 5.2). Given that such complications afflict the instruments on which the research hangs, it becomes unclear whether a divergence in the results of such questionable instruments may be trusted to construct accounts of implicit bias such as those endorsing the dualistic alignment hypothesis.

Let me summarize what I have discussed so far in this chapter. I highlighted the complexities involved in the interpretation of the results on attitude measurement devices.

[49] The Go-No Go Association Task (GNAT) was developed to overcome the relative nature of scores on standard IAT measures (Nosek and Banaji, 2001). On this measure, participants are asked to press a key (go) in response to some stimuli and another key (no-go) in response to other stimuli. For example, in a block on the GNAT measuring racial bias, participants are instructed to press (go) in response to a photo of a black face or a positive word but not to any other stimuli which may include pictures of white faces or negative words. In another block, participants are instructed to press (go) in response to a photo of a black face or a negative word but not to any other stimuli. The same task is repeated for white instead of black faces. The time given to the participant to respond is usually limited (600ms) and scores are calculated based on the rates of error the participant makes. GNAT has the advantage of capturing scores based on the ‘attitude’ towards an *individual* target object (attitudes towards blacks) instead of relative scores involving two target objects (relative preference of whites over blacks) as is the case with standard IATs. However, the reliability estimates of GNAT are rather low in comparison to the IAT (Gawronski and Brannon, 2017; Gawronski and De Houwer, 2013; Gawronski, 2009).

[50] For more on issues of validity and reliability on these and other indirect measures see De Houwer et al. (2009), Bar-Anan and Nosek (2014), Gawronski and De Houwer (2014), De Houwer and Moors (2010), Lundberg and Payne (2014), Gawronski, LeBel and Peters (2007), and Schimmack (2019).

Although one interpretation for the dissonance could be a genuine distinction between two mental elements underpinning behaviour, two alternative explanations of the dissonance are also plausible. The first alternative explanation refers to methodological differences in the structure and format of the two measurement instruments which may result in a divergence in the responses they generate. The second explanation involves methodological shortcomings related to the validity and reliability of direct and indirect measures of attitude. This short review of the methodological shortcomings around direct and indirect measurement instruments of attitude leads to the conclusion that, at the very least, we must be wary of possible *post hoc* dualistic explanations of the divergence in the responses on these instruments.

4.3 **More on validity: Predictive validity**

I finally turn to discussing the predictive validity of attitude measures. *Predictive validity* refers to the extent to which results on a measurement device predict behaviour. When the results of a measurement procedure are not good predictors of behaviour, this may indicate a methodological (or a conceptual) flaw. To begin with, I briefly list some evidence showing that indirect measures of attitude are presented as having satisfactory predictive validity. I discuss counter evidence showing that their predictive power is in fact very weak. I then explicate two ways of understanding the weak predictive validity of implicit measures. It may be interpreted as a weakness in the methods used, or it may be due to incorrect assumptions of the explanatory theories.

Various studies have found that there is a low to modest positive relationship (correlation) between implicit measures and behaviour; in other words, that implicit measurement devices have some power to predict discriminatory behaviour (Friese, Hofmann and Wanke 2008; Perugini, Richetin and Zogmaister, 2010; Cameron, Brown-Iannuzzi and Payne, 2012; Greenwald et al., 2009; Greenwald, Banaji and Nosek, 2015; Kurdi et al., 2019; Oswald et al., 2013). However, this is a very controversial finding. There are methodological reasons to question that implicit measures are good predictors of behaviour. For example, Schimmack (2019) attempted to reproduce the predictive validity scores of many of these studies but failed. The reason, he argues, is that his statistical analyses account for a prerequisite dimension of predictive validity, namely, *discriminant* validity while the original studies do not. When discriminant validity is low, predictive validity will also be low. In other words, when results on differing instruments, namely explicit and implicit measurement instrument correlate, this indicates low discriminant validity and poses risk to predictive validity.

For example, one of the studies that Schimmack (2019) examines is a well cited study by Greenwald and colleagues (2009) in which the authors tested the predictive validity of the race IAT in voting behaviour. The main aim of this study was to predict whether the race of the 2008 US presidential candidates influenced subjects' voting behaviour. Given that it was the first time that one of the presidential candidates was black and given that there was concern that subjects would not reveal their true attitudes in polls, implicit measures were considered ideal to track whether implicit bias explains voting behaviour. The explicit measures used were Likert type rating scales and feeling thermometers of Blacks and Whites. The implicit measures used were the Race IAT and the Race AMP. The behaviour that was tracked was the voting intention. Greenwald and colleagues reported that the IAT "predicted vote choice independently of the self-report race attitude measures, and also independently of political conservatism and symbolic racism", and they concluded that "these findings support the construct [and predictive] validity of the implicit measures" (Greenwald and Banaji, 2009, p. 242).

Schimmack (2019), however, reassessed the data while addressing discriminant validity of the implicit measures and he found different results. Firstly, he found that there was a high correlation between explicit and implicit measures suggesting that discriminant validity is low. Secondly, in his analysis, the strongest predictor of voting intentions was political orientation (e.g. subjects were more likely to vote Republican if they were in fact Republicans). But while the candidate's race did influence voting behaviour, this was explained simply by explicit attitudes. Implicit attitudes did not significantly explain behaviour above and beyond what the explicit attitudes explained. Schimmack explains that predictive validity ought to be considered along with discriminant validity, otherwise, it may be overestimated. From his analyses of the studies, Schimmack concludes that there is "no evidence that IAT scores predict behaviour above and beyond explicit measures" (Schimmack, 2019, p. 10).

Furthermore, there are other methodological explanations for why the predictive validity of implicit measures is low. One of these involves the principle of correspondence discussed briefly in 4.1. Attitude measures are better predictors of behaviour when there is a clear correspondence between the attitude object and the behaviour. For example, Oskamp and colleagues (1991) (as discussed in Brownstein, Madva and Gawrosnki, 2020) report that prediction of recycling behaviour is low when the considered attitude is generic such as *having an environmentally positive attitude*. Better prediction of recycling behaviour, however, is attained when the attitude being measured is specific towards recycling. Thus, when measuring predictive relations between an attitude and a type of behaviour, and the principle of correspondence is ignored, the expected predictive validity will be low.

Sidestepping the methodological explanations, scholars like Brownstein, Madva and Gawronski (2019) argue that orthodox theories may do better to consider the variables involved in attitude-behaviour relationships. To the extent that theories consider implicit measurement results as direct indicators of attitudes, the predictive power of implicit measures is *expected* to be high. But if equating implicit measurement results with implicit bias is resisted, then the expectations for prediction of behaviour will be reduced. Furthermore, Brownstein, Madva and Gawronski (2019, p. 7) point out that what is of critical importance for prediction is not merely how well an implicit or an explicit instrument predicts behaviour, rather it is how well research can discover the "domains" and the "conditions" under which the instrument may predict behaviour. Said differently, the stress is on the role of the numerous variables (alongside the attitude) involved in predicting behaviour. Such variables include (in addition to the particular behaviour being measured), the conditions under which the behaviour takes place and the person who undertakes this behaviour (ibid).

Indeed, research acknowledges the fact that behaviour prediction is a difficult task (Brownstein, Madva and Gawronski, 2019; 2020; Gawronski, 2019). As Brownstein, Madva and Gawronski (2020, p. 4) argue, "the thought that any specific attitude will predict a range of behaviour, regardless of behaviour specific, context-specific, and person-specific variables, conflicts with basic long-understood truisms about the mind". In chapter 6, I propose an interpretive view of the data which takes into consideration this interaction of a range of variables in behaviour as it aims to resolve the conceptual difficulties facing orthodox theories brought on by the dualistic hypothesis.

4.4 Conclusion

My main motive in this chapter has been to throw further scepticism on what I consider to be a misguided dualistic interpretation of the data. The finding of dissonance in the evaluative responses tracked by dual types of measurement instruments, I argued, doesn't (only) lead to the conclusion that there is a corresponding explicit-implicit dualism at the level of underlying mental elements. An explicit-implicit dualism in the underlying mental elements is only one interpretation of the data. The dissonance may be interpreted as resulting from a difference in the structural parts of the measurement instruments (they lack structural fit). It may also be interpreted as the result of weak psychometric properties of the measurement instruments. If the tools employed are psychometrically dubious, then the results they produce may be construed as the result of error.

It may be that the methodological difficulties just reviewed are of little concern to many dualistic philosophers; that might explain why they are often neglected. More alarming is the fact that even empirical research often dismisses these difficulties and instead continues to employ weak measurement instruments. When suboptimal direct measures are used to control for explicit attitudes, it is likely to lead to an overestimation of the predictive utility of indirect measures. If attitude measures have significant problems in their psychometric attributes, one is at a loss why more philosophers have not practiced a wary scepticism on dualistic interpretations of bias. Clearly, the research needs to get its psychometric house in order before it goes off making grand conclusions and discussions involving the architecture of the mind or even moral culpability. In any case, it must build itself on strong foundations before the discussion continues.

Chapter 5

UNITY, HETEROGENEITY, COMPLEXITY

Three Objections to the Dualistic Alignment Hypothesis

Introduction

After having raised scepticism around the theoretical development (chapter 3) and the methodological basis of the dualistic hypothesis (chapter 4), in this chapter I raise scepticism around its conceptual basis. I put aside the psychometric limitations and for the sake of argument, I assume that alternative explanations explain only some, but not all, the variance in the results. I grant the assumption that the responses on direct and indirect measurement instruments reflect a difference in psychological underpinnings as per the dualistic alignment hypothesis. My aim here is to show that the assumption that these psychological underpinnings form two *distinct* and *uniform* categories is nevertheless mistaken. There is no category of uniform mental elements that is to be identified as *implicit attitudes* and as that which guides micro-behaviours. There is also no clearly distinct category of uniform elements that underpins what is referred to as *explicit attitud*es, generates results on explicit measurement instruments, and guides macro-level actions. Rather, the psychological underpinnings are heterogeneous and differ in degrees of *explicitness* and *implicitness* and guide behaviour in complex ways. (I will come back to this in chapter 7 when I discuss eliminativism as an ontological implication of my alternative view, *the mosaic view*).

To this end, I raise three challenges which are difficult for the dualistic hypothesis to overcome. In 5.1, I raise the first challenge that there is no feature of cognitions which allows them to be carved neatly into the explicit-implicit distinction that the dualistic view requires. I call this the *unity challenge.* In 5.2, I pose the second challenge that what the dualistic view classifies as two categories of uniformly structured mental elements (the explicit and the implicit) are not uniform. In fact, there is much heterogeneity that is unaccounted for by the dualistic view. I call this the *heterogeneity challenge*. Section 5.3 is where I raise the last challenge which is that there is much complexity involved in the performance of behaviour including various moderating and mediating factors. This makes the relationship between attitude and behaviour very complex, and that's why I call this the *complexity challenge*.

5.1 <u>The Unity Challenge</u>

I begin with the first challenge. One of the main assumptions of the dualistic hypothesis is that there is a principled distinction between explicit and implicit mental categories. Against this, I raise the unity challenge and I argue that although there may be several conceptually coherent ways of characterizing the explicit-implicit dualism at the level of psychological underpinnings, there is enough empirical evidence to show that none of these ways is successful at maintaining a principled distinction (Stammers, 2016).[51] In a recent article aptly titled 'What is Implicit Bias?' Holroyd, Scaife and Stafford (2017, p. 3) review various characterizations of the implicit and conclude that "no one view unproblematically carves our cognitions into implicit and explicit". Below I review three of the most researched characterizations of the explicit-implicit distinction, namely the features of control (5.1.1), introspective awareness (5.1.2), and reason-responsiveness (5.1.3). In 5.1.4, I discuss the explicit-implicit distinction as it relates to the type of measurement procedure (explicit/direct vs implicit/indirect). I conclude that none of these considered characterizations succeeds at drawing a principled distinction between explicit and implicit attitudes.

[51] See Holroyd, Scaife and Stafford (2017) for more on how there isn't any view which can unproblematically draw a divide between implicit and explicit cognitions.

5.1.1 The explicit-implicit distinction characterized by control

I begin with the first characterisation of the explicit-implicit distinction: control. The dualistic hypothesis generally describes explicit attitudes as involving controlled processes, e.g. as beliefs one can deliberate on and control. On the contrary, implicit attitudes are described as mental elements (e.g. associations) involving automatic processes over which one lacks control. Control, in effect, is taken to set the distinction between what is captured by direct and what is captured by indirect measures (explicit beliefs are controlled, implicit biases are not). I argue that this characterization of the distinction is not as simple as the defender of the dualistic hypothesis (henceforth *the dualist*) presents it. The interplay between explicit – as involving controlled processes – and implicit – as involving automatic processes – is, in fact, rather complex. There are varying ways in which control can be exercised on implicit attitudes thereby allowing them to be characterized not strictly as spontaneous and uninhibited, but also as controlled.

To examine the varying ways of control, I start by drawing a distinction between (1) the *activation* and (2) the *expression* of implicit attitude, and then I argue that although control may not feature over the activation of an implicit attitude, it can feature in its expression. By *activation* of implicit attitude, I mean the activation of the psychological element implicated in the attitude, e.g. the association of WOMEN and WEAK. By *expression*, I mean the influence of the activated psychological element on behaviour, e.g. interrupting women.

The predominant discourse on implicit bias considers control strategies to be generally operative only in direct measures involving explicit attitudes.[52] The True Attitude stream, for

[52] See for example Devine (1989) and Fazio (1990) for work on the True Attitude line of research, and Frankish (2016) for philosophical discussion of dual-process theories of control. For a detailed discussion on the different modes of control strategies, see Holroyd and Kelly (2014).

example, claims that the difference between explicit attitudes and implicit attitudes is that the former involves controlled psychological processing (like deliberation and inhibition) while the latter involves automatic processing (spontaneous and inescapable). Explicit beliefs, for example, are understood as states whose influence on behaviour involves control even when relevant mental elements are *activated* automatically. In this sense, the *expression* of the behaviour involves some form of control, even if the *activation* of its psychological underpinnings does not (see 3.3.2 on the MODE model).[53] On a self-report, for example, a subject may engage in control strategies to inhibit or reduce the influence of bias, i.e. its expression (e.g. the sexist professor may repress his sexism and claim his egalitarianism towards women).

However, strategies of control are not typically thought to be operative in indirect procedures. Indeed, these procedures are thought to capture the sort of thing that is automatically (defined as inescapably) *activated*, i.e. the implicit attitude (e.g. the association WOMEN and WEAK) and that is automatically *influential* in behaviour (e.g. the association automatically influences how the professor will speak to women – he will interrupt them) (Payne and Gawronski, 2010). Said differently, the implicit attitude (WOMEN and WEAK) will be automatically *activated* upon encounter with a stimulus (woman colleague) and it will also *influence* behaviour automatically (the professor has no control over his interrupting women).

Clearly, the overarching understanding of implicit attitudes is that they are characterized by automatic processes or "inescapable habits expressed despite attempts to bypass or ignore them"

[53] The MODE model (3.3.2) describes explicit attitudes as indicated by the responses on measures where control is possible (e.g. on a self-report). The respondents have ample time and cognitive resources and perhaps also have the motivational factors not to appear prejudiced. And although the mental element (the association BLACKS and DANGER) involved in the self-report is automatically activated for the subject, its *expression* in action can be controlled by the subject (e.g. he can report that blacks are no more dangerous than any other race).

(Dasgupta, 2013, p. 238). On this understanding one can claim, as Saul (2013, p. 55) does, that individuals “do not [when made aware of their bias] instantly become able to control it”. Such an understanding gives the impression of immutability. The question is whether this immutability is characteristic of both the *activation* of the mental elements and its *expression*, i.e. its influence on behaviour.

To answer this question, consider whether control may be exercised over the *activation* of implicit attitudes. Are implicit attitudes activated spontaneously and independently of any factor such as attention and agent awareness (apart from minimal perceptual acuity of the stimulus)? Empirical research seems to suggest as much. For example, if implicit attitudes are taken to be associative, then the presentation of a concept (ISLAM) will (in certain individuals) automatically elicit the activation of the concept (DANGER) (Fazio, 2001; Gawronski and Bodenhausen, 2007). The activation of these associations may very well be inescapable. That generally *is* what typifies the characterization of implicit attitudes; that they are spontaneously activated. Consequently, we may agree that control does not feature in the activation of implicit attitudes. But what about the influence of implicit attitudes on action? It is here in what I labelled as the *expression* of implicit attitudes where control may be exercised. Or so I argue.

Before continuing, let me clarify an important notion: any picture characterizing implicit attitudes as akin to nervous system processes responsible for bodily functions such as knee jerk reactions, arm spasms, and eye dilation is a misconceived picture (as I argued in 2.2). To adopt it is to misdescribe implicit bias as “the spasm of a hand, or perhaps the intrusion of compulsive or phobic thoughts” as Holroyd and Kelly (2016, p. 3) would agree. And as Mandelbaum (2016, p. 649) aptly argues, micro-behaviours or the expression of implicit bias in action (such as seating distance, interruptions, eye contact, shooting behaviour, CV choice etc.) are “decidedly different than paradigmatically reflexive behaviours such as the deep tendon reflex that controls knee jerks”. In effect, the expression of implicit bias in action involves a multitude of processes that are significantly more complex than simply the uncontrolled activation. If this is so, then there are ways in which control may be applied on any one of these various processes. In fact, the research suggests that the expression of implicit bias is indeed malleable, i.e. there are ways in which implicit bias can be controlled (Amodio and Swencionis, 2018). I discuss some of these in the remainder of this sub-section.

Firstly, control may feature in an agent’s expression of the otherwise spontaneously activated attitude by *inhibiting* its influence on behaviour. Take for example, an agent who is self-motivated (and socially motivated) not to be sexist, but whose score on indirect measures suggests that he harbours the implicit bias involving WOMEN and POLITICAL INCOMPETENCE. This individual is not inevitably doomed to express his implicit bias in behaviour. He may, for instance, gather counter stereotypical information (in this case about women) which invalidate his automatically activated biases (Holroyd and Kelly, 2016, p. 9). This individual might think of strong women leaders like Angela Merkel and Jacinda Arden to mitigate the influence of the automatically activated association (WOMEN and POLITICAL INCOMPETENCE) on his behaviour. In this way, although the implicit attitude is automatically activated, it nonetheless may not exert its influence on behaviour because other information will invalidate the activated association and instead influence the judgement (Monteith and Pettit, 2011).

Secondly, control may take the form of *correcting* the biased behaviour by adjusting the responses prior to the act. Such strategies are employed as interventive techniques in anti-prejudice training programs and have shown some efficacy in one's ability to control implicit bias.[54] In shooter bias simulation experiments, for example, participants who adopt "if-then" strategies to behave in non-biased ways (also known as *implementation intentions*) have their biased responses decrease significantly (Mendoza, Gollwitzer and Amodio, 2010; but see also Holroyd and Kelly, 2016; Stammers, 2017; Brownstein, 2017). Other experiments which expose participants to images or short movie clips showing members of stigmatized groups acting in stereotype-discordant ways also show significant decrease in implicit bias response (Blair, Ma and Lenton, 2001; Dasgupta and Greenwald, 2001). Holroyd and Kelly (2016, p. 6) discuss these strategies as a form of "ecological control" or automatic type of control that intervenes in the expression of implicit bias.[55] The literature on control is vast and complex but for my purposes, suffice to note that (at least) the influence of implicit bias on behaviour is not spontaneous in any reflexive or inescapable manner. Rather, there is a real sense in which it can be controlled.

Furthermore, external conditions have been shown to affect/moderate how implicit bias plays out. For example, several factors play a part in influencing the expression of implicit bias making it inaccurate to consider it as spontaneous and uncontrolled. Such factors include

- Contextual factors such as the context in which a stimulus is encountered (say encountering a black person in a basketball court will elicit different responses than encountering the same person in a back alley) (Gawronski and Bodenhausen, 2011; Gawronski et al., 2014; Gawroski and Sritharan, 2010).

[54] For an elaborated discussion on strategies for exercising control on bias see Monteith, Wookcock and Gulker (2013) and Holroyd and Kelly (2014).

[55] Ecological control is a form of control which involves cognition being distributed along environmental 'props' – it is a sophisticated term that describes an umbrella of the strategies discussed above such as exposing oneself to counter-stereotype images, adopting if-then strategies, actively committing oneself to egalitarian goals, all of which help in manifesting less implicit bias (Holroyd and Kelly, 2016 citing Andy Clark, 2007).

- External variables specific to the target object (such as prototypical Afrocentric features are shown to elicit more negative responses on indirect measures) (Livingston and Brewer, 2002).[56]
- Internal variables specific to the subject/respondent such as negative emotions (positive emotions in the respondent elicit less expression of implicit bias than do negative emotions) (Dasgupta et al., 2009; DeSteno et al., 2004).[57]

It is the interaction of the activated association with moderating variables (contextual factors as well as factors external and internal to the subject) which has an effect on the expression of the activated association. To the extent that such variables influence the expression of implicit attitudes, there is a sense in which these attitudes may be controlled.

To sum up, implicit attitudes do not always involve spontaneous and automatic processes in the sense that dualistic alignment theorists describe, at least some may involve controlled processes. If this is right, then there is reason to be sceptical about drawing a sharp divide between the explicit and the implicit on the basis of control. Perhaps the feature of introspective access may do better at characterizing the explicit-implicit distinction. This is what I examine next.

5.1.2 The explicit-implicit distinction characterized by introspective awareness

I argue that there is no reason to think that what distinguishes explicit from implicit attitudes is that the former is open to introspection while the latter is not. First, I explain why one might think that the explicit-implicit is characterized by introspective awareness and then I argue that empirical studies do not support such a characterization.

[56] Take for a clearer example that the physical attributes of members of stigmatized groups can have different effects on implicit bias. Targets with prototypical Afrocentric features elicit greater negative implicit bias than targets with less prototypical Afrocentric features (Livingston and Brewer, 2002).

[57] For example, negative emotions such as anger increase implicit anti-Muslim bias, and disgust increases implicit anti-gay bias (Dasgupta et al., 2009).

The general understanding of the dualistic hypothesis is that explicit attitudes are introspectable (subjects can become conscious of them) while implicit attitudes are not open to introspection (subjects are typically unaware of having them). The conscious-unconscious distinction is of particular significance to implicit bias research. For example, the terms *implicit* and *unconscious* often typify the discussion of implicit bias to the extent that both terms are used interchangeably in academic research (for example Saul, 2013b) as well as in public discourse.[58] The conscious-unconscious distinction is also the main idea behind the 'Driven Underground' stream of research (discussed in 3.4). The Driven Underground stream gathers support from anecdotal evidence that individuals are (or seem to be) *surprised* at 'discovering' their implicit bias after taking an indirect attitude test (Brownstein, Madva and Gawronski, 2019) and from the low correlation found between results on direct and indirect measurement instruments (Nosek, 2005, 2007; Hofmann et al., 2005; Cameron, Brown-Iannuzzi and Payne, 2012).[59] I examine this evidence below.

First, I consider the anecdotal claims of respondents who are surprised at their results on indirect measurement tools. To be sure, it is unclear to what extent individuals *genuinely* experience surprise (as opposed to merely *perform* surprise) at the findings, given that there is so much social pressure not to appear biased. It is further unclear what percentage of people do in fact experience surprise as this information is anecdotal (Brownstein, Madva, and Gawronski, 2019). Moreover,

[58] For different views on the public discourse on the unconscious nature of implicit bias see for example Sirois (2017) and Cherry (2019).

[59] It is interesting to note that people who are surprised to learn about their scores on implicit tests often flatly deny the accuracy of the results and even deny having any implicit bias (Howell, Gaither and Ratliff, 2015, Howell and Ratliff, 2017). Brownstein and colleagues (2019) humorously explain that subjects claim surprise at 'discovering' they harbour implicit bias in the same way they would be surprised at discovering they had high cholesterol levels through relevant blood tests. This sounds as though one had never encountered or could possibly know without the intervention of sophisticated testing (like blood tests or an IAT) (Brownstein, Madva, and Gawronski, 2019, p. 3).

as Brownstein, Madva, and Gawronski (2019, p. 3) argue, it could be "that some people are less disposed to introspective self-examination than others" and thus feel more surprise to learn about their biases than others who are more disposed to examine their cognitions more regularly. Relatedly, while many people may be able to introspect into their attitudes, it could be that only a few do so (refer back to 4.2.2 for a discussion on the psychometric properties of explicit measures). It is also unclear to what extent individuals are merely being defensive (and thus surprised) in light of the 'accusations' of bias which runs counter to their experience of themselves and their explicit self-conception (Howell and Ratliff, 2017). Thus, subjects' surprise at discovering their implicit bias (on implicit measures) does not show that their bias is unconscious.

Second, I examine the idea that low correlation (or divergence) between the results on direct and indirect measures supports the explicit-implicit as conscious-unconscious distinction. I have already argued in chapter 4 that this divergence is suspect. Here I add that, even if it were not suspect, a divergence is just not informative about the awareness question (Hofmann et al., 2005; Blanton and Jaccard, 2015; Gawronski, 2019). For to claim that indirect measures do not require the respondent to introspect doesn't mean that what is being measured is itself not open to introspection. If the participant is unaware of what the instrument is tracking, it simply doesn't follow that what is being measured by the instrument (i.e. the implicit attitude) is itself not introspectable. As Gawronski (2019, p. 575) puts it, while implicit tasks may have the "potential" to capture unconscious mental elements, this doesn't imply that subjects are unaware of what is driving their responses on these measures.

Furthermore, there is much ambiguity in the claim that the divergence between the explicit and the implicit is characterized by awareness. To be sure, it is not clear exactly what the subject is unaware of. Gawronski, Hofmann, and Wilbur (2006) argue that people may be unaware of the *origins* and the *influences* of their implicit bias on behaviour but there is little evidence that they are unaware of the *content* of their implicit biases. They draw distinctions between three aspects of awareness of indirectly accessed attitudes: (a) *source awareness* or a subject's awareness of the origin of the bias, (b) *content awareness* or a subject's awareness of the bias itself, and (c) *impact awareness* or consciousness of the consequences this bias has on other mental processes. With respect to (a) *source awareness,* the authors show empirical evidence that it is a feature lacking in both indirectly assessed (implicit) attitudes as well as self-reported (explicit) attitudes. In other words, people are often unaware of the causes of their attitudes, whether these are explicit or implicit. For example, in explicitly reported preferences, such as *love for wine* or *hatred for pickles*, people may not be aware of (a) the source of these preferences. With respect to (c) *impact awareness*, the authors found some evidence that it differs for the explicit and the implicit. Individuals may not try to curb their prejudiced behaviour, simply because they are unaware of the influence their implicit bias has on their behaviour. For example, subjects may be unaware of the impact that their implicit racism has on maintaining a larger distance in interaction with black people than with white people, thus they may not try to control it. There is interesting empirical work on source and impact awareness, but for my purposes, I focus only on (b) *content awareness* insofar as that is what seems to be at issue in the implicit bias literature.[60]

[60] Holroyd (2015) draws a different distinction for three dimensions of awareness. The distinction comes in the context of moral responsibility for action and her paper she addresses the question whether or not individuals should be held accountable for their implicitly biased behaviour. The first dimension she calls *introspective awareness* and it refers to the subject's awareness of the "cognitive processes that produce action" or the "way in which an association is activated" (Holroyd, 2015, p. 516, 519). It is not entirely clear from Holroyd's (2015) discussion whether *introspective awareness* also refers to awareness of the contents of the attitude. If so, then it would overlap with Gawronski and colleagues (2006) *content awareness*. The second dimension she calls *inferential awareness* and it refers to the subject's awareness of the body of knowledge involving how people harbour and exhibit implicit bias (e.g. empirical evidence from social psychology such as the studies reviewed in section 2.3). The third dimension of awareness she

Let us then follow the interpretation that the divergence in the results on the two types of measures reflects a difference in (b) *content awareness*. The dualistic hypothesis, recall, favours a dissonance in content awareness between explicit and implicit attitudes: that one is conscious of their explicit attitudes but unaware of their implicit attitudes. This interpretation, however, needs re-examination given developing empirical evidence. Take for example, Gawronski, Hoffman and Wilbur (2006) who build an empirical case against the hypothesis that implicit attitudes are unconscious. They argue that people are generally able to report attitudes tracked by indirect measures. However, whether or not these attitudes are reflected in the responses on self-reports depends on "cognitive and motivational factors" (Gawronski, Hoffman, and Wilbur, 2006, p. 490). In other words, people are generally "conscious of their attitudes as they are reflected by indirect measures", but various factors may "undermine the influence of these attitudes" on direct measures, hence the divergence (ibid).

They cite a seminal study by Jason Nier (2005) which had participants believe that their responses on the IAT are "the closest thing to a lie detector that social psychologists can use to determine [one's] true beliefs about race" (Nier, 2005, p. 43). In this study, the convergence between participants' self-reports (explicit attitudes) and their indirectly assessed (implicit) attitudes was significantly higher when participants thought their responses were being monitored by a lie detector (i.e. when they thought inaccurate reports would be detected). What this suggests

calls *observational awareness* and it refers to the subject's awareness of the influence of the automatically activated attitude on her behaviour. *Attributional awareness* (a subset of *observational awareness*) refers to the awareness that some feature of a subjects' cognition is influential in the biased behaviour. *Observational* (along with *attributional*) *awareness* overlaps with what Gawronski and his colleagues call *impact awareness*.

Holroyd (2015) argues that while individuals cannot (and ought not) be expected to have awareness of the operation of their cognitive processes (introspective awareness) or to engage in the complex body of academic around implicit bias (inferential awareness), they should nonetheless have knowledge of the morally relevant features of their actions (observational awareness).

is that subjects are not unaware of the biases reflected by indirect measures but social desirability concerns, to a large extent, undermine the influence of the biases on a subject's self-report (i.e. explicit attitude). If subjects were not aware of their bias, the bogus pipeline would lead to slight "shifts in the mean values" of explicit attitudes but not to a higher convergence between explicit and implicit attitudes (Gawronski, Hoffman, and Wilbur, 2006, p. 490).

Another much-cited series of studies conducted by Hahn and colleagues (2014) found that agents were surprisingly accurate in predicting their scores on the IAT, lending further support to the suggestion that people may be aware of the *content* of their implicit biases (see also Rivers et al., 2017). In this study, participants were asked to predict their scores on multiple IATs towards different social groups and then participated in the IATs as well as in direct measures. Half the subjects were "informed" that implicit attitudes refer to "cultural associations" that may or may not reflect their true attitudes (this was meant to remove the influence of any motivational factors), while the other half were told that implicit attitudes reflect their "true attitudes" (Hahn et al., 2014, p. 6). It was hypothesized that if respondents were influenced by motivational factors of social desirability, then their predictions will be worse on the "true attitudes" condition than on the less threatening "cultural associations" condition. The respondents were asked to make their predictions by judging the degree of ease by which they would sort various stimuli between two categories as with the IAT. They were encouraged to look at the stimuli (for example pictures of faces of black and white people) and "listen to their gut-feelings" to predict which sorting task would be easier (the one involving the black or the white faces) (ibid, p. 8). On a modified study in the same experiment, participants were asked to predict *directly* their "true implicit attitudes" or their "culturally learned associations" by rating the statement "I predict that the IAT comparing my reactions to BLACK vs. WHITE will show that my true implicit attitude [culturally learned association] is…" on a Likert type response scale (p. 12). While on further trials, participants received very minimal instruction on what the experiment was about. They were just told that a new study called the IAT was supposed to examine people's implicit and explicit attitudes, and they were instructed to predict how they would score on the IAT. Hahn and colleagues (2014) concluded that the high level of accuracy of the subjects' predictions revealed that they had awareness into the association strengths measured by the IAT. Accuracy was high regardless whether participants had prior experience with the IAT, whether they received information about the IAT, or whether they were told the IAT reflected their 'true attitudes' or 'cultural associations'.[61]

Moreover, recent studies (e.g. Hahn and Gawronski, 2019, p. 3) suggest that simply telling participants to attend to their "spontaneous affective reactions" to certain minority groups significantly increases their acknowledgment of harbouring bias towards those groups. Finally, studies where participants construed implicit biases towards minority groups (e.g. gay men) as their own (i.e. as belonging to them) reported those same 'attitudes' explicitly (Cooley et al., 2015).

To sum, there is good empirical reason to think we can be conscious of the content of (at least some of) our implicit biases. As Hall and Payne (2010, p. 3) report "hard evidence that people have attitudes and beliefs that they don't know about, or can't know about when they try, is difficult

[61] This was so even when there was a divergence in the results of participants' explicit and implicit measures indicating that the predictions were not being made on the basis of subjects' explicitly reported attitude of which they were aware.

to find”. The lack of convincing evidence has led leading psychologist in the field Gawronski to conclude that,

> counter to a widespread assumption in the literature, there is currently no evidence that people are unaware of the mental contents underlying their responses on implicit measures. If anything, the available evidence suggests that people are aware of the mental contents underlying implicit measures [meaning results on implicit devices]…which allows them to predict their implicit bias scores with a high degree of accuracy (Gawronski, 2019, p. 5).

Of course, this is not to deny that people have unconscious beliefs, desires, and other mental states, nor that a great deal of our mental lives is consciously inaccessible. The point, however, is that indirect/implicit measures do not track and capture a mental state that is unconscious and introspectively inaccessible.[62]

[62] The flipside of this, and only briefly for the sake of space, is something that was suggested in the previous chapter in the section critiquing direct measures. It is the contestable idea that explicit beliefs may themselves be characterized as those which are introspectively accessible, or whether they may *always* be characterized as accessible to introspection. It isn’t obvious that if we profess to know our beliefs, we have a clear understanding of what these are. It is plausible, in fact possible, as the research on introspection and self-enhancement suggests, that we are mistaken about our beliefs. This may be the case when we misrepresent how non-prejudiced we are (as we engage in self-enhancements and believe we are egalitarian when in fact we either have not properly introspected into our beliefs or we just are unable to properly introspect into them). I don’t delve into this claim because it is not taken up seriously by implicit bias debate (For more on knowing one’s own beliefs, see Carruthers, 2013, 2017, and for additional information on self-enhancements of one’s beliefs, see Palhaus and Reid, 1991).

5.1.3 The explicit-implicit distinction characterized by underlying processing structures

In this sub-section, I examine whether a principled distinction between explicit and implicit attitudes may be drawn on the basis of a difference in their underlying structure: namely that the latter is associatively structured while the former is propositionally structured. I argue that such a distinction does not hold. (At least) some mental elements constituting implicit attitudes and described as associatively structured may be responsive to reason in a similar way as propositionally structured explicit attitudes. If (at least some of) these elements are responsive to reason, then they are not associatively structured.

While orthodox dualism advances associative accounts of implicit bias (most philosophical accounts of implicit bias take it to be associatively structured), there is reason to question this. For dualist theorists, implicit biases are activated independently of whether the subject endorses them, and because they are not validity-apt nor reason responsive in the same way as explicit beliefs are, they are distinct from explicit attitudes. Although as I illustrated in section 2.5, the predominant view is that implicit bias is associative and not responsive to reason in the same way as beliefs (e.g. Gendler, 2008a, b; Madva 2016; Brownstein, 2017), some authors have argued that implicit bias is responsive to reason (e.g. Mandelbaum, 2016 and his argument that the enemy of my enemy is my friend), while others have argued that it is patchy in its responsiveness (e.g. Levy, 2015). I believe that all these theorists are convincing in their arguments about the nature of some of the mental elements responsible for our micro-level behaviour. And if indeed they are right in their description of what constitutes implicit bias, it suggests that there is no *one* way for implicit bias to be.

Take Brownstein (2017, p. 11) who claims that the explicit-implicit dualism reflects two "faces" of prejudice: one face (the explicit) reflects mental states and processes which are "validity-apt" and the other (the implicit) reflects states and processes which are "not validity-apt" (refer to section 2.5). By *validity-apt* (or the explicit face), Brownstein refers to the kind of processing that reflects what the person "takes a stand" towards, what she considers "true and false, mine or not mine, valid or invalid" (Brownstein, 2017, p. 12). In other words, states and processes which are validity-apt are ones which respond to some form of logical reasoning. The explicit face of prejudice involves *propositional processing* because it is concerned with validation of propositions like 'Are Arabs dangerous?'. Propositional processing underlies what are typically called *explicit beliefs* and concerns whether these are true or false (Is it true that Arabs are dangerous?). The implicit face of prejudice involves *associative processing* which does not concern validation of propositions. Rather it involves associations between things like ARABS and DANGER, and since there's no relation of syntax, there is no concern for whether this association is true of false. Mental processing which is associative is not validity-apt because it is not responsive to any form of logical reasoning nor to mental states we consider to be reason-guiding, like our beliefs, commitments, normative standards, and other values.

Brownstein's discussion is built on the Associative-Propositional model of evaluation (APE) developed by Gawronski and Bodenhausen (2006; 2007; 2011; 2014a, b; 2017) (refer to footnote 42 for a brief discussion of the APE model). The APE model claims that unlike propositional processing involved in explicit measures, what underpins responses on implicit measures is associative processing which is independent of the assignment of truth values. Brownstein (2017, p. 12) calls associative processing exemplary of "arationality" because it is "insensitive to rules we ourselves set down; that is, our beliefs, values, and ideals". Measures of our implicit attitudes, he claims, do not show our true attitudes, nor do they show our "sublimated underground attitudes. Rather they reveal our arational attitudes" (ibid). This resonates with Gendler's (2008a, b) view which conceives of implicit attitudes as reality insensitive aliefs (a unique kind of mental associative state that does not fall into any of our existing mental categories).

So how does the propositional-associative or rational-arational characterization fail to carve cognitions into the explicit and the implicit? I turn to two of Eric Mandelbaum's (2013; 2014; 2016) arguments against aliefs to illustrate. The first argument Mandelbaum (2013) calls the *binding argument*, the second, he calls the *argument of 'inferential promiscuity'*.

I begin with the *binding argument* which presents a convincing case against the claim that implicit attitudes are associatively structured and not responsive to reason. Mandelbaum (2013) claims that the content of an associatively structured alief (implicit bias) must contain propositional content, essentially it must be validity-apt in some sense for it to affect behaviour in any meaningful way. Take for example the content of the associative alief 'BLACK, DANGER, AVOID'. The question is if these components have an associative structure, i.e. are tokened in succession, then what motivates the subject experiencing them to behave in any particular way towards *that particular encountered black person* as opposed to anything else? Why would the subject 'AVOID' that particular 'BLACK' person, rather than 'AVOID' anything else that is around her? What makes the behavioural component 'AVOID' bind to the particular 'BLACK' person encountered? The binding argument concerns the notion that the avoidance behaviour has no way of attaching (or binding) to *that particular encountered black person* as opposed to anything else. For the content of the alief to bind, it needs to have some structure, and an associative kind of structure doesn't allow for the kind of structure needed for the motivation to act.[63] The structure must have some propositional content of the type: 'this is a black person', 'black people are dangerous', 'avoid crossing his path'. At least, in some sense and to some extent, (some) implicit attitudes have propositional structure. And since propositions are truth apt, (some) implicit attitudes are not arational.

Mandelbaum's (2016) second argument (which he may have adopted from Jan De Houwer's 2009 work) begins with the explanation that propositions give relational information, in other words, they inform how two concepts are related. Consider the proposition that 'blacks are dangerous', there is relation between the two concepts ('blacks' and 'dangerous'), namely, that 'blacks *are* dangerous'.[64] Associations, on the other hand, only inform that certain concepts are linked together. A simple association between BLACKS and DANGER, for example, does not provide information on the direction of relation: it only informs whether activating the concept BLACKS activates the concept DANGER or the other way around.[65] The nature of the relation is unidentified in associative structures. Generally, in associative models of implicit bias, the relation

[63] It is worth quoting Mandelbaum's argument at length.

> …when I token the alief with content CYANIDE, DANGEROUS, AVOID, what am I thinking? If I am just tokening these concepts in succession (which is what Gendler's 'associative state' talk implies), then why would I show any behavior whatsoever toward the bottle (and its contents) and not, say, the window, my left foot, or the experimenter's forehead? Since the behaviour is bottle/bottle-content specific, the putative alief must somehow bind to the bottle (and its contents), or else participants would not show the avoidance behaviour toward it. Merely saying that the alief's content is associated with the bottle does not explain why the alief binds to the bottle (and its contents) alone (Mandelbaum, 2013, pp. 7-8).

Refer back to section 2.5 for more details on this example.

[64] For a short discussion on *the unity of the proposition and its problems* see McGrath and Devin (2018) and for a detailed discussion refer to Gaskin (2008).

[65] Most theorists discuss two place associative relations (between concept and valence or between two concepts) and although Gendler talks about paradigmatic aliefs as being four-place relations, she allows for more loose usage in some contexts (Gendler, 2008b, p. 559).

between associated concepts is unidentified; the concepts are activated together regardless of the kind of existing relation. The associations are merely products of conditioning. They are not formed on the basis of logic or reasoning; they are simply the co-occurrence of two concepts (or a concept and a valence) together as the result of repeated pairings (e.g. SALT and PEPPER). This implies that 'blacks are dangerous' and 'blacks are not dangerous' both are activated by the concepts BLACKS and DANGER. For associations to interact with any mental states to cause behaviour in the way they do, they cannot be truth insensitive. Mandelbaum calls this the *argument of 'inferential promiscuity'* perhaps mostly in response to Levy's (2015) conception of implicit bias as patchy endorsements.[66] Implicit attitudes, according to Mandelbaum, need to be inferentially promiscuous in order for them to interact with other mental states and beliefs and to affect behaviour in the way that empirical work shows that they do. (For a state to be inferentially promiscuous means that it interacts with, and responds to the semantic content of, other mental states).[67]

If Mandlebaum's arguments are right, and if (at least some of) what the literature calls implicit attitudes are reason-responsive (and not associatively structured), then the explicit-implicit distinction based on the rational-arational or propositional-associative distinction is not as robust

[66] Mandelbaum is generally concerned with critiquing the research that considers the structure of the mental representation responsible for implicit bias as purely associative, he nonetheless maintains a clear distinction between implicit bias and explicit attitude as two different mental states.

[67] The activation of the concepts BLACK, DANGER, AVOID in succession in my mind as I encounter a black man on the street doesn't tell me which path to avoid crossing. Mandelbaum (2012) questions the extent to which such an activated association guides the individual one way or another. What would it mean for an individual to have the association 'BLACKS' and 'DANGER' or 'BLACK' and 'AVOID' automatically activated without syntax or relation attached to it? To think of implicit biases as associations does not allow these biases to affect behaviour in the way that they do. Mandelbaum (2016) provides an extended critique based on empirical studies to show how implicit bias follows a 'logic' that is not merely associative. He reviews various research findings from social and cognitive psychology to show that implicit biases are sensitive to inferences, responsive to reasons, and to the strength of arguments.

as the dualistic view requires it to be. A principled distinction between a uniform single kind of mental state described as explicit and a uniform and unified kind of mental state described as implicit is thus not achieved.

5.1.4 The explicit-implicit distinction characterized by measurement results

In this final sub-section, I examine (and reject) the understanding of the explicit-implicit distinction as reflecting a divergence in the results detected by two different measurement procedures (and the behaviour they predict). I call this the *parallel alignment* pattern. *Parallel alignment* indicates a dissociated alignment, between on the one hand explicit procedures, the explicit attitudes they track, and the behaviour they influence, and on the other hand, implicit devices, the implicit attitudes they capture, and the behaviour they generate (see figure 2 below). I argue that both explicit and implicit measures may be involved in predicting/explaining different types of behaviour in what I call a *cross alignment* pattern (figure 3). *Cross alignment* pattern fits the data better than the *parallel alignment* pattern of the dualistic view (figure 2). Both measures may predict/explain behaviour which, it is important to stress, is moderated by various factors that involve situation and individual level variables.

<u>Figure 2</u>

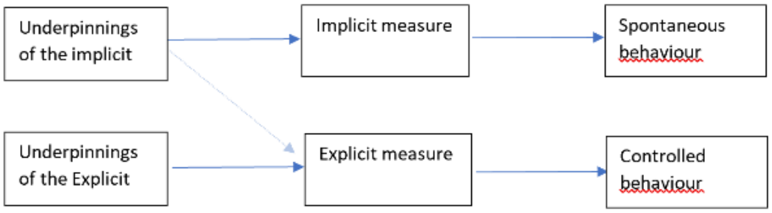

Figure 2: Parallel alignment pattern of relations between mental elements, measurement procedures, and behaviour.

Figure 3

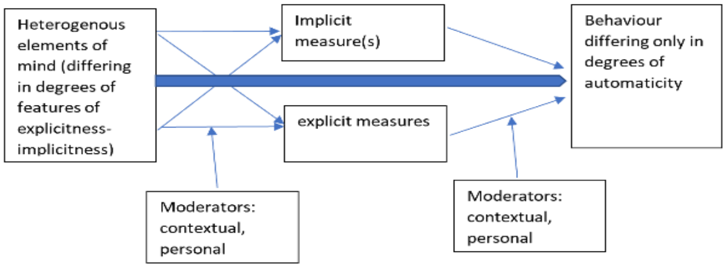
Heterogenous elements of mind (differing in degrees of features of explicitness-implicitness)
Implicit measure(s)
explicit measures
Behaviour differing only in degrees of automaticity
Moderators: contextual, personal
Moderators: contextual, personal

Figure 3: Cross alignment pattern of relations between mental elements, measurement procedures, and behaviour.

Consider studies which have suggested that responses on explicit measures may also be influenced by what is typically thought of as implicit attitudes. Such studies illustrate the cross-alignment pattern. Take for example, Jonathan Haidt's (2001) case of the two consenting adults, a sister and a brother (Julie and Mark), who for the sake of curiosity (and because they have never experienced it before), decide to have consensual sex with one another.[68] They take all precautionary measures of birth control (Julie uses contraceptive pills and Mark uses a condom), and although both enjoy the experience, they decide they won't repeat it, and both agree that the experience is to be a secret they both know has brought them emotionally closer together. The majority of people who are asked whether this behaviour is acceptable, respond with a negative. Their reasons for saying the behaviour is unacceptable (which include that incestuous sex leads to offspring with genetic abnormalities, or that such a relationship leads to emotional problems) are consistently debunked by the facts of the story, but this still does not seem to change their minds. On the basis of this, Haidt argues that it is the subject's implicit emotions, such as implicit repulsions, that are producing people's explicit responses in this case, whereas the explicit beliefs they give, as though arrived at through reasoning, are just *post hoc* justifications for those

[68] I thank Jussi Suikkanen for referring me to this example.

responses. Such justifications are also often discussed in the confabulation literature (see Bortolotti, 2018).

Although Haidt's research was intended to support his social intuitionist model, the results can be taken in favour of the argument for a cross alignment of prediction. Deliberate/explicit responses may be explained/predicted by implicit attitudes. In this case, measures instructing participants about their explicit attitudes (towards the story of Julie and Mark) tracked mental elements (perhaps disgust or intuition) better thought of as captured by implicit measurement devices.

Consider further, research on voting behaviour. Galdi, Arcuri, and Gawronski (2008) and Lundberg and Payne (2014), for example, report similar relations in their findings; namely, that implicit attitudes predict future voting decisions (a behaviour thought to be deliberative, intentional, and captured by explicit measures).[69] These studies report that at high levels of decidedness or confidence in one's vote (when voters know what their explicit attitudes are towards the presidential candidates), explicit measures are good predictors of voting behaviour (i.e. explicit response). But, at low levels of confidence in one's votes (when voters are still undecided who to vote for), both explicit and implicit attitudes are equally predictive of the voting behaviour (i.e. the explicit response). If this is not convincing of a cross alignment pattern of relations as shown in figure 3, think of people who report higher level life satisfaction on rainy days, or those who are kinder to others when they smell fresh bread than when they don't. In both these cases, subjects' 'implicit' moods unknowingly influence their explicit evaluations.

[69] What this means is that (for certain individuals) both measures give better predictions of behaviour than one measure alone.

Taken together, such evidence suggests a challenge for dualistic models, namely that the prediction and explanation of behaviour does not reflect a parallel alignment pattern as described in figure 2. Rather, it follows a cross alignment pattern (figure 3). Implicit attitudes (such as implicit repulsions in Haidt's example) may be involved in responses on explicit measures. Moreover, it suggests that the underpinnings of behaviour are not dissociated (or distinctive binary elements) but vary only in degrees of automaticity. The cross alignment pattern indicates that various mental underpinnings may be captured by a variety of implicit and explicit measures each of which may be implicated in adding to the prediction of behaviour.

To sum up this lengthy section, the challenge of unity against the dualistic hypothesis claims that no characterization of the psychological underpinnings guiding responses on direct and indirect measures may unproblematically mark our cognitions into two distinct explicit and implicit categories. Firstly, claiming an explicit-implicit distinction as reflective of controlled-uncontrolled shows to be unsatisfactory. At least some mental elements considered *implicit* may be within direct or "ecological control" (while at least some mental states typically described as *explicit* may be beyond direct control). Secondly, claiming a distinction reflective of conscious-unconscious does not work. As Gawronski (2019, p. 2) put it "there is no evidence that people are unaware of the mental contents underlying their implicit biases", in many cases, subjects may become aware of the content of their biases. Thirdly, claiming an explicit-implicit distinction on the basis of propositional-associative (rational-arational) underlying processes also does not hold. There is convincing evidence to show that (at least some of) the mental states described as associatively structured and arational (thus taken to be implicit) may actually be responsive to reason and interactive with other mental states in a manner typical of propositional structures. Finally, claiming the explicit-implicit distinction as indicated by a parallel alignment between, on the one hand, direct/explicit measurement devices, what they aim to track, and the behaviour they guide, and on the other hand, indirect/implicit devices, what they tap into, and the actions they influence is unsatisfactory. Direct measures may track states and processes which are considered implicit. In effect, there is no evidence to *clearly* distinguish two uniform and single kinds of mental items underlying behaviour. What the research does show, however, as we will see shortly, is that the relationship between attitude and behaviour is rather nuanced and complex. Keep this in mind as it will come into discussion again in the next chapter (section 6.6.2).

5.2 <u>The Heterogeneity Challenge</u>

In this section, I discuss the second difficulty that the dualistic view faces and it relates to the assumption that implicit attitudes make up a *uniform* mental category. This difficulty is related to what Holroyd and Sweetman (2016) call the *heterogeneity of implicit bias.* I argue that there is reason to think that what the dualist calls *implicit attitudes* is not a unorm category of mental kinds that is distinct from the unified mental category the dualist calls *explicit attitudes*.

My aim, thus, is to question the common treatment of implicit attitudes as a uniform psychological kind. To do this, I first explicate the notion of heterogeneity and I argue that the puzzling data (which shows low correlation between results on different implicit measures) gives reason to suspect that implicit measures tap into a *homogeneous* mental category (5.2.1). I draw on Holroyd and Sweetman's (2016) discussion of heterogeneity of implicit bias (5.2.2) and offer my own claim that what the literature calls *implicit bias* is not only grounded in associations as some authors suggest but may also be grounded in heterogeneous mental elements including beliefs, affects, gut-feelings, and so on (5.2.3).

5.2.1 The notion of heterogeneity

This sub-section explains what the notion of heterogeneity involves. Here, I discuss how heterogeneity is used in reference to the array of different kinds of psychological underpinnings involved in the production of responses on indirect/implicit measures and in the generation of certain micro-level behaviour. The prominent paradigm understands implicit biases as (homogeneously) associative in structure (Madva, 2016; Gendler, 2008; Brownstein, 2016; 2017; Saul, 2013a, b; Frankish, 2016).[70] On this conceptualization, for example, a preference towards a male's CV compared to an (identical) female's CV, the speed of categorization of white over black faces on an IAT, and the excessive interruptions of female students in comparison to male students are all underpinned by the same homogeneous *type* of mental state.

However, I argue (with Holroyd and Sweetman, 2016, p. 84) that subsuming implicit bias under one unified and uniform category encourages the tendency to view the term *implicit bias* as a 'catch-all'. Moreover, describing implicit bias as a uniform category also dismisses the possibility that the responses on indirect measurement procedures pick out a range of social (and often non-social) cognitions and affective states including stereotypes, categorizations, likings, dis-likings, gut-feelings, intuitions, beliefs, and others. Such a broad characterization of implicit bias, although may be effective for some general purposes, nevertheless leads (and may have led) to

[70] For example, Madva, Gendler, Saul, and others consider the structure of the underpinnings of implicit bias as strictly associative. In a similar manner, Mandelbaum (2015) considers the structure to be strictly propositional.
Notably, Mandelbaum (2016) and Schwitzgebel (2010) challenge the associative paradigmatic framework and defend belief or belief-like structure for the underpinnings of implicit bias. Interestingly, Mandelbaum (2016) contends that his commitment to structured beliefs does not "preclude commitments to other entities in the mind's unconscious". He further suggests that these entities might be "free-standing associations" but he insists that such "entities" "do far less causal work than often supposed, especially in the implicit bias literature" (p. 636).

counterproductive normative recommendations and unwarranted generalizations (Holroyd and Sweetman, 2016).[71]

The argument for heterogeneity of implicit bias was introduced by Holroyd and Sweetman (2016) as a critique of the unification of implicit bias under a homogeneous category. The contention is that any functional definition of implicit bias such as Brownstein and Saul's (2016, p. 7) "implicit bias" as "unconscious" and as affecting the way we "perceive, evaluate, or interact" with people from a target group

> leaves open the matter of precisely what processes constitute implicit bias, and in particular whether we are dealing with a singular entity or a range of psychological tendencies (Holroyd and Sweetman, 2016, p. 82).

In other words, defining implicit bias as *unconscious* and/or *uncontrolled* does not give us insight into the exact type of processes that constitute implicit bias and whether these processes are singular or multiple.

To support their heterogeneity argument, Holroyd and Sweetman (2016) draw on the well-known but often neglected anomaly in the literature regarding the lack of correlation between results on different indirect measurement procedures (see also Machery, 2017).[72] Indirect measurement instruments are thought to be interchangeable when it comes to measuring attitudes (Gawronski, 2009) and this assumption feeds into the idea that the underlying psychological bases

[71] Holroyd and Sweetman (2016) argue that broad usage of implicit bias serves (1) to focus attention on mental activity that is not encompassed by reflection and deliberation and its role in our mental lives, and (2) to better formulate the widespread effects of implicit bias including micro-discrimination.

[72] This lack of correlation in the outcomes of different indirect measures (of the same attitudinal object) is an awkward puzzle in the research on attitudes.

of attitude are a uniform and homogeneous kind.[73] Holroyd and Sweetman (2016, p. 87) argue that although measurement error may explain this lack of correlation between different implicit measurements instruments, the more likely account is that these instruments access "discreet and non-unified implicit associations (such that they do not all cluster together to form an implicit attitude), or perhaps different kinds of implicit processes". Thus, the likely explanation according to these authors is that a heterogeneity in the content of the associations or a heterogeneity in the processes underpinning these associations is involved in the responses on the different implicit procedures. As I shall argue shortly, this is a narrower type of heterogeneity than I will be suggesting insofar as Holroyd and Sweetman (2016) take (strictly) different kinds of *associations* to constitute the heterogeneity.

In the next sub-section, I draw further on Holroyd and Sweetman's (2016) account of heterogeneity. Bracketing the various features which characterize the *implicit* and how they differ from those which characterize the *explicit*, Holroyd and Sweetman (2016) suggest that implicit biases are heterogeneous, and this heterogeneity may be described along two dimensions: functional and structural. I elaborate on each of these below.

5.2.2 Two dimensions of heterogeneity

Two possible dimensions of heterogeneity of implicit bias are discussed in Holroyd and Sweetman (2016). These may be crudely taken as follows: (a) a *functional* dimension because it refers to what implicit bias *does*, and (b) a *structural* dimension because it refers to what implicit bias *is*.

[73] Philosophers and psychologists in this field treat implicit biases as a single category, or a single type or sui generis mental state, notable exceptions are Holroyd and Sweetman (2016), Del Pinal and Spaulding (2018), and to a certain extent Sullivan-Bissett (2019).

A) Functional heterogeneity

The first dimension of heterogeneity, then, is *functional*. There are several ways to understand functional heterogeneity and they involve the idea that there are differences in how implicit biases *operate* in relation to other mental states. If the mental elements typically taken to be implicit relate differently to other mental states (like explicit beliefs), then there is reason to suspect heterogeneity in implicit bias (Holroyd and Sweetman, 2016).

One way to illustrate this (as discussed by Holroyd and Sweetman, 2016) is to consider how implicit bias operates in relation to explicit beliefs. We can ask the question: do implicit biases and explicit beliefs systematically relate to each other in a similar manner? In other words, does a change in an individual's explicit beliefs have an influence on some of her implicit biases but not others? There is empirical evidence to suggest that conceptions of homogeneity in explicit-implicit relations are not warranted; indeed, there are substantial varying differences in the relation between explicit attitudes and implicit bias.

On the one hand, some studies reveal no relationship between explicit attitudes and implicit bias, i.e. that a change in explicit attitudes does not lead to a change in implicit bias. For example, Banaji and Hardin (1996) show implicit bias – e.g. associations between pronouns (SHE/HE) and stereotypical roles (NURSE/DOCTOR) – to be strong for both individuals who are highly explicitly sexist as well as those who are not explicitly sexist. On the other hand, other experiments do demonstrate a relationship between explicit attitudes and implicit bias. For example, Devine and her colleagues (2002) show that individuals who prioritize the importance of *anti*-racism – as shown by their explicit non-racist attitudes – manifest implicit *anti*-racism in the scores on their indirect measures. While those who don't believe anti-racism to be important don't show such changes in their implicit bias.

Gawronski and Brannon (2017) also discuss the complicated relationship between explicit and implicit attitudes. They cite research (e.g. Brochu, Gawronski and Esses, 2011; Gawronski et al., 2008) showing high correlation between explicit and implicit attitudes for individuals who are motivated to control their implicit attitude and who perceive that there is high discrimination against a stigmatized group. For example, when a subject is highly motivated to control his racist (implicit) attitude towards Muslim hijabi women, this will show in his anti-racist explicit attitude; his self-reports will be positive towards Muslim hijabi women. Moreover, he may notice that in his town, Muslims are highly discriminated against. So, when he witnesses a hijabi woman on the bus being discriminated against, his implicit bias (insofar as he is motivated to control bias *and* he observes high discrimination against Muslims) will align with his explicit attitude towards hijabi women. In effect, he will likely choose to take the seat next to the discriminated-against hijabi woman. Being attentive to how implicit bias relates to explicit beliefs is especially important with respect to the kind of generalizations often made about "individuals being afflicted by implicit bias irrespective of their explicit beliefs" (Holroyd and Sweetman, 2016, p. 88).

A second way to consider this heterogeneity in the relationship between implicit attitudes and other mental states, suggested by Del Pinal and Spaulding (2018), is to consider implicit attitudes' *dependence* on other mental states. These authors draw a distinction between (1) features (e.g. FEMALE) of a social category (e.g. NURSE) which have *weak* relations of interdependence with other features (i.e. they are *peripheral* to understanding the concept of NURSE) and (2) features (e.g. EMPATHIC) which have *strong* relations of interdependence (i.e. they are *central* to understanding the concept of NURSE).[74] Changing the peripheral features (FEMALE to MALE) has little effect on other features of the concept (NURSE) while changing central features (EMPATHIC) often makes the concept difficult to comprehend. Paradigm measurement procedures of implicit bias (e.g. IAT and priming tasks) typically focus on associations involving peripheral features (like NURSE and FEMALE or NURSE and MALE) because these are often more salient (i.e. prominent and available for the subject). Measuring these types of associations turns out to be unreliable because they don't have strong relations of conceptual dependency. This explains the low correlation between different measures of the implicit association/bias. Central features (e.g. EMPATHIC) of a social category (NURSE) may not be statistically salient in common contexts given by current implicit measures and so they often go untracked, but they are more stable and robust than peripheral features (MALE and FEMALE).

The reason Del Pinal and Spaulding's (2018) discussion is important for my purposes is because it allows for two classes of implicit biases: implicit biases which involve salient (but peripheral) associations (FEMALE and NURSE), and those which involve features that hold strong centrality (EMPATHIC and NURSE). Echoing the main message of their work, they write,

> ….even if we focus just on the class of biases that involve relations between concepts and cognitive features, implicit biases can be encoded in different ways… It follows that we should not assume that this broad class of implicit biases has a uniform underlying nature (Del Pinal and Spaulding, 2018, p. 108).

The important take home message is the authors' conclusion that there is a significant reason to consider heterogeneity in implicit bias.

[74] Implicit biases with weak relations of interdependence are organised in terms of their *salient or statistical associations*, while those with strong relations of interdependence are organised in terms of their *conceptual dependency*.

Finally, and perhaps this is the simplest way to understand functional heterogeneity, consider how implicit bias with different content influences behaviour differently. While certain kinds of associations influence behaviour in one way, other kinds of biases may influence social behaviour in a different way (e.g. the association of BLACKS and MUSICAL influences one's behaviour towards a black man in a different way than the association of BLACKS and DANGER). It's not clear *a priori* why one would consider relata with different contents to behave in homogeneous ways. As Mandelbaum (2016, p. 631) explains, treating implicit biases with different content as sufficiently similar is "de rigueur in the literature" but "it may ultimately be misleading". He cites Nosek and colleagues (2007) to illustrate the lack of correlation in implicit racism across different indirect measurement tasks (such as the scores on IAT and error rates on Payne's Weapon Bias task). What these results suggest, he argues, is that indirect instruments may capture different psychological underpinnings of implicit bias.

To sum up the discussion on functional heterogeneity, there are considerable accumulating arguments for differences in how implicit biases relate to other mental states, to each other, and to behaviour. This makes it difficult for implicit bias to be described as a uniform category guiding behaviour, thus mounting the pressure on dualists who dismiss the evidence against the homogeneity of implicit bias.

B) Structural heterogeneity

A second dimension of heterogeneity in implicit bias is what Holroyd and Sweetman (2016) call *structural heterogeneity*. This dimension involves a diversity in the mental processes underpinning the results on different indirect measures. One way to examine structural heterogeneity is to consider the structure of indirect instruments. Different indirect instruments involve diverse cognitive processes (and produce diverse results) suggesting that there might be a diversity in the implicit biases involved. If this is so, then claiming that implicit bias is a uniform psychological kind is misguided.

Consider, for example, the variation in the structure of the processes underlying the IAT and priming tasks. In the IAT, the cognitive processes involved are those of *categorization*. Respondents to an IAT are required to *categorize* the stimuli presented to them into one of two groups based on their category memberships. The tracked association may be considered as reflective of the association between faces (images of black or white faces) or words (JOY, ANGER, PEACE) and a generic social category (BLACKS, WHITES). The IAT is thus considered to be a measure of "category-based associations" (Gawronski 2009, p. 146). In contrast, in priming tasks, the processes involve an *evaluation of valence*. Here, respondents are not required to categorize pictures or words into a category like in the IAT. Instead, the respondents are subliminally presented with a prime stimulus (say a black or a white face) and required to *assess the valence* of a target picture or a word, making their assessment amenable to the idiosyncratic features of the prime stimulus (the black or white face). In effect, priming tasks require the prime stimulus to be the *exemplar* of the category (as opposed to a generic category) which means that these measures track "exemplar-related associations" (Gawronski, 2009, p. 146).

Category-based associations and exemplar-related associations are – at least *prima facie* – different, suggesting that making any generalizations about the underpinnings of implicit bias as a uniform kind may be misleading. Gawronski (2009) takes this form of heterogeneity in the structure of the processes underlying implicit bias to explain the perplexing data of low correlation between different indirect measures. He argues that the variance in results on different indirect measures might, in large part, be due to these measures tapping into different underlying processes, i.e. to a heterogeneity in the processes underlying implicit bias.

Insofar as indirect measures involve different processes, they may be taken to capture different implicit psychological underpinnings. If this is so, then researchers are advised not to generalize indirect measures under a single and uniform category. Some theorists in the field of implicit bias warn against this. For example, Nosek and his colleagues (2007) promote a nuanced categorization of indirect measurement instruments based on the particular psychological processes they engage. De Houwer (2003) argues that, since different indirect measures employ different mechanisms, the psychological processes involved could also be heterogeneous. And Mandelbaum (2016, p. 631) claims that "the ultimate discussion of implicit bias should discuss the cognitive structures that each test reveals, and these needn't be the same structures".

In summary, the second challenge against the dualistic view, i.e. the heterogeneity challenge, targets the uniformity of the underpinnings of implicit bias. It argues for heterogeneity in implicit bias. There is good reason to think that the underpinnings of implicit bias differ in many ways. They differ in how they relate to other mental states such as beliefs and behaviour, and they differ in terms of the processes and mechanisms they involve.

5.2.3 Beyond a heterogeneity in associations

At this point, I want to expand on the notion of heterogeneity and suggest that what is involved in the results on the different indirect/implicit measures is not only not uniform but also not a uniform *mental kind*: that it is neither exclusively associatively structured, nor is it exclusively propositionally structured. I want to suggest the possibility that a heterogeneity in the underpinnings of the results on indirect measures may not only be reflective of differing functions or structures of these underpinnings, but also reflective of a heterogeneity in the *kinds* of mental elements that are involved in these biases. Let me explain.

Many researchers claim the heterogeneous mental underpinnings of implicit bias are associatively structured (Holroyd and Sweetman, 2016; Gawronski, 2009) or propositionally structured (Mandelbaum, 2016; De Houwer, 2003; Del Pinal and Spaulding, 2018). Given this fact, the heterogeneity of implicit bias is taken to be descriptive of a single psychological *kind.* For example, for many philosophers this psychological *kind* is associatively structured but heterogeneous in its relation to other mental states (functional heterogeneity) or in the structure of its underlying processes (structural heterogeneity) (Holroyd and Sweetman, 2016).[75] Thus, although the current discourse around heterogeneity claims a diversity in function and structure of underlying elements, the *architecture* of implicit bias is typically a unified kind. Implicit bias is described as either associative (e.g. aliefs), propositional (e.g. unconscious beliefs) or *sui generis* kind (e.g. patchy endorsements).

The suggestion I propose, however, (and this is a suggestion which has not before been entertained) is that the heterogeneity of implicit bias may extend beyond the functional and structural heterogeneity discussed above, and towards a heterogeneity in the psychological *kinds* that are involved.[76] Various arguments lend support to the claim that what is being recorded by

[75] Holroyd and Sweetman (2016) question the often-employed framing of associations as semantic vs affective as a main source of heterogeneity in implicit attitudes. My suggestion need not involve such strict dualities. In fact, it is much more complex than that.

[76] There have been recent attempts to address the heterogeneity problem. For example, Sullivan-Bissett (2019) presents a heterogeneous account of implicit bias and argues that implicit biases as unconscious imaginings can be structured associatively and non-associatively. Brownstein (2018, p. 92) also explicitly endorses heterogeneity in implicit bias but he maintains that their “variety differs [only] in degree, not in kind”. Brownstein’s (2018) focus is on the degree and intensity of affect that is part of “all implicit attitudes”, meaning that the heterogeneity of implicit bias stems from a difference in the intensity of affective response that is inherent in every implicit bias. This affective intensity,

indirect/implicit measurement instrument are not a unified *kind* of mental elements (neither exclusively associative nor exclusively propositional).[77] For my purposes, suffice to note that since at least some of our implicit biases are grounded in mental elements considered to be associations while others are grounded in elements with propositional content, then there is reason to question the assumption that implicit bias forms one unified mental *kind* (refer back to 5.1.3). This will come to inform the discussion in the next chapter.

To sum up the heterogeneity challenge against the dualistic view, I presented arguments that implicit bias is not a uniform category; rather, it is heterogeneous. I discussed evidence from empirical work showing that (a) different biases *function* differently in relation to explicit beliefs and behaviour and (b) indirect measures employing different processes track different *structures* underpinning implicit biases. I further suggested a heterogeneity in the psychological *kinds* comprising implicit bias.

5.3 The Complexity Challenge

In the previous section I argued for the claim that implicit attitudes are a heterogeneous psychological kind. In this section, I discuss the complexities involved in explicit attitudes. If explicit attitudes, as I argue, are complex and difficult to measure, it becomes unclear how they are to be distinguished from the heterogeneous implicit attitudes. I call this the complexity challenge

Brownstein explains, is a function of content-specific, person-specific and context-specific variables which serve to moderate implicit biases one way or another. It isn't clear, however, how this addresses the heterogeneity problem. He describes four components of implicit 'attitudes' that he calls FTBA, "a salient Feature in the environment, the experience of bodily Tension, a Behavioral response, a felt sense of Alleviation" (Brownstein, 2018, p. 23). For more refer to his book *The Implicit Mind: Cognitive Architecture, the Self, and Ethics* (2018).

[77] Recall for example Mandelbaum's (2016) convincing arguments that implicit biases are grounded in beliefs. Hahn and his colleagues (2014) have also argued that at least some of our implicit biases are underpinned by, or function as, gut feelings, and Sullivan-Bissett (2019) argues that at least some of our implicit biases are grounded in unconscious imaginings.

and I explain it in two parts. The first part (5.3.1) involves the difficulties that measurement instruments are faced with when tracking a complex concept as explicit attitudes. Here, I illustrate the complexity involved in the notion of explicit attitude which I argue may not be tracked by the uni-dimensional measurement tools currently used. The second part (5.3.2) draws attention to the multiple moderating variables involved in behaviour. I argue that although dualists are likely aware of these complexities, the picture often drawn simplifies the relation between attitude and behaviour and dismisses complex moderating factors.

5.3.1 The complexity of explicit attitudes

I first discuss the complexities involved in how explicit attitudes are understood and the ambiguity in how they are tracked. I argue that if there are complex and multifaceted dimensions to what the literature calls an individual's *explicit attitude*, then it isn't clear what the direct/explicit measures track. Given that a subject's explicit attitude involves multiple aspects of the target object in different contexts, the direct measurement instruments that are currently in use in the research (e.g. surveys and questionnaires) cannot fully capture this complexity.

Consider a target object that many of us can relate to, chocolate cake. Imagine a subject who enjoys eating chocolate cake. Imagine that she also knows sugar is bad for her health especially given that there is recurrent diabetes in her family, and she knows she should be watching her weight given her doctor's recommendation. Meanwhile, eating chocolate cake makes her feel energetic and raises her overall mood. When she is asked on a direct measure what her explicit attitude is towards chocolate cake, what is she to respond? She has many differing (mixed) cognitions about chocolate cake, and given that she is well aware of them, is she expected to average out all these beliefs when she is asked to report her attitude? What would such an average or summary mean? Is she to report the one she considers most favourable to her physical wellbeing? Alternatively, is she to consider her psychological wellbeing as primary? Where is she to commit to on the dimension of her evaluation of 'liking' or 'disliking'? Is she to respond with '*feeling favourable towards'* or *'unfavourable towards'* chocolate cake?

These questions underscore the complexity involved in considering a person's explicit attitude. Effectively, the responses elicited by direct measures presuppose that individuals can simplify, or as Fazio (2007) suggests, can summarize their multifarious 'attitudinal stance' along a single evaluative dimension. A summarized version of this, however, would simply miss out on much information.

Now recall Rachel, the professor who believes herself to be highly racially egalitarian (section 2.6). Rachel believes in the equal intelligence of blacks and whites as her research shows her this fact. She has daily interactions with her black students, she has some black friends and colleagues and at least in part because of this, she considers herself not to be prejudiced against blacks. She professes her egalitarianism and most likely, on survey questions she would score low on racism. It is helpful to closely examine Rachel's very broad and vague belief that:

p: 'Blacks and whites are equally intellectual'.

What does the belief that *p* entail? Does it follow from the belief that 'blacks and whites are equally intellectual' that:

(a) Blacks and whites achieve equal scores on IQ tests?
(b) Black students write academic essays as eloquently as whites?
(c) Black people have the potential for high academic achievement, but social structures impede this?

(d) Black students can ideally write academic essays as eloquently as whites, but they are often faced with systemic impediments such as stereotype threat?
(e) Blacks and whites should have equal standing in academia? Or
(f) Blacks and whites have the same potential for intellectual success, but black people don’t try hard enough?

Which of these different proposals follows from Rachel’s belief that *p*? How is the researcher to tell? Might her belief guarantee multiple contents (a-f)? If so, then the uni-dimensional scoring typical of explicit measures does not accurately capture the multi-dimensional aspect of her belief.

Perhaps Rachel has all these beliefs, which *prima facie* look to be subsumed under the belief *p*, yet on closer inspection they are different insofar as they may lead to different behaviour. Should we think of Rachel as having all these beliefs when we consider her egalitarian explicit attitude? For example, if Rachel believes that (c) ‘Black students have the potential for high academic achievement, but social structures impede this’, then there wouldn’t be surprise at her disposition to be more impressed by a black student’s exceptional paper compared with a white student’s exceptional paper (as her case study in section 2.5 suggests). In the dominant dualistic way of thinking of an individual’s attitude towards social identity groups, it isn’t clear what to make of the belief *p*: ‘Blacks and whites are equally intellectual’, nor is it clear how (a-f) above are to be understood in relation to the belief *p*. In the situation where an explicit attitude is expressed on a unidimensional scale or a self-report (as with the current explicit measures) it simply is not clear how a researcher is to establish what Rachel’s self-report entails, her belief *p* is radically indeterminate.

Indeed, one can see that Rachel has complex mental elements that incorporate her belief in egalitarianism yet are gravely more sophisticated and nuanced than the claim that ‘Blacks and whites are equally intellectual’ or the self-reported state of liking towards black students. If we consider all the nuances of Rachel's attitude towards blacks, then pointing to a gap between her micro-level behaviour in class and her professed belief *p* just doesn't serve to express an accurate picture, neither of Rachel's racism nor of her egalitarianism. In effect, to make a distinction between explicit attitude and implicit bias (based on results on different measures) and to limit their description to evaluations as the literature does, leaves no room for the complexity and heterogeneity that are part of a person's attitude towards a target object.

Nor is the use of the more modern direct measures that employ Likert-type scales any better at capturing the complexity that I mean to stress. If a person's attitude towards an object or social group is complex, as I am suggesting, then it is problematic to assume that it may be measured on a Likert type scale format for the following reasons. Firstly, if explicit attitude is complex, then what do results on a single scale indicate? For quantitative data on Likert scales such as the Modern Racism Scale (MRS) for example to be useful and interpretable, a score of one's explicit attitude should represent a single 'construct' (Carifo and Perla 2007). Specifically, a core assumption underlying measurement scales of the Likert type is that there is some conceptual or empirical underpinning that connects all the items of the scale, and moreover that these items are replications of each other. These are essential for the scale to be interpretable (ibid). Thus, given that explicit attitude is complex and multi-dimensional, a Likert scale will not be a feasible way to measure it. Using such a scale is likely to produce inadequate data.

Secondly, if explicit attitudes are taken as evaluative summaries, what would a summary evaluation look like on a unidimensional scale? This issue concerns the concept of 'attitude' (which I will come back to in section 6.1). The question is, can the different mental elements/states concerning the target object or social group (e.g. blacks) be taken as similarly structured to be subject to summation? Not exactly. It is unclear how this information is only quantitatively different in such a way that the subject may be able to give an evaluative summary as expected by direct measurement instruments. The issue also concerns the measurement scales. For example, what is the researcher to make of the zero point on a Likert scale (or scales ranging from -5 to 5) where zero is considered neutral? The zero point is ambiguous presumably because a subject will respond with a zero when (1) she sums up differing (conflicting) cognitions regarding the questionnaire item *or* when (2) she has a neutral attitude towards the target. So when a subject is instructed to respond to a statement on the MRS ('I would rather not have blacks live in the same apartment building I live in'), she might choose the point closest to zero (on the Likert scale) whether she experiences a sense of ambivalence towards people of colour or she doesn't really care either way. Individuals with very differently distributed scores might have the same summarized score even though they hold very different attitudes towards the target object (Samra, 2014; Carifio and Perla, 2007). This is problematic.

To sum up, as we consider the complexity involved in a person's explicit attitude towards a target object, we can see that there simply is not enough structure on current direct measurement instruments to portray this complexity. It is not something that can be summarized or typified by single dimensional scales. Current explicit measurement tools can only capture a summarized point on a unidimensional linear scale. Add to that the unclarity that comes with understanding what exactly is meant by a *summary* evaluation the individual is instructed to report. Moreover (as we learned from the heterogeneity challenge in 5.2), multi-dimensional psychological kinds are responsible for our judgements and behavioural responses, and (as we learned from the unity challenge in 5.1), the relevant psychological kinds cannot be neatly divided into the explicit and the implicit. What is needed is a clearer and more representative notion of the complexity. I develop a model that accounts for the complexities and the other challenges in the next chapter.

5.3.2 The complexity of factors involved in explicit measures

This is the part where I draw closer attention to the complexities (imparted by extraneous factors) that are involved in the measurement of attitude. I argue that given that numerous variables moderate the measurement results, philosophical accounts such as those that endorse the dualistic alignment hypothesis, which do not fully attend to such complexity, are misguided.

The challenge against the dualistic alignment hypothesis is not that it does not account for the complexities involved in attitude measurement. Granted, dualistic theorists will likely be in accord with the idea of mediating and moderating influences. However, the picture they paint involves questions of moral responsibility of persons who have been found to score as biased on indirect/implicit measurement devices (i.e. who harbour implicit bias). The challenge for those theorists is to explain the *adequacy* of their accounts in light of the complexities of extraneous factors. To be sure, contextual factors are rarely, if ever, discussed in their accounts. Yet these cannot be overlooked if we are to have a satisfactory explanatory theory of bias.

Empirical research has reported robust results suggesting that various factors influence the responses on measurement devices and influence the relationship between these responses. For example, Greenwald and colleagues (2009) point to varying factors influencing the relationship between the responses on direct and indirect measures. Among these factors are:

- *The importance of the attitude* - the more important and relevant the attitude is, the more the responses on explicit measures converge with those on implicit measures (Karpinski, Steinman and Hilton 2005; Nosek, 2005, 2007).[78]
- *The motivation for positive self-presentation* - the more interested the respondent is to present herself in a positive light, the more likely her explicit response will differ from her implicit response (Nosek, 2005, 2007).
- *The spontaneity of response on direct measures* – when spontaneous responses are instigated on direct measures, they are more likely to converge with responses on indirect measures (Hofmann et al., 2005).
- *When there's low motivation and opportunity to engage in deliberate processing* – when the respondent does not have the opportunity or the motivation or is not given the opportunity to deliberate on their explicit response, it most likely will converge with their implicit response (Fazio et al., 1995).
- *Perceived discrimination* – when perceived discrimination is low, the respondent is more likely to respond in a similar way on the direct measure as they would on the indirect measure (Gawronski et al., 2008).

The claim is that individual and contextual factors moderate the relationship between responses on explicit and implicit measures. This reflects the importance of theory in understanding the complexities involved in what measurement devices track, the relationship between the results on these measures, and what constitutes social cognitions in general. It equally highlights the naivety of positing the question 'Do explicit and implicit attitudes diverge?'. As Fazio and Olson (2003) conclude in their review of implicit attitude research:

> We already know enough to be able to say that the question has no simple answer. That is, the answer is 'it depends'… We need to be asking a 'when' question: When, under what conditions, and for what kind of people are implicit and explicit measures related? (Fazio and Olson, 2003, p. 304).

To get a better grip on what I mean by the complexities inherent in delineating explicit from implicit attitudes, reflect on the following questions: if I am opposed to affirmative action, does

[78] Attitude importance refers to the "subjective sense of concern about an attitude" as well as "the psychological significance" that the respondent attaches to it (Karpinski, Steinman and Hilton, 2005, p. 950).

that mean I’m *explicitly* or *implicitly* prejudiced? If I consider the success of Asian-Americans as evidence showing that racism in America is not a variable limiting the progress of African Americans, am I explicitly or implicitly racist? If I am a male CEO of a large company who calls women employees by their first names, but men by their last names, am I considered explicitly or implicitly sexist? Can I be anti-prejudice and support the measures taken by Trump to build a wall at the Mexican border? Can I be egalitarian and support anti-Muslim immigration measures taken by the Trump administration in the US? If I think that Muslims are partly to blame for the bigotry and discrimination they face in the West, am I considered explicitly or implicitly biased? These questions are just meant to illustrate the kinds of intricacies, nuances, and ambiguities surrounding our understanding of what prejudice means and what discrimination towards stigmatized groups involves. To posit a dichotomy between two uniform mental kinds (an explicit and implicit) as an explanatory theory is to dismiss these complexities.

From this brief review, we can see that one needs to be very careful when discussing the nature and causes of responses on measures of bias. Numerous moderating factors could account for variations in the responses. Research needs to resist the simplification and reduction often required by psychological measures, as well as the simplification of the concept of prejudice in general, and explicit prejudice in particular. In the absence of a robust understanding of the complexities of biases such as sexism and racism and the evolving nature of their manifestation, and in the absence of valid and reliable measures that control for various moderators, it isn’t clear what is being measured. And if it isn’t clear what is being measured by explicit and implicit instruments, then drawing a principled distinction between explicit and implicit attitudes just does not hold.

5.4 Conclusion

This chapter confronts philosophers who endorse the dualistic alignment hypothesis with three challenges. Firstly, the unity challenge, contests the principled distinction between the mental underpinnings of explicit and implicit attitudes. The mechanisms underpinning explicit attitudes may be characterized as having features shared by those underpinning (at least some) implicit attitudes, features such as *controllability, introspective awareness* and/or *responsiveness to reason*. Secondly, the heterogeneity challenge, argues that there is good evidence to suggest diversity in the structure of the underlying processes of implicit bias as well as a diversity in the way these cognitive underpinnings relate to other mental states. Lastly, the complexity challenge objects to the treatment of an explicit attitude as a unified phenomenon. Indeed, a cluster of mental states and processes are implicated in the production of responses measured by direct instruments.

These challenges will turn out to be crucial for understanding the notion of ‘attitude’ in social and cognitive psychology because, unlike the dualistic hypothesis, the relationship between heterogeneous psychological underpinnings and behaviour prove to be very nuanced and complex. To the extent that philosophical accounts dismiss the disunity, the heterogeneity, and the complexity involved in social attitudes, they miss out on much of the nuances and intricacies involved in social behaviour and create unnecessary challenges which they face difficulties resolving.

From this point forward, I use the terms ‘implicit’ and ‘explicit’ with caution because as I hope to have demonstrated in this chapter, the distinction between what is ‘explicit’ and what is ‘implicit’ is not clear enough to substantiate a dualism. In the next chapter, I construct a complex picture of bias which is far more heterogeneous than that envisioned by Holroyd and Sweetman (2016) and Del Pinal and Spaulding (2018).

Chapter 6

THE MOSAIC VIEW

A Novel Approach to Understanding Bias

Introduction

In this chapter, I use the three challenges of unity, heterogeneity, and complexity that I put forth against the dualistic view to develop an alternative account of how the mind works around biases and prejudices. The view I advance reconciles these challenges, incorporates various mental states into our conception of bias, and considers a novel approach to understanding bias in which heterogeneous mental elements interact dynamically in various contexts to produce behaviour. I call this the *mosaic view*.

Before I get to elaborating on the mosaic view, I start off the chapter by questioning our understanding of the notion of attitude and its relation to behaviour. It is important to understand that a significant reason for the controversies around implicit bias stems from difficulties in describing what we are measuring. That is why in 6.1, I examine the historical development of the understanding of the notion of 'attitude' in psychology. I show that the concept of 'attitude' transformed over time from a multi-dimensional concept to a uni-dimensional model of evaluation. I argue that this simplification of the notion of 'attitude' and its relation to behaviour has limited the concept's power to represent what it was intended to describe.

In 6.2, I advance my account, the mosaic view, as an alternative theoretical framework of understanding human behaviour. I suggest that underwriting our social behaviour is a complex conglomeration of elements of mind that together form what I label an evaluative *stance* towards the target object.[79] A stance is not the sort of thing that can be measured on a unidimensional scale such as the Likert scale which current direct measurement instruments employ (e.g. the explicit attitude section of the IAT). Treating a stance as though it can be measured in such a way leads not only to mistaken measurement results (or at the very least an incomplete understanding of what people's stances - or what the literature calls *attitudes* - are like), but it also generates a variety of needless puzzles. Sections 6.3 to 6.5 are concerned with expanding on the mosaic view and explaining how it is superior to the dualistic view. Finally, section 6.6 addresses two possible objections to the mosaic view.

6.1 The orthodox notion of 'attitude' in psychology: A brief historical outlook

I start off this section by claiming that simplifying the notion of 'attitude' along uni-dimensional scales is not conceptually driven. Although such simplification may be parsimonious, it does not reflect the reality nor the complexity of our mental lives. To support my claim, I trace the historical development of the notion of 'attitude' in psychology. I explain that an 'attitude' was originally regarded as a multi-dimensional concept (i.e. as involving conceptually different dimensions: affective, behavioural, and cognitive). However, to simplify quantitative analysis, it was reconsidered as an 'evaluative summary' capturable by uni-dimensional scales. I argue that this reconsideration may have paved the way for the dualistic hypothesis to posit homogeneous and static elements of mind responsible for behaviour.

The convention among psychologists in the field of implicit bias is to use the term 'attitude' to designate what Eagly and Chaiken (1993, p. 1) call a "psychological tendency that is expressed

[79] As discussed in chapter 2, there is confusion arising from the use of the term 'attitude' as it has different referents in philosophy and in psychology. Instead, I adopt the term 'stance' to refer to what the dominant view indicates as 'attitude'.

by evaluating a particular entity with some degree of favor or disfavor”; for example ‘I like women’. Such an understanding assumes a unidimensional structure for ‘attitude’ in such a way that it can be measured using a single evaluative dimension ranging from positive (favour) to negative (disfavour). Along similar lines, Fazio (2007, p. 4) understands the term ‘attitude’ as “an evaluative summary” of information about a target object and similarly relies on a unidimensional structure of evaluation.

If we consider an attitude as an evaluative summary capturable by uni-dimensional scales, then it becomes more plausible to consider its nature as uniform and static. Current measurement instruments are based on the assumption that an attitude can be tracked along a single point on a preference scale between two points (i.e. a uni-dimensional scale). This is the kind of data currently available to theorists in the field and it is data which should be questioned if the aim is to fully understand social cognition and behaviour.

It is important to be clear that this kind of thinking about ‘attitudes’ has not always been dominant in psychology. The attitude literature from the 1940s until the late 1960s was under the influence of what may be considered a *tripartite paradigm* (Samra, 2014). The main idea behind this paradigm is that an ‘attitude’ is conceived of as lying along multiple dimensions realized in affective, behavioural, and cognitive states. In effect, the tripartite model maintains that an attitude consists of three classes of information: (1) an affective class relating to how the subject *feels* towards the target object, (2) a behavioural class relating to how the subject *is disposed to behave* towards the object and (3) a cognitive class relating to the subject’s *beliefs* and stereotypes about the target.

These three dimensions were considered *conceptually* different and as such were measured separately using separate scales with different formats (Samra, 2014). For example, the Likert scale may be used to track the *affective* dimension. Here the subject may be instructed to indicate how favourable or unfavourable she feels towards group G (e.g. Muslims). Particular behaviours may be used to track the *behavioural* class of information constituting the attitude. Here the subject may be asked to indicate on a dichotomous scale (agree/disagree) whether she would approach someone from group G (e.g. If I see a woman in hijab on the bus being bullied for her clothing, I would interfere to resolve the situation). Finally, a totally different format using semantic differential scales may be used to measure the *cognitive* dimension. Here the subject may be instructed to rate group G (say Muslims) on certain traits (e.g. hostility, kindness, pleasantness, etc.). The measurement of three different dimensions came out of the understanding that the three components are *qualitatively* and *structurally* different, and the model received statistical validation that the measures tracked *qualitatively* unique elements (Ostrom, 1969). Empirical findings supported the use of this tripartite model as a conceptual model to frame the investigation of *complex* attitudes (Samra, 2014).

According to Samra's (2014) historical account of the definition of attitude in psychology, influential psychologists of that era (e.g. Fishbein, 1967) called for a reinterpretation of this tripartite model in a way to better suit the structure of the measurement instruments and to make for less complex analyses. The reasoning behind the reinterpretation rested on the idea that research using uni-dimensional scales is easier and more rigorous than research using multi-dimensional scales. Here's Fishbein on the matter:

> a conceptual system in which only the affective component is treated as attitudinal and the other two components are linked to beliefs, should permit a more productive approach to the study of attitudes (Fishbein, 1967, p. 257).

The suggested reinterpretation, which later came to be known as the *ABC model of attitudes*, adopted the three dimensions as *quantitatively* different components of a uni-dimensional concept (as opposed to *qualitatively* distinct dimensions).[80] Effectively, the subject's attitude, (say 'women and men are equally competent leaders') although grounded in affective, behavioural, and cognitive components, was reconceptualized so as to be a *generalized evaluative summary* of these three components (Samra, 2014).

A cooking metaphor is helpful here. Think of the three components of the tripartite model (the affective, behavioural, and cognitive components) as three ingredients in a cake, say eggs, flour, and butter. The ABC model's summary evaluation would be analogous to the cake itself which does not resemble any of the ingredients, but it is based on these ingredients. The multiple dimensions of the tripartite model are analogous to the ingredients of the cake, whereas the summary attitude is analogous to the cake itself. The complexities of the cake are not clear from the cake itself, just as the complexities of an attitude are not clear from a summary evaluation.

Since current research typically produces summated scores on a single scale, the evaluative dimension being measured (preference for a target object along a continuum) represents a simplified and uni-dimensional understanding of the attitude which does not reflect the complexity of that attitude (nor the earlier thinking about attitudes). As we have seen from the discussion on the complexity challenge (section 5.3) and as we shall shortly discover, attending to one dimension

[80] This may remind the reader of similar understandings of attitude in the philosophical literature, namely Gendler's (2008a, b) notion of alief as having three components (see chapter 2 for more on Gendler's alief model).

misses valuable information about a person's attitude and reduces the validity of the measurement devices. This is just to say that the uni-dimensionality of the structure of the attitude concept is something that cannot be *a priori* assumed. In fact, what the literature considers as an 'attitude' could well be multi-dimensional.[81]

The point of this historical account is to raise concern around the theoretical conceptualization of the concept of attitudes assumed by the dualistic hypothesis. Linking back to the three challenges of unity, heterogeneity, and complexity raised in the previous chapter, one can see how psychology's historical interest in quantitative research has reduced the ability of the concept of attitude to be a placeholder for a variety of psychological underpinnings and it has limited its measurement to a point on a unidimensional scale. The danger in this is that the notion of what the dualists call *attitudes* as an evaluative preference of likings or dislikings does not reflect the reality of our mental lives nor the complexity involved in our consideration of a given target object. Further, it may result in erroneous theories of prejudice and discriminatory behaviour as those endorsing the dualistic hypothesis. In the following, I discuss a novel approach which I argue describes our psychological reality more accurately than the dualistic alignment hypothesis.

6.2 <u>The mosaic view</u>

If an attitude is not best described as "liking or disliking" or as the psychological disposition to make an evaluative summary statement, then how is it best construed? Here, I suggest a novel and more realistic description of what is traditionally taken to be an 'attitude' and used as the fundamental element of the dualistic alignment hypothesis. I call this novel framework the *mosaic view*. According to the mosaic view, what underpins social behaviour is a complex conglomeration

[81] To be clear, I'm not committing to the three-dimension or tripartite view, insofar as there could very well turn out to be more than just three dimensions.

(or mosaic) of interacting elements of mind within a network I call a *stance*. I expand on this view and give several analogies and an example to illustrate what I mean by a stance. Finally, I argue that the notion of a stance is better suited to describe the complexities involved in our approach towards social categories than are the monolithic phenomena depicted (dualistically) as 'explicit attitudes' and 'implicit attitudes'.

Therefore, from this point forward, I refrain from using the term 'attitude' so as to relieve my suggestion from the baggage that the term holds. Instead I will use the term 'stance'. I begin with the claim that what we refer to as someone's 'attitude' is best understood as their *stance.* By *stance*, I mean to refer to a multi-dimensional mosaic of mental elements with a shared relation relevant to a target object. The notion of stance which I propose is intended to replace orthodox talk of explicit and implicit attitudes.

What is a stance?

When considering an individual's stance, especially with respect to social identity groups, it is important to note that it is not simply that one holds (all things considered) two kinds of *uniform* and *stable* mental elements, whatever their structure may be, i.e. associative, propositional, or something else entirely. Elements of a stance are not isolated static units as such, they are not unconnected discrete elements (explicit attitudes or implicit attitudes) elicited in isolation as responses to an encountered stimulus. This is the type of thinking broadly reflected in the dominant literature, insofar as it erroneously supposes that measurement instruments track two uniform, stable, and distinct explicit and implicit mental kinds. As Holroyd and Sweetman (2016, p. 89) suggest, in the orthodox view "there is a tendency to suppose that implicit biases are unrelated to explicit beliefs”.[82] However, my preferred view, the mosaic view, not only avoids such a tendency, but it goes further to explicate the relations between various elements of mind that constitute a stance and underpin behaviour in a way that is nuanced and dramatically more complex than orthodoxy supposes.

A stance involves a complex conglomeration of mental elements – perhaps associations, beliefs, gut-feelings, emotions, aliefs, unconscious imaginings, and so on – which are brought together in virtue of the role they play in an agent’s psychology; i.e. their role with respect to the target object.[83] Here, I do not mean to claim that this list of mental elements is exact or exhaustive. These psychological elements, in fact, will include some of the same elements any proponent of the dualistic view considers. It is simply that, according to the mosaic view, these elements are numerous in kind and arranged differently than they are under the dualistic framework (i.e. they are dynamic and distributed along a continuum, rather than composing a principled duality with features of explicitness and implicitness).

This requires some fleshing out, so let me explain. A subject’s stance towards a target object encompasses many different mental elements that are *about* or *concern* that target object. A fundamental part of what makes some mental elements that are about a target object part of a stance is their functional role: namely, that they all *can*, when activated, take part in motivating behaviour

[82] At least when it comes to philosophical discourse on implicit bias, this assumption is often taken at face value and not critically questioned.

[83] If the reader does not agree with some of these elements (e.g. unconscious imaginings), this is fine, my view is not wedded to any one element of mind. But if these elements do exist, they’re the kind of thing which would be constituents of a stance.

towards the relevant target object.[84] These elements are dynamic units which interact with each other and influence each other reciprocally.

The heterogeneous elements of mind constituting a stance, however, cannot be understood in isolation, rather they can only be understood as elements within a stance. There's a level of explanation which would be missing if the focus is only on the individual elements of mind. That is because at the level of a stance, there are patterns of psychological and behavioural facts which we miss if we only looked at the individual elements constituting the stance (Dennett, 1969). A stance thus is not identical to any of the elements which constitute it, nor to any given set of these elements.

Describing the psychological underpinnings of bias in this way, as belonging to a network of dynamic elements (i.e. a stance), denies the notion of 'attitude' as a *summary* evaluation. What underpins social behaviour is not something that can be captured with an averaged number score or a rating of warm or cold, nor by any of the current measurement instruments that are available to us. To the extent that a stance encompasses heterogeneous mental elements interacting together and underpinning social behaviour, it resists being captured in such a way.

In varying contexts, elements of a stance (on their own, in interaction with other elements in a stance, or in combination with various background beliefs, values, commitments, and other elements outside the stance) produce social behaviour towards a target group. Whether this social behaviour is discriminatory or non-discriminatory depends in part on the elements of the stance

[84] Recall that a target object ranges from objects such as food, clothing, people, to other matters such as social and political ideas, events, activities, and groups. Thus, the mental elements which form the conglomeration that is a person's stance towards the target object are such that their content is involved in motivating some form of behaviour towards the target object.

that are activated and on their dynamic interaction. But it also depends on personal background mental elements not belonging to the stance such as goals, motivations, values, desires, and commitments (I call these *background elements* of mind) as well as on the context which serves to activate the elements. (I will have more to say on how the mosaic view explains behaviour in the next section, but for now please refer to figure 4 below for a visual illustration of the mosaic view.)

Figure 4

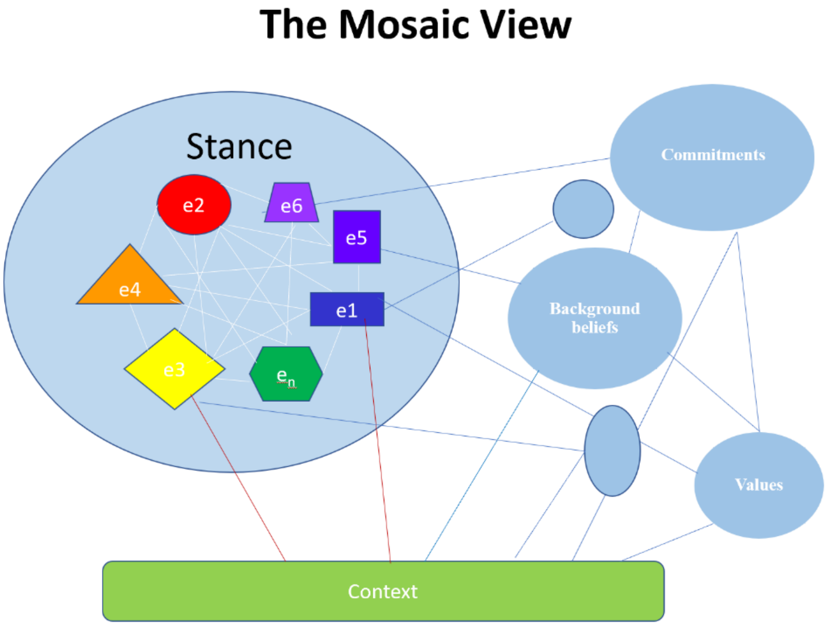

Figure 4: Illustration of the mosaic view.

A stance is constituted by heterogeneous, non-uniform, and dynamic mental elements (e1, e2, e3, en…). For example, a subject's stance towards women constitutes elements like e1: the belief that women have strong leadership qualities, e2: the association WOMEN and WEAK, e3: the gut-feeling of disgust towards women, and so on. The elements constituting the stance (e1, e2, e$_{n…}$) interact with one another, with various generic commitments that the subject holds (e.g. 'the commitment to justice and equality'), background beliefs (e.g. 'gender does not define strong leaders'), and values (e.g. 'I value treating everyone equally'), and all of these are activated differently depending on context and interaction with each other.

I use two helpful analogies to illustrate. Consider first an analogy with the colour spectrum that makes up a rainbow. The rainbow here is analogous to a stance, just as the different colours of the rainbow are analogous to the mental elements that constitute a stance. The colour green is not by itself the rainbow, but it is part of the rainbow. What binds the different colours of the rainbow together is the fact that they have certain effects in our visual field: i.e. their wavelengths of light are visible to the naked eye. Similarly, what binds the diverse constellation of mental elements together is that they (alone, in combination, or in interaction with other mental elements) motivate behaviour towards the relevant social target. What any given measurement instrument tracks is only one element within a stance or one colour of the rainbow (more on what measurement instruments track in section 6.4).

A second analogy from football may capture the idea further.[85] Here, a stance constituted by heterogeneous mental elements is analogous to a football team constituted by individual players. To understand a football game, one does not observe each individual player in isolation. Individual players cannot be understood when they play except as elements within a team playing football. Accordingly, team level facts are essential to the nature of individual players on the team just as stance level facts are essential to the nature of the psychological elements constituting the stance. If we just talk about individual players without discussing teams, we would be missing some important phenomena that are really in the world when it comes to football; patterns that are real but are constituted out of heterogeneous realizers (Dennett, 1969).

What I mean to stress is that to understand the game of football, it is important to understand (1) the football team, (2) its standing among other teams, (3) the playing characteristics of each individual player on the team *as part of that team*, (4) how the players interact with each other and with players on opposing teams, and (5) how the players interact with the coach, the referee, the crowd, the cheering, and so on and so forth. In an analogous way, understanding behaviour towards a social group (e.g. women) involves understanding one's stance, the individual mental elements

[85] Thanks to Scott Sturgeon for suggesting this useful analogy and for helping the mosaic view take shape.

within that stance, the interaction of these elements with each other, and their interaction with background mental elements (commitments, values, etc.) (see figure 4). Context plays a role in understanding social behaviour just as it does in understanding how the game is played. For example, a team may play better on its home field, when crowds are loud, or during the last half when under pressure, it may play better during finals than during preliminaries. The point here is that there is much complexity involved.

I illustrate the mosaic view with the example of Rachel, the philosophy professor discussed in section 2.6. Rachel exhibits various behaviours towards black people in different contexts. She manifests microaggressions towards her black students while conveying genuine egalitarian concerns towards blacks. On the mosaic view, she harbours heterogeneous psychological elements which are about or concern the group ‘black people’, some of which are captured by implicit measures, others by explicit measures, but it’s not the case that all the constituents of her stance are (or can be) captured by our current measures. Let me consider (by stipulation) what *some* of those mental elements constitutive of her stance involve:

e1. An *evaluation* of blacks as a traditionally disempowered group in the US.
e2. An *association* between BLACKS and LOW INTELLIGENCE.
e3. A *categorization* of black people as a mysterious, exotic and interesting group (reflecting aspects of exoticizing and fetishizing).
e4. A *gut feeling of disgust* at seeing black individuals in positions of power (in political posts e.g. Barack Obama and Ilhan Omar, or in high academic positions).
e5. A *fear* that blacks are trying to seize power in America.
e6. *Unconscious imaginings* of a black person as incompetent.
e7. A *belief* that blacks are just like any other group of people, some of whom are good/smart, others bad/not so smart.
e8. An *association* of BLACKS and ATHLETICISM.
e9. An *association* of BLACKS and RYTHMIC SKILLS.
e10. A *belief* that blacks are intellectually equivalent to whites.

and so on (e_n). Note that what these elements have in common, i.e. what makes them part of Rachel's stance is that they are all relevant to the target group 'blacks' and they can be involved in motivating judgement and behaviour towards this group in one way or another.

The mosaic view suggests that these heterogeneous psychological elements constitute Rachel's stance towards blacks insofar as these elements can lead to forming evaluations of and producing discriminatory or non-discriminatory behaviour towards black people. Any one, or a combination, of these elements does not *represent* Rachel's stance towards blacks, rather it reflects only a *partial* view of her stance besides other elements which, taken together, form a mosaic picture that is her stance. Effectively, each element of mind within this constellation is just one element within the mosaic that is Rachel's stance, but it is not identical to her stance (just like any one football player does not represent, nor is identical to, the football team). As discussed above, there are patterns of behaviour which we miss if we look only at the individual elements of Rachel's complicated stance towards blacks, just as there are patterns of behaviour which we miss if we only looked at individual team players.

Importantly, the elements of a stance may not cluster along the ends of an explicit-implicit dichotomy (or along its two polar sides), each of which captured by explicit or implicit measurement instruments. Rather, the elements of a stance may be situated anywhere along a continuum of properties (typically attributed to implicit attitudes or explicit attitudes by the dualists) at any given moment in any given context. But these elements do not reliably cluster to make up a stable polarized dichotomy.[86] Recall that when I talk of properties typically attributed to

[86] I use the term 'reliably' here to acknowledge that I'm not denying that there could, in principle, be cases where the mental states cluster around the explicit and the implicit dichotomy, but to elevate this mere possibility to a principled theory as with the accounts endorsing the dualistic hypothesis is unwarranted.

explicit and implicit attitudes, I mean features of automaticity, mainly *control* and *introspective awareness*. Given the dependence on various factors (background beliefs and context), the elements of a stance may show notoriously unstable characteristics on the dimensions of *control* and *introspective awareness*. Unlike the rather static dualistic alignment framework where distinct mental states cluster around two separate poles with differing properties (the explicit side having features of *control*, *introspective accessibility*, and *intentionality* and the implicit side having features of *lack of control*, *introspective inaccessibility*, and *spontaneity*), the mosaic view posits that having these properties is not an all-or-none characteristic, but they can come in a matter of degrees along a continuum.

To varying degrees, some of the elements of a stance may have any (or a combination) of these functional properties which have been typically used by the dualistic view as earmarks to describe the explicit side of the explanatory duality. Other elements may have any (or a combination) of properties typically taken by the dualistic view to describe the implicit side of the explanatory duality. Nonetheless, there are many elements of mind which lie somewhere on the continuum in the large area between the two polar sides of the duality. Most importantly, there is no reliable cluster of mental elements that may be thought of as explicit and as explaining a full range of homogeneous macro-level behaviours and there is not another cluster that may reliably be thought of as implicit and as explaining a full range of homogeneous micro-level behaviours. An analogy might help make things clearer. Think of the North and South Pole as the explicit and implicit poles of the attitude dichotomy. While hardly any people live in the North and the South Poles, life exists mostly somewhere in between. Similarly, as the mosaic view conceives it, few mental elements occupy the explicit and the implicit poles, while most can be considered in the large area in between these poles.[87]

It remains crucial that wherever the psychological element is located on this continuum of control/introspective awareness/intentionality, it is not a stable property of this element: for example, one is not always and in all emotional states and contexts unable to control their biases (refer to 5.1.1). In other words, the degree of control (or introspective awareness) is highly contingent on factors that are external and imposed by the context, as well as factors internal to the subject. Consider from the list of psychological elements of mind above, the disorienting emotion or aversion: (e4) ‘the gut-reaction of aesthetic disgust at seeing a black person in public office’. This gut-reaction, as I am suggesting, is expressed in degrees along a continuum for the various functional properties which don’t always come together. The gut-feeling may lie anywhere on each of the dimensions of control/introspective awareness/etc. So, when Rachel is watching a speech by Barack Obama, this gut-feeling may be introspectively accessible to her, and she may to a certain extent be able to control its influence on her decisions, thus she continues to listen to Obama’s speech. Where this element lies on the continuum of properties depends on various factors including the conditions of the target (e.g. idiosyncratic features of the target object), the subject (e.g. Rachel’s mood), and the context (e.g. when Obama is discussing the enhancement of health care measures, a topic of importance to Rachel). In this manner, over time and context, the properties of the elements may change, and the elements themselves may also change, but the stance remains relatively constant, just as the team players may change but the team itself is relatively stable.

[87] Thanks to Ema Sullivan-Bissett for this helpful analogy.

Different contextual factors activate different elements of mind for the same individual in differing intensity, especially if the individual's elements of a stance have conflicting content as is the case with Rachel, the biased egalitarian. By intensity, I mean to describe how well-established a belief is, how strong a gut-feeling, how well-learned and automatic an association, and so on.[88] (This is not out of line with traditional dualistic alignment thinking, see Higgins and Bargh, 1987 for an example). This is not to say that there is no stability in a stance. While individual differences lie in the differences of activation, intensity, and interaction of the various elements, what holds relatively constant is the stance itself. What this means is that there are elements of mind which cluster one way or another and which change arrangement over time and context. How any one element fires in any given context can itself change depending on various personal and contextual factors (experience, situation, history, etc.), but the stance itself remains relatively constant.

To summarize, a stance is a conglomeration of heterogeneous mental elements responsible in part for our behaviour towards social groups. Its elements are dynamic and non-uniform and they do not reliably cluster around the explicit-implicit poles, rather, they exhibit features of control and introspective accessibility in varying degrees along a continuum and depending on personal and contextual factors. A stance interacts with background elements of mind such as commitments and values at any given context to produce behaviour. Being constituted by heterogeneous mental elements featuring different degrees along the continua characteristic of explicitness and implicitness, a stance resists measurement by current uni-dimensional instruments.

[88] What I mean by 'strong' and 'well-established' psychological elements is how likely these elements are to significantly influence processing and behaviour.

As I conclude the discussion of the mosaic view, I note that the notion of a stance is meant to replace talk of explicit and implicit attitudes (it is in this sense that my view could be broadly described as eliminativist, something I turn to in the next chapter). Talk of a range of homogeneous explicit or implicit psychological elements (within a stance) which explain a wide range of behavioural phenomena misses out on the explanatory role of the stance in behaviour. This is analogous to discussing how football is played by making a distinction between team members who are on the defence and those who are on the offence. This distinction may hold, but as an analysis of how football is played, it misses out on much of the nuances needed to understand the game. In what follows, I explicate how the mosaic view explains behaviour.

6.3 **How does the mosaic view explain behaviour?**

The mosaic view posits that a stance with its various components interacts in a complex manner with various factors, including contextual and personal factors, to produce behaviour. Before I expand on this, I first describe what I mean by *background elements of mind* as these will prove to be essential in my account of behaviour production.

Behaviour, according to the mosaic view, is produced by way of interaction of activated mental elements within a stance, like (e1-e10) above, and background mental states. By background states I mean such states as:

b1. The *belief* that everyone in society, regardless of race should have equal opportunities to education, career, advancement in life, etc…

b2. The *commitment* to support and empower individuals who belong to disempowered groups.

b3. The *motivation/desire* to behave ethically according to societal norms, and to one's morals and beliefs.

b4. The *value* of cultural diversity and the affirmation of cultural identity.

b5. The *rationalization* that low intellectual performance of people who grow up in unjust circumstances is predicted and justified.

These background states, like the constituents of a stance, are differentially activated in virtue of context. Behaviour is the result of a rich system of interrelations between various elements of mind which guide and constrain the activation of other elements of mind whether these are the constituents of a stance or background mental states.

The interrelations show the level of complexity involved such that:

i) The mental elements of a stance vary in their degree of intensity (automaticity and strength).
ii) There are various possible patterns of relations between these elements.
iii) The activation of the elements depends on the situation and on other background states (only the activated elements are functionally important for behaviour).
iv) Certain patterns of activation become sustained via environmental and personal feedback. The more a set of elements is activated (in combination), the faster the activation becomes.

Elements of a stance such as (e2) ‘the *association* of BLACKS and LOW INTELLIGENCE’ may to a certain degree be outside of introspective awareness, and their activation may also be uncontrolled. Their activation, however, depends on the interrelations such elements have with other mental states within a stance, as well as with personal and contextual factors.

How does the mosaic view explain differences in behaviour within and across subjects? To reiterate, the mosaic view considers more than just individual elements of mind for explanation. Given that the heterogeneous elements of mind work as part of a stance, within a network of interrelations with other elements, they are considered within a system.[89] This is analogous to understanding a football game within a network of a range of factors besides the individual players. A mosaic approach to understanding football games involves the analysis of the individual players

[89] By considering the mosaic view as a system, I follow Michael Esfeld’s (1998) argument that elements which are part of a system have some of their characteristic properties only within a whole system. “With respect to the instantiation of these properties each of these things is dependent on there being other things together with which it constitutes a whole of the kind in question” (Esfeld, 1998, p. 366).

as elements within a team and as influenced in their playing by facts about that team, as well as by external factors such as the referee, the crowd, the weather, etc.

The mosaic view understands the difference in behaviour *within* an individual (including the behavioural responses of direct and indirect measures) by appeal to the different elements of a stance and their interaction with background mental states outside of the stance. When Rachel, the biased egalitarian, requires more credentials from a black candidate for research opportunities than from a white candidate, it is not because she harbours an implicit bias, e.g. (e2) certain underlying associations of ‘BLACKS and LOW INTELLIGENCE’ and an explicit attitude, e.g. (e10) ‘a belief that blacks are intellectually equivalent to whites’, and/or any other element of mind, no matter how structurally or functionally heterogeneous. Rather, it is that these elements of mind are differently activated in certain situational conditions, and it is that their activation restrains and enhances the activation of other mental elements within the mosaic that is a stance.

To wit, the activation of the elements of the stance may inhibit the activation of, for example, certain background beliefs or values (if Rachel harbours them) such as (b2) the commitment to support and empower individuals who belong to disempowered groups. Thus, the activation of elements (e2) and (e10) may concurrently trigger some affects such as (e4) ‘gut feeling of disgust’ or (e5) ‘fear that *blacks are trying to seize power in America*’. This interaction feeds back into the processing system to reactivate the elements (e2) and (e10), thereby strengthening them, but also inhibits the activation of (b2). Keep in mind that this pattern of activation and inhibition of the various psychological states is contingent on the situation. For example, it changes when the situation is such that there is no external monitoring on Rachel’s behaviour, when she is not required to justify her decisions to an audience, or when she is not held accountable for any unfair decisions she makes (again, refer back to figure 4 for a graphic illustration of this interaction).[90] Indeed, the cognitive processing involves a very complex and sophisticated network of interrelations between psychological elements, the constituents of a stance as well as background beliefs, values, and commitments. That's how the mosaic view explains the differences in behaviour (including results on measurement instruments) within subjects.

Across subjects, the difference in behaviour similarly depends on the complex interactions between the stance, the background mental states, and the context. Consider the difference between Rachel and an (imaginary) purely egalitarian person. The egalitarian subject, according to my view will have more of the egalitarian background beliefs, values, and commitments than Rachel, even if he has the same exact underlying mosaic of mental elements. The egalitarian subject's background mental states will interact with his activated elements of the mosaic to produce different behaviours. For instance, element (e2) of the stance 'the association of BLACKS and LOW INTELLIGENCE' may be activated in both Rachel and the egalitarian if both subjects score as biased on an implicit measure. However, (e2) in its position in the network of activated mental elements will likely be involved in the production of micro-aggressive behaviour in Rachel but not in the egalitarian. The reason this is so is because the egalitarian, unlike Rachel, has certain moral beliefs and values such as (b1) 'the value in equal opportunities' or (b3) 'the desire for impartiality' which interact with the association (e2) to produce a different behaviour. Rachel either lacks many of these psychological elements (b1-b5) or they are not well-established, whereas the egalitarian

[90] For more on empirical research showing that these techniques motivate subjects to make fair decisions see Lerner and Tetlock (1999, pp. 256-258, 263) and Quinn and Schenker (2002).

has many more of those elements, and they are strong enough to inhibit the association or to activate egalitarian elements within the stance.

What I mean to stress here is that there aren't two relatively independent and homogeneous – explicit and implicit mental kinds – that explain Rachel's wide-ranging behavioural responses. Rachel's purported *implicit* bias (like her egalitarian *explicit* belief) is just another aspect of her overall stance. Her responses on the measurement instruments (be they direct or indirect) may be as multifaceted and complex as the elements of her stance, for the very reason that they reflect the elements constituting her stance. In some situations and contexts (e.g. immediately after walking out of a dark-lit parking garage) her response may exhibit one element of her stance (e.g. fear of blacks); in others (e.g. after watching the movie *Hidden Figures* – a movie about black women who worked in NASA), her response may exhibit another aspect of her stance (e.g. grading objectively her black female students). The important thing to keep in mind is that when it comes to the psychological bases of behaviour, there is much complexity to fathom, and as a result of this we should question the dualistic architectural structure of the mind that dualistic theories presuppose and aim to legitimate empirically.

Finally, on the mosaic view, the explanation of the behaviour of a biased egalitarian (e.g. Rachel) rests on the interaction between the various activated psychological elements within a stance and the contextual variables pertaining to the person and her environment. Behaviour is the result of a rich system of interrelations between the various elements of mind within a stance as they guide and constrain the activation of other elements of mind (including background mental states). While it is important to understand the functioning of individual elements of mind, it is their interrelations within a larger network that is at issue for the mosaic view.

6.4 What do explicit and implicit measures track?

I here explore how the mosaic view explains what is tracked by direct and indirect measurement tools. I argue that the dissonance found in the results tracked by these measures is better explained by the mosaic view than it is by the orthodox dualistic explanatory framework.

Consider again the rainbow analogy. If a certain device or filter is used to track the colours of the rainbow (say an orange filter), what is captured by this filter is only one part of the rainbow (the orange colour) and any claim that considers a rainbow as orange is clearly an incomplete and mistaken claim. Even if a more delicate orange/yellow filter is used, it remains mistaken to claim that a rainbow is orange/yellow. To fully capture a rainbow in all its wavelengths, one not only needs lenses to track the seven colours but one may also need more advanced devices to track wavelengths of light that are not captured by the naked eye, such as ultraviolet rays. Analogously, it may turn out that some elements of a stance, like associations, beliefs, or gut-feelings (i.e. like colours of the rainbow) are capturable by current devices like those covered in chapter 4 akin to the way certain wavelengths of colour are captured by the naked eye. But it may also turn out that some constituents of a stance (perhaps unconscious imaginings) are not trackable by our current devices and require different measurement instruments. This is analogous to certain infrared wavelengths of a rainbow which need specialized devices to capture them.

Similarly, current measures employed by social psychologists may capture any one of several mental states depending on various factors including those imposed by the measurement procedures as well as the situation, the emotional status of the subject, and so on. The reason why these different measurement devices show different and heterogeneous results (as discussed in section 5.2.2) is because they capture different pieces of the mosaic conglomeration of mental elements at different times and in different situations. This means that with our current techniques, what is being measured is a snapshot image of a complex and dynamic mosaic picture that is the *stance*. For example, the IAT tracks only one element of mind – an association – that, depending on context, may be activated with varying degrees of controllability, intentionality, awareness, and so on. Similarly, a priming task, an affect misattribution procedure (AMP), a survey question, and other measurement devices would, according to my view, track only one psychological element of an individual's stance towards a target object (at any one time and in any one context) but none of these elements is exclusive in guiding any kind of behaviour. Rather it is the interaction of the activated elements with background elements and situational factors that explains behaviour. The narrow focus on distinct psychological elements guiding action (as with the dualistic hypothesis) misses the mark and provides an inadequate understanding of human behaviour.

Even if we grant that our current measurement devices have good validity at capturing something systematic about the individual's responses, and hence about the elements of mind concerning the target object, any one of the measurement devices captures what may be only one element of the stance or a combination of more than one. Yet none of the current measurement tools captures what may form a homogeneous 'implicit' category distinct from a homogeneous 'explicit' category, nor do they capture the full mosaic of psychological elements that is the *stance*. Moreover, even if in principle, measures existed that could capture every isolated constituent element of a stance, the combination of the results would not make up the stance. Just like understanding the individual players of a football team would not capture how the team plays. There are facts about the team which cannot be understood by examining its individual members in isolation. In effect, any of the current measurement devices tracks an incomplete picture of a person's stance towards a target object. In line with work on connectionist networks, the constituents of a stance are characterizable by reference to their relation to other states within a larger network, as well as to environmental conditions (Mischell and Shoda, 1995).[91]

Thus, talk of direct/explicit measurement devices capturing explicit attitudes and indirect/implicit devices capturing implicit attitudes has no room in a fully accurate theory of mind. The mosaic view I present does not characterize states piecemeal, rather they are characterized within a system. As such, the mosaic view rejects the explicit-implicit dichotomy; it has it that what the different measurement tools track are only partial snapshots of the heterogeneous mental elements implicated in complex ways in behaviour. (I will say more on how the mosaic view is eliminativist when it comes to implicit bias in the next chapter).

6.5 The mosaic view explains the case studies from chapter 2

How does my view fare at explaining the three case studies of the professors introduced in section 2.6? In this section, I address this question and argue that the mosaic view fares better at explaining the behaviour of the three cases, namely the case of Barb the bigot (i.e. the racist), Sam the self-justifying sexist (i.e. the protocol observing sexist) and Rachel the receptive racist (i.e. the biased egalitarian).

Before I elaborate on how these cases are best explained by the mosaic view, it is worthwhile assessing the manner in which such cases are presented by the dominant dualistic alignment framework. It is important for the reader to understand that the vignettes as presented in the literature are decontextualized snapshots; they are constructed and presented in such a manner

[91] Mischel and Shoda (1995) discuss is detail their connectionist associative model of personality, the cognitive-affective personality system, CAPS which might be helpful in understanding the mosaic view.

as to support the dualistic view. This is to say that typically two psychological kinds (an explicit attitude and an implicit attitude) with differing properties of controllability, introspective awareness, intentionality, and so on, are presented as producing different kinds of behaviour: macro and micro-behaviours. Such vignettes are what I call *decontextualized snippets* of the lives of these agents, and they are meant to establish basic and simple ways of describing these agents' position or stance towards complex social entities such as those involving race and sex.

I first consider the case of Barb who, according to dualists, has what Holroyd (2016) might call '*harmonious kinds of explicit and implicit prejudices*' towards black people given that both her overt and her covert behaviours show prejudice towards black people. According to the dualistic thinking, such as Holroyd's (2015; 2016, p. 159), it is Barb's explicit bias which generates her macro-behaviours; i.e., it is her outright bigotry and hostility towards blacks which leads her, for example, to refuse to accept any black person as her colleague and to share her hatred of black people openly. Moreover, Holroyd spares no effort to convince us that we must pay attention to the fact that it is not Barb's *explicit* attitude that is responsible for her micro-aggressions towards blacks (e.g. her discomfort and irritability in the presence of black people). Rather, it is a distinct psychological kind, Barb's *implicit* attitude, which is responsible for this purportedly distinct kind of behaviour.

My view, however, rejects this dichotomy of kind in the psychological underpinnings of behaviour. Given that a stance is a mosaic of heterogeneous mental elements, the mosaic view considers Barb's mosaic with respect to black people as heterogeneous in kind but harmonious in content. In other words, the heterogeneous elements captured by various direct and indirect measurement devices, i.e. her beliefs, motivations, desires, habits, associations, affects, imaginings, and so on, align in their prejudicial content towards blacks. Moreover, Barb lacks many of the background beliefs considered as egalitarian and just. And thus, any which way these elements of mind interact, they produce prejudiced behaviour, some of which is epistemically easier to track, like overt discriminatory behaviour, while other behaviour is epistemically more difficult to track, like the micro behaviours discussed in chapter 2. Barb's stance is stable, and it indicates hostility towards blacks.

Second, I consider the case of Sam, the ethics professor, who is strongly motivated to adhere to norms of equality of the sexes. The dualistic view considers Sam to harbour implicit attitudes which generate such behaviours as hiring more men than women in his department in addition to the protocol-conforming explicit beliefs which reflect his reported motivation for egalitarianism. Sam is motivated to consider himself egalitarian when it comes to gender. According to the dualist, then, Sam's implicit attitudes lead him to behave in subtle sexist ways. For the sake of simplicity, I take the dualist to endorse implicit attitudes as associations, and to include strong associations between WOMEN and DOMESTICITY and weak associations between WOMEN and CAREER. At the same time, the dualist assumes that Sam adheres to the protocol of egalitarianism. This is expressed by his reported explicit attitude which leads him to behave in what may be considered overt egalitarian ways.

The mosaic view teaches us that this description of a very complex person is unsatisfactory. In its place, the mosaic view attempts to elaborate on Sam's case. Insofar as Sam is a protocol adhering sexist, he is motivated by protocol not to act in sexist ways, and since he is a professor of ethics, he follows what he considers egalitarian ways which he does for the sake of the ethical importance of egalitarianism. As such, he applies what he considers egalitarian ways to his personal life and in his parenting activities. For example, Sam does not try to enforce gender norms on his toddler whom he dresses in gender neutral colours, buys her trucks and trains as toys. He allows his older male child to play with dolls. At home, he will do his share of the house chores; for example, he often washes the dishes and takes the kids to the playground. He also considers practicing egalitarianism at work where he considers that showing benevolence towards women is not a way to show that they are equal to men. He deliberately tries to give as much attention to his female students as his male students. In effect, he genuinely considers himself to be egalitarian when it comes to gender because he follows what *he considers* the 'protocol' of treating women and men equally.

As part of this protocol, or so he believes, the right moral act of egalitarianism is to choose randomly between an identical male and identical female CV for a professorial job in engineering where a woman has never previously held such a post. He does not consider affirmative action (which could potentially improve women's representation in STEM fields particularly in the engineering faculty) to be the right move because he believes that this gives women an unwarranted advantage over men. He believes, along with his strict desire to observe egalitarian protocols, that men and women should have equality, but no gender should have an advantage over another. He also thinks that if a woman feels intimidated in an all-male panel of judges for a competition in computer software, then she should not be on the panel to begin with. If she feels intimidated by the lack of representation of women in the field, then her emotions will interfere with her judgments, and thus, she just should not be there.

The point here is that the complexity is not just a complexity of Sam's stance. Granted, Sam's stance towards women is relatively stable (and negative), but the complexity also involves other mental states which don’t exactly fit Sam’s stance but involve notions of egalitarianism and sexism (e.g. Sam’s egalitarianism means that both sexes are equal in all respects and women should get no opportunities that are not given to men; i.e. he considers that the playing field is currently levelled for both sexes, and there should be no affirmative action, quotas, or such measures that would ensure equal representation in various fields). On close inspection, there are many self-professed egalitarians who espouse such narratives and consider them indications of egalitarianism, but if my argument for complexity is right, then it is clear that such considerations are not to be judged rashly or in a simple manner. It is likely that our understanding of egalitarianism (as much as our understanding of prejudice) is in need of deeper analysis and wider discussion.

Finally, I consider Rachel whose egalitarianism reflects her liberal stance on issues such as race. Rachel’s case is the most puzzling and it presents the typical unexplained phenomenon that dominates the literature on implicit bias. Section 2.6 covers some of the particular versions of the dualistic stories explaining the biased egalitarian’s behaviour like Rachel’s, and this gives reason for the focus I have been giving to this particular case. Recall that Rachel’s scientific research motivates her conviction of racial equality in intelligence and her avowal of this belief is purportedly explained by her explicit belief of egalitarianism. Her puzzling subtle behaviour, i.e. her microaggressions towards black students, however, is thought to be explained by unconscious, uncontrolled, and unintentional implicit bias towards blacks.

According to the dualist, Rachel holds two types of “attitudes”: an explicit attitude and a conflicting implicit bias. Unconscious, uncontrolled psychological elements underpin Rachel’s implicit bias towards black people, and it is this implicit bias that is responsible for Rachel’s microaggressions towards her black students in class. Whereas her explicit attitude is egalitarian insofar as she professes and is convinced that there is absolutely no evidence to suggest social differences in intelligence. Her explicit belief does not explain her biased micro-behaviours towards her black students, whereas her implicit bias towards blacks does.

Again, the vignette used by dualists to describe Rachel is dramatically and problematically simplistic in its portrayal of the biased egalitarian. This is because the vignette presumes a dualistic paradigm. Here, I suggest a more sophisticated account of what might be happening in Rachel's case. Rachel is someone who, like a considerable segment of the population, including perhaps the reader of this work, disavows prejudice and common social stereotypes and opposes explicit discrimination. Moreover, Rachel aims to be unbiased in all her actions. Yet, she is susceptible to the kinds of biased behaviour that concern implicit bias researchers, namely the ones I have been calling *microaggressions*.

The mosaic view takes Rachel's heterogeneous mental states (those relevant to her judgement of black people) as involved in typifying her stance towards blacks but does not take distinct mental states to be manifested in distinct behaviours. Importantly, a summary of these psychological elements in question may not give a proper understanding of Rachel's attitude towards blacks, it is not clear what it would mean to summarize various affects, beliefs, motives, intentions, etc. into one overall characterization of Rachel's take on black people. Neither can a group or combination of these psychological elements share properties which make them cluster along two poles indicating an implicit or an explicit kind of attitude. According to the mosaic view, Rachel's stance is complicated. It is contextually interactive and includes a mosaic of differing psychological composites that are activated at different times in different situations. Rachel may well have affective states along with beliefs (even contradictory beliefs) and various other mental states, all of which may contribute – interactively – in one way or another to form her complicated stance towards blacks. The different elements of the stance integrate and interact in a complex manner to produce behaviour (both micro and macro) depending on contextual and individual factors.

6.6 Objections to the mosaic view

I discuss two possible objections from dualists against the mosaic view. (Although I'm focusing mainly on the dualist as my opponent, I also mean to include other potential objectors). The first objection is that the mosaic view is not illuminative. The second objection is that there is nothing in the mosaic view that makes it superior to the dualistic view. I outline and respond to these objections below.

6.6.1 Objection 1 – The mosaic view is uninformative

The first worry is simply that the mosaic view is not illuminating of human behaviour. My opponent might agree with the notion of a stance but still object that it tells us anything new or instructive about puzzling human behaviour. If all the mosaic view posits is that an individual has heterogeneous mental elements which in some way or another produce behaviour, it doesn't get us very far in terms of an explanation of that behaviour. In effect, the mosaic view does not do much to help us make sense of the phenomenon of microaggression that we started out trying to explain, nor with the puzzle of what allows for or facilitates continued prejudice and discrimination in the face of reduced explicit bias. The objection continues that while both the dualistic approach and the mosaic approach agree that our mental lives are complicated, the dualistic approach configures this complexity in terms of a *dual alignment* of explanation: explicit attitudes inform beliefs and guide macro-level actions, and implicit attitudes influence micro-level behaviour. The mosaic view on the other hand, just points to the complexity without arranging it in a way which could support explanations of behaviour. The dualistic view is thus informative about our behaviour whereas the mosaic view is not.

My response here takes two steps. In the first step I offer some unusual cases that do not fit the orthodox dualistic view of implicit bias and in the second step, I argue that whatever explanation of these cases the dualist offers, the mosaic view does better. The unusual cases I have in mind are what I call *atypical cases*. Imagine, for instance, contexts in which characteristic properties of the explicit (e.g. control, introspective awareness, and so on) or the implicit (e.g. lack of control, lack of introspective awareness, etc.) dissociate from one another. An *atypical* context includes ambiguous situations as the following two cases:

Atypical case 1:

> A subject is given limited time to respond on a questionnaire or survey; in this case the subject is aware of the contents of their mental states (introspective awareness) but the time for responding is limited (lack of control).

Atypical case 2:

> A subject is given ample time to respond to an IAT; in such a case, it is unclear what psychological states the subject is aware of, but it is clear that the response involves control and reflectiveness.

It isn't clear how the dualistic view aligns the mental elements and the behaviour in such cases. Does the questionnaire in the first case capture an explicit attitude (since it is measured by a self-report) or an implicit attitude (since it involves lack of control)? The same with the second case. Does it reflect the subject's implicit attitude (since the mode of measurement is the IAT) or an explicit attitude (since the subject has ample time to deliberate on the response)? Perhaps my objector is able to explain such atypical contexts and fit them within a dichotomous framework. But before I get to the objector's possible response, let me explore other atypical contexts.

Artificial laboratory settings, case studies employed in research, and dualistic discussions on bias are shaped in such a way as to isolate distinct biases from each other. The sexist is often presented as just a sexist, without reference to his take on other social groups, e.g. people of colour, the elderly, Muslims, Asians, gay men, people with disabilities, and so on (suggesting a lack of understanding the person in a more comprehensive manner). Effectively, this isolation of bias is an artificial representation of real cases and it does not reflect real life scenarios. The vast literature on the intersectionality of axes of oppression reveals to us that claims (such as what dualists would call 'attitudes') cannot always be examined and understood as resulting from discrete sources of bias (Kimberle Crenshaw, 1989). Imagine, therefore, the following complex situation

Atypical case 3:

> A subject is required to choose between two equally suitable candidates to fill a job vacancy. The competing candidates happen to be a young black man and an older white woman (or perhaps a white gay disabled man and an Asian heterosexual abled woman).

Such atypical cases involve more than one dimension of bias, and the multiple intersectional and interactive mental elements need consideration. My objector will have to explain behaviour in such atypical cases.

One possible response may be to claim that both kinds of mental elements, the explicit and the implicit attitudes explain behaviour in some way or another in the atypical cases. Such a response will have two consequences. The first is that it makes the objection against the mosaic view (i.e. that it is uninformative) not a very good objection. If rival views explain behaviour in atypical cases as the result of all explicit and all implicit elements of mind, and by the idea that these together explain behaviour in some way, then by the same reasoning, their explanation *also* doesn't get us very far. Moreover, an appeal (by the dualist) to the distinction between explicit and implicit attitudes will just not help to explain what the agent does in such situations. Consequently, rival views could be accused of being equally uninformative. The second consequence is that if my opponent goes further to give some account of how explicit and implicit attitudes produce behaviour, I'll use the same accounts they use, and my view will be as explanatory as theirs. Whatever account of behaviour they use to address this objection, the mosaic view will do the same.[92]

The opponent of the mosaic view may further claim that their view does indeed specify the influence of complex real-life factors (including other mental states and environmental factors) in simpler terms by spelling out the usual 'other things being equal' clause. By doing that, the analysis of the phenomena becomes easier. Such a move, however, turns out to be inadequate for two reasons. Firstly, while the dualistic view may marginally reference some role to other mental states, or use 'the all else being equal' clause, its focus remains on homogeneous and uniform elements of mind. And so even if my opponent gives a minor role to the influence of context and varying psychological states on behaviour, that role is insubstantial and bracketed. Secondly, an analysis which isolates important factors within a network of mental states and external or environmental factors for the sake of simplicity is fragmentary, inadequate, and faces many challenges.

To be sure, the objector may insist that the mosaic view is not illuminative because it is 'too broad' and complex as an explanation. While I disagree that the mosaic view is not illuminative, I

[92] The mosaic view is (as orthodox views) wedded to the idea that mental states cause behaviour, but it is not wedded to any detailed account. For example, the dualist may adopt a dominant account of action to explain behaviour, and the mosaic view would be compatible with these dominant accounts of action. At least two accounts are available: the standard causal theory of action CTA or the interpretive theory of action. (For more see Ruben, 2003, pp. 77-155) If either is adopted, it is equally compatible with the mosaic view.

concede that, with its core notion of a stance, it is indeed complex. But it is not *impossibly* complex. There is ample theoretical work in areas of personality and cognitive science which are similarly complex, but which are nonetheless very illuminative. (The Cognitive-Affective Personality System, CAPS model advanced by Mischell and Shoda's (1995) is one example of a complex connectionist associative model of personality). Moreover, the mosaic view reflects real life situations such as atypical cases, whereas orthodox views do not. Although my opponent will engage with less complex homogeneous kinds of elements of mind such as explicit and implicit attitudes, her explanation misses out much of the complexity involved in a stance and in behaviour in various contexts. That is why dualistic accounts, for example, are faced with contradicting data and lack of repeatability of findings. It is hard for dualistic models to account for the subtle and hugely complex ways in which we, in fact, function. The mosaic view does better with that.

In summary, the mosaic view offers an explanation of behaviour that is complex and that involves sophisticated interactive relations of mental states and environmental factors. Effectively, rather than setting aside the influence of varying mental states and context, it builds them into the description of the causal role of a stance. Any causal role that a stance plays in behaviour depends indeed on these varying factors.[93] Yet, such an approach does not make the analysis un-illuminative; it just makes it complex. And as it happens, it is a *matter of fact* that our mental lives are themselves complex and involve a multitude of states and processes active in complicated cases and contexts. This is something we cannot dismiss for the sake of simplifying the discussion.

[93] Among contextual variables are time limits, the physical context or environment and among internal variables are the agent's cognitive capacity, her mood, her motivation to control prejudice, her working memory capacity and others (Perugini, Richetin, and Zogmaister, 2010).

6.6.2 Objection 2 – The mosaic view is not superior to dualistic view

The second objection takes the following form. The dualist can dismiss the claim that the mosaic view is in any way superior because she can claim her view will have no problem accommodating the challenges of unity, heterogeneity, and complexity. The notions of explicit and implicit attitudes, she can add, may still be thought of as distinct clusters of heterogeneous and complex elements without discarding talk of explicit and implicit. Even with the heterogeneity and complexity of the mental underpinnings of behaviour, these will nevertheless still cluster into two distinct categories: an explicit and an implicit. Such a view, the dualist can add, remains a far better explanation of the data than the mosaic view. The objection then is that there is no theoretical gain in positing a stance. Even with disunity, heterogeneity, and complexity, the dualist can persist in claiming that the elements cluster around an explicit-implicit pole with various elements being anywhere along the continuum. Effectively, the dualist can claim that heterogeneous mental elements within a stance cluster along two poles of the explicit-implicit dichotomy.

I have three responses to this objection. Firstly, it is not clear that the relevant mental elements reflect a cluster along explicit and implicit sides of the dichotomy. In any case, this is an empirical question. What the evidence currently suggests, however, (and this has been discussed at length in section 5.1), is that the elements of mind are rather *non-systematically dispersed* along a spectrum between the explicit and the implicit and there are no meaningful clusters as the dualistic alignment hypothesis would hope for. It is helpful to bring to mind the North Pole-South Pole analogy from section 6.2. While some elements may lie closer towards the explicit end of the spectrum (just as some humans live in the North Pole) and other elements may lie at the implicit end (few humans live in the South Pole), most of the mental states involved in social behaviour lie somewhere in between (just as most human beings are dispersed in the area between the North and

South Poles). Given that the explicit-implicit distinction is controversial, i.e. given that there is no principled divide between what is explicit and what is implicit (refer to 5.1), describing the heterogeneous mental states underpinning behaviour as either explicit or implicit is inadequate, if not outright false.

Secondly, the data also suggest that the psychological elements considered responsible for behaviour are *dynamic*: they change across time and across different contexts. Dynamic elements do not cluster systematically along the explicit and the implicit. The mental elements captured by direct and indirect measurement instruments make up only a part of the heterogeneous elements of mind one harbours towards a social group. Moreover, these elements are not stable across time and contexts, and there is much to consider in the results than a purportedly summarized version of a subject's so-called attitude.

Thirdly, there's a level of explanation which would be missing if individual elements of mind are examined in isolation rather than within a system that is a stance. Even supposing that elements of mind constituting a stance happen to cluster into the explicit and the implicit, it isn't the case that macro-level behaviours are those explained by the explicit and micro-level behaviours are those explained by the implicit. Again, it is the interaction of individual elements (just like individual team players) within a stance (a football team) that instructs the explanatory framework.

Indeed, empirical research methodologies desperately require that a complex notion, such as one's *take* on a target object (even if it is chocolate cake) be simplified in order to make it amenable to empirical testing (section 6.1). This kind of analysis has contributed drastically to a simplistic and dichotomous manner of thinking about racism, sexism, and all manners of prejudice.

An agent is considered either on the extreme side of being prejudiced – i.e. she is a bigot – or she is on the extreme side of egalitarianism – i.e. she is an egalitarian – or she is the puzzle in between.

If dualists argue that this somewhere in-between makes up a straightforward dichotomy between two conflicting explicit non-prejudiced and implicit prejudiced mental states, each of which constitutes a homogeneous category, then they may be misinterpreting the data. An agent's avowed singular statement about her 'explicit attitudes' towards a stigmatized target group does not tell us much about the complexity involved in her 'attitude, neither does a response on an indirect measurement procedure. These don't tell us much about the area in-between the bigot and the egalitarian.[94] And if the dualists go on to develop models to help them decide whether persons who score high on implicit racism, for example, ought to be considered racists (see for example Levy, 2017), then they are unjustifiably simplifying the complexities. Because as we have seen, this area 'in-between' is a multi-shaded grey that is much more complex than dualistic theories make it out to be.

To sum up, the dominant way of thinking about intergroup bias is rather unsophisticated and inadequate. Mostly, it leaves no room to formulate the complexity I have discussed. If we are convinced by this argument, then we are left (at the very least) sceptical about dualistic explanatory theories. As philosophers in this field, we need to overcome the 'all-or-none' notions in our understanding of prejudice, and instead of asking 'Am I racist?', or questioning the intent of one's egalitarianism, we ought to realize that many of us lie in the grey area somewhere in-between.

[94] It might be the case that dualistic alignment theorists will agree to these arguments, but insofar as their work and the vignettes and case studies they present involve a simplification and a dualism of explicit and implicit attitudes, they are complicit in the oversimplification.

6.7 Conclusion

The mosaic view is a novel and alternative account of understanding bias. It explains what happens with the biased egalitarian who exhibits microaggressions (the phenomenon in need of explanation at the outset of this thesis). The explanation of the mosaic view incorporates the notion of a stance that is meant to replace talk of explicit and implicit attitudes. A stance is a conglomeration of heterogeneous mental elements responsible in part for our behaviour towards social groups. Its elements are dynamic and non-uniform and they do not reliably cluster around the explicit-implicit. Rather, they exhibit features of control and introspective awareness in varying degrees along a continuum depending on personal and contextual factors. Being constituted by heterogeneous elements of mind which feature different degrees along the continua characteristic of explicitness and implicitness, a stance resists measurement by current uni-dimensional instruments. A stance interacts with background elements of mind such as commitments and values at any given context to produce behaviour. Thus, the mosaic view rejects the explicit-implicit dichotomy; it has it that what the different measurement tools track are only partial snapshots of the heterogeneous mental elements implicated in behaviour. In effect, the mosaic view considers behaviour as the product of interactions between the various activated psychological elements constituting a stance and contextual factors pertaining to the person and her environment.

Chapter 7

CONCLUSIONS AND IMPLICATIONS

Summary of thesis/Conclusion

In this thesis, I have argued that orthodox philosophical and psychological accounts of social attitudes and bias endorse a dualistic alignment hypothesis (what I called the *dualistic view*). The dualistic view commits to various types of dualisms which align in a parallel manner. Implicit attitudes or implicit bias is taken to explain the responses on indirect measurement tasks (e.g. IAT and priming tasks) and the behaviours they influence (e.g. microaggressions). Explicit attitudes are said to explain the responses on direct measurement tasks (e.g. questionnaires) and the behaviours they guide (section 2.4). At its core, the dualistic view posits distinctively unified homogeneous explanatory states that are responsible for subtle modes of behaviour such as microaggressions (section 2.5).

I suggested that the instigation of dualistic thinking about attitudes was not data driven but relied on various factors contingent to science. I argued that the wide acceptance of the dualistic view came to be in virtue of its modelling of dual-process theories. The acceptance of dualism rests on the idea that the explicit-implicit dichotomy reflects a controlled-uncontrolled and a conscious-unconscious dichotomy (sections 3.3 and 3.4). But as I argued, such thinking was unjustified given that no empirical evidence to support it (section 3.5).

I further highlighted the psychometric limitations underscoring the empirical research into bias and suggested that the dualistic framework delineating distinct mental elements responsible for our behaviour towards social groups relies on a divergence in the results on two questionable types of measurement instruments. I examined alternative explanations for this divergence and suggested that these include measurement error (section 4.2) and a lack of structural fit between the two types of measures (section 4.1) on which the edifice of the dualistic view is constructed.

After raising empirical scepticism around the adoption of the dualistic view, I raised conceptual scepticism. I presented the dualistic view with three challenges: the unity challenge showed that carving our social cognitions into two principled categories of explicit attitudes and implicit biases is not possible (section 5.1). The heterogeneity challenge illustrated the non-uniformity in the structure and function of what is being tracked by both measurement instruments (section 5.2). And finally, the complexity challenge exposed the cognitive intricacies involved in the expression of behaviours such as microaggressions (section 5.3).

With the concerns that the dualistic view is an inadequate explanatory framework, in chapter six, I advanced an alternative framework, the mosaic view, in which I argued that dichotomizing attitudes into the explicit and the implicit is not the right way to think about social cognition and bias. Rather, what the dualistic view considers as attitudes are best thought of as constitutive of a *stance* which holds a mosaic of different kinds of mental elements (section 6.2). Some of these elements may have characteristics typically considered as uncontrolled/unconscious, and some may have characteristics described as controlled/conscious, but these features are not as stable as the dualistic view presents them. The activation and the characterization of these mental elements depend largely on a network of connecting elements within the stance whose activation is subject to contextual and personal factors and whose manifestation in behaviour is equally determined by the interaction of these variables.

The mosaic view, I argued, is better suited to respond to the challenges as well as to the psychometric limitations implicated in the empirical work. According to the mosaic view, predicting discriminatory behaviour using current measurement techniques is difficult because the cognitive underpinnings of behaviour are multifaceted and involve the interaction of various elements of mind and various personal and contextual factors (sections 6.3- 6.6).

One of my central conclusions is that relying on any one type of measurement instrument such as the IAT or self-reports is like relying on a snapshot photo (or a series of snapshots) to understand an animated movie. Current measurement devices capture transient elements of mind at a given time, in a given situation, rather than something stable about the mind. In other words, attitude measurement techniques track what's on a person's mind at a particular place and time, and what may be needed to address discrimination are advanced theoretical understandings of bias, how it functions, and how its cognitive underpinnings may be tracked.

Developing hypotheses to test the mosaic view empirically is not as simple as developing hypotheses to test the dualistic models. Clearly, testing how one kind of mental element (implicit mental state) produces one kind of behaviour (either an overt discriminatory act or a microaggression) is more precise and specific. But it is inadequate if we wanted an accurate picture of mind. Future work will require focus on the interaction between the various elements of mind and the various moderating and mediating variables that play a central role in producing behaviour. This fact of complexity further suggests that theoretical effort is needed just as much as empirical work.

Another central conclusion of my work is that thinking about biases and their underpinnings within a dichotomous framework results in a misconception of the notion of bias, an undermining of the harm that micro-types of discriminations cause, and directs the focus of researchers away from understanding human behaviour as part of a complex interaction of various individual, social, and contextual phenomena towards thinking in categorical terms. An underlying theme throughout has been towards a paradigm shift in our conception of bias.

In the remainder of this concluding chapter, I examine the ontological implications of rejecting the dualistic model in favour of the mosaic view. In 7.1, I look at the ontological status of the mental elements labelled as *explicit attitudes* and *implicit bias* if we are to reject the family of theories which posit them. I ask the question: if the dualistic alignment hypothesis turns out to be a seriously mistaken theory, and if it turns out that there are not two distinct, unified, and homogeneous sets of psychological explainers responsible for guiding two different kinds of behaviour, then what do we make of *explicit* and *implicit attitudes*? In 7.2, I briefly discuss the practical implications of the mosaic view and I suggest how it may guide future empirical research on how discriminatory behaviour may be mitigated.

7.1 Ontological implications of the mosaic view

If the dualistic view is rejected in favour of a mosaic view, and if heterogeneous mental states and processes with no precise unifying feature are responsible for the responses on direct and indirect measures, then this raises a worry for the ontology of implicit bias as a distinct category in our mental life. The ontological status of implicit bias is the concern of this section. To address it, I draw on Stephen Stich's (1996) discussion of eliminative materialism, the rationality of ontological inference, and the possible strategies one may take in settling ontological questions. I show that adopting the mosaic view implies being realist about implicit bias in the ordinary common sense of the term but eliminativist about implicit bias in the theoretical sense described by the orthodox view.

Stich asks: “How do we settle questions about the existence of things spoken of in theories that we no longer take to be correct?” (Stich, 1996, p. 4). The response Stich gives is that there are a few strategies available for philosophers looking to understand the ontological status of entities posited by theories which turn out to be mistaken. One of these strategies is *revisionism*, the other is *eliminativism*.

I will not spend much time discussing *revisionism* because the mosaic view is not revisionist about implicit bias. An example of a revisionist account comes from Payne, Vuletich, and Lundberg (2017a) who claim to reconcile the puzzling evidence facing orthodox notions of implicit bias as the product of individual mental underpinnings. Payne and his colleagues (2017a) propose a reconceptualization of *implicit bias* as a “social phenomenon that passes through the minds of individuals but exists with greater stability in the situations they inhabit” (Payne, Vuletich, and Lundberg, 2017a, p. 236). Implicit bias according to this revised conception is a property of the person’s environment as opposed to belonging to that person’s mind. Implicit bias as such does not describe unconscious, uncontrolled, unintentional bias *residing* with some degree of permanence in the minds of individual persons. Rather, it reflects the momentary accessibility of certain mental concepts which are shaped by the level of prejudice in the person’s environment (for example, the common stereotype found in American society that blacks are dangerous). For a detailed understanding of this revisionist view of implicit bias, I refer the reader to *The Bias of Crowds hypothesis* introduced by Payne and his colleagues (2017a).

Eliminativism is Stich’s second strategy. In principle, a theory which denies the existence of some posited ‘thing’ is eliminativist about that ‘thing’ (in our case, the ‘thing’ a theory would be eliminativist about is the psychological elements *explicit* and *implicit attitudes*). In science, posits are alluded to by a certain scientific theory for prediction, explanation, and description. If this scientific theory turns out to be mistaken, then these posits, just like “phlogiston”, “caloric fluid”, or “the gods of ancient religions” would be considered as “the fictional posits of a badly mistaken theory” (Stich, 1996, p. 1). If my arguments are right and the dualist alignment hypothesis turns out to be a mistaken explanatory theory, then what will become of the posits *explicit attitude* and *implicit bias*? Would they be considered “fictional posits” whose ontological existence as two distinct mental categories within our mental architecture is false?

In order to answer this question, let me first consider the ordinary claim:

There are tables.

I want to suggest that there are two ways to understand this claim. (There may be more ways, but here I want to consider only two salient ways). I call these two ways the *light-touch reading* and the *heavy-duty reading*.[95] On the light-touch reading, the claim *there are tables* is read as:

There are table-y objects that occupy regions of space.

This commits the light-touch reading to the existence of certain things that occupy regions of space. Nonetheless, it is not a hugely detailed claim. That is why it is light-touch. The heavy-duty reading on the other hand claims what the light-touch reading does, that:

There are table-y things which occupy regions of space.

But it goes on to claim that

> *the table-y things occupy regions of space by occupying every point that makes up that region of space.*

[95] Thanks to Scott Sturgeon for this helpful way of thinking about the ontological implications.

The heavy-duty reading entails the light-touch reading but not the other way around. If the heavy-duty reading is true, the light-touch reading is also true, but not the other way around. The light-touch reading may be true without the heavy-duty reading being true. If tables can occupy regions of space without occupying every point, then the light-touch reading is true but not the heavy-duty reading.

Now consider that we learn from science that on a micro-level, space is almost completely empty of matter, that atoms consist of (mostly) space. The interesting ontological question becomes: are there tables or not? If space is almost completely empty, should we claim that there are no tables? To answer this question, I return to the distinction I made between the two types of readings. If we take the light-touch reading, it is perfectly consistent with the common-sense claim that there are tables. Science doesn't rule that out. But on the heavy-duty reading in which we assume that an entity occupies a region of space by occupying every point in that space, the answer is that there are no tables.

Now let's return to the ontological status of implicit biases and explicit attitudes. Are there implicit biases (and explicit attitudes) or not? Again, I make use of the distinction I introduced above to suggest that the notion of implicit bias can be heard in one of two ways: a light-touch reading and a heavy-touch reading. On the light-touch reading implicit bias is just a bias which in some presently salient way, involves implicit elements of mind. It is not a very detailed description of what implicit describes. Rather, it is whatever common-sense ordinary people have in mind when they talk about implicit bias (simply that implicit bias is involved in the marginalization and discrimination against certain social groups). Given how often we hear about implicit bias on the news, in magazine articles, at companies who work on changing policies to avoid implicit bias, etc., it has become so absorbed into our everyday lives. This is what I mean by the light-touch reading.

The heavy-duty reading, however, involves filling out the light-touch reading with the dualistic alignment hypothesis in some specific way. Generally, it involves the idea that there is a range of elements responsible for social behaviour, some of which are explicit, and some are implicit. The explicit involve properties like *conscious awareness, control, deliberation*, etc., and the implicit involve properties like *lack of conscious awareness, lack of control*, and *spontaneous processes*, etc. On this type of reading, implicit bias describes a uniform kind of element (or range of uniform static elements) of the kind found in some detailed account that endorses the dualistic view (see 2.5). It also commits to a heavy-duty set of claims which I've argued against in my thesis.

Are there implicit biases? When read in the light-touch way, then yes, there are implicit biases. But when read in the heavy-duty way implicit biases don't exist. This makes my ontological stance (like my view) slightly complex. I am realist about implicit bias when read in a light-touch way, i.e. in the ordinary common-sense understanding. But I am eliminativist about implicit bias when read in a heavy-duty way as the posit of orthodox theories which endorse the dualistic view. As a realist about implicit bias, I accept that there are implicit biases in the ordinary sense spoken of on the news or in magazine articles. As an eliminativist, I consider the posits of orthodoxy, namely *explicit attitudes* and *implicit bias* as non-existent. There is no dimension of our human psychology that corresponds to these psychological posits. I both reject the notion of implicit bias as a unified, homogeneous category of mind which features the distinctive characteristics of being introspectively inaccessible, uncontrolled, arational, and reject the notion of explicit attitudes as a unified homogeneous category of mind which features characteristics of being introspectively accessible, controlled, and rational. And although my view has it that there are mental elements constituting a stance, some of which may be unconscious, others conscious, some uncontrolled, others controlled, some spontaneously activated, others not, and so on and so forth, these do not cluster into two reliably distinct categories.

The mosaic view takes biases, prejudices, and stereotypes as part of an individual's cognitive architecture. But it does not frame these biases into the explicit and the implicit as does the dualistic view. The qualifiers *explicit* and *implicit* just do not apply to a stance, nor do they distinguish the elements of a stance as such for two reasons. Firstly, the features of automaticity are not stable features of these elements as they vary in degree depending on various factors. Secondly, the elements of a stance do not cluster reliably onto an explicit and implicit bipolar continuum although individually some of the elements may be said to be either automatic or deliberative. On the heavy-duty reading, there is no *implicit racism* and *implicit sexism* (or explicit racism and explicit sexism), there is only racism that manifests in different behaviours. In that sense, my view is eliminativist about the posits *explicit attitudes* and *implicit bias* as such. The orthodox literature tends to concern heavy-duty mental states and so, the mosaic view is eliminativist about its underlying psychology. Nevertheless, I can happily talk about implicit biases so long as they are heard in the light-touch way.

7.2 <u>Practical implications</u>

If, as the mosaic view proposes, the ontological status of implicit bias is such that these psychological posits don't exist, how does the mosaic view propose to mitigate microaggressions (considered by the dualistic view to be the effects of implicit bias)? I finish off by briefly sketching the practical implications of the mosaic view in light of its ontology. Working this out in detail, however, is a subject for future work.

As I have been arguing, the psychological underpinnings of microaggressions are not dichotomous psychological kinds. The mosaic view provides an explanation for microaggressions and biases without the need for positing two separate clusters of mental elements, be they mental states or mental processes. But if this is so, what mental elements ought to be targeted and what strategies and methods are to be used in our attempt to mitigate bias?

Current methods aimed at mitigating what the dualistic view describes as *implicit bias* (i.e. debiasing methods) are largely ineffective when long term changes are sought after. A meta-analysis of intervention strategies conducted by Calvin Lai and his colleagues in 2016, examined the effects of 17 interventions on ‘implicit’ racial preferences (see also Lai and his colleagues 2014 meta-analysis for similar results). Interventions included such strategies as counter-stereotyping training (or the envisioning or imagining of exemplars that are noncongruent with common stereotypes) and engaging in deliberative processing (or implementation intentions) (see section 5.1.1 for more on ways to control implicit bias). The results showed that only some strategies were effective at reducing preferences measured by the IAT, but the effects were short lived. None of the methods used induced any sustained changes/reduction in implicit bias after some time delay. The common explanation for the failed long-term interventive effects refers back to the tenacity of implicit biases: that because of their characteristic nature, these biases are just very difficult to eliminate or change.

Perhaps, the most important results of this and other meta-analyses, however, were the null findings. Not only did the debiasing interventions fail to *sustainably* reduce implicit bias, they also failed to affect any long-term reduction in explicit bias. Moreover, they failed to affect change in one's support for affirmative action, or in the motivation to respond without prejudice. In effect, none of the interventions in the meta-analyses had any influence on what I called the *background mental states*, including beliefs, commitments, desires, values, and others.

My speculation as to why the interventions are ineffective is that the standard model for interventions targets only isolated cognitive mechanisms, for example, associative mental states. The aim of such interventive techniques as Amodio and Mendoza (2010, p. 362) affirm is "to change the underlying associations that form the basis of implicit bias". The target is strictly the associations at hand, e.g. BLACKS and DANGER. For example, debiasing techniques include strategies to eliminate the association BLACKS and DANGER through habit-breaking (extinction of the association), changing it through counter-stereotyping methods (counter-stereotyping), or controlling it with implementation intention techniques (if-then strategies).

But if the mosaic view is correct, intervention recommendations would not be limited to eliminating/changing/controlling underlying associations or biases (e.g. associations of BLACKS and DANGER). Rather the target of intervention would also be the enhancement and/or habituation of egalitarian background beliefs, commitments, and values (a commitment to being anti-prejudiced or an internal motivation to overcome prejudice). Clea Rees (2016), for example, argues that egalitarian commitments can inhibit the influence of biased associations on cognition outside of conscious awareness and without the need for deliberative control. To support her argument, she cites empirical work by Rudman and colleagues (2001) that shows success in the reduction of prejudice in persons who voluntarily enrol in a diversity education class. The class is designed to raise awareness of racial prejudice, to motivate students to overcome their racism, and provide them with opportunities for social contact with outgroup members in a safe and supportive environment. The interventions here included standard strategies of awareness of bias, as well as the motivation to overcome racism and the commitment to anti-racism (presumably because the decision to enrol in the class suggests motivation and commitment to overcome racism).

Although the matter of how effective intervention strategies turn out to be is an empirical matter, it remains a requirement for philosophical models to instruct empirical research. This is why the mosaic view opens an interesting avenue for experimental psychology to inform the research on whether stocking the background mental states with more egalitarian ‘tools’ is a good way to reduce bias.

Bibliography

Ajzen, I. and Fishbein, M. (1977) 'Attitude-behaviour relations. A theoretical analysis and review of empirical research', *Psychological Bulletin*, 84(5), pp. 888-918.

Allport, G. W. (1935) 'Attitudes', in Murchison, C. (ed.). *Handbook of Social Psychology*. pp. 798–844.

Allport, G. W. (1954) 'Formation of In-Groups', in Allport, G. W. *The Nature of Prejudice*. London: Addison-Wesley Publishing, pp. 29-47.

Alwin, D. and Krosnick, J. (1991) 'The reliability of survey attitude measurement', *Sociological Methods & Research*, 1(20), pp. 139-181.

Amodio, D. and Mendoza S. A. (2010) 'Implicit intergroup bias: Cognitive, affective, and motivational underpinnings', in Gawronski, B. and Payne, B.K. (eds.) *Handbook of Implicit Social Cognition.* Guilford Press, New York, pp. 353-374.

Amodio, D. M. and Swencionis, J. K. (2018) 'Proactive control of implicit bias: A theoretical model and implications for behavior change.', *Journal of Personality and Social Psychology*, 115(2), pp. 255–275.

Baker, E. (2002) 'Flying while Arab - Racial profiling and air travel security', *Journal of Air Law and Commerce*, 67(4), pp. 1375-1403.

Banaji, M. and Hardin, C. (1996). 'Automatic stereotyping', *Psychological Science*, 7, pp. 136-141.

Banaji, M. R., Nosek, B. A., and Greenwald, A. G. (2004) 'No place for nostalgia in science: A response to Arkes and Tetlock', *Psychological Inquiry*, 15(4), pp. 279-310.

Bar-Anan, Y. and Nosek B. A. (2012) 'Reporting intentional rating of the primes predicts priming effects of the Affective Misattribution Procedure', *Personality and Social Psychology Bulletin*, 38(9), pp. 1194-1208.

Bar-Anan, Y. and Nosek, B. A. (2014) 'A comparative investigation of seven indirect attitude measures', *Behavior Research Methods*, 46(3), pp. 668–688.

Bertrand, M. and Mullainathan, S. (2004) 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination', *The American Economic Review*, 94(4), pp. 991-1013.

Blair, I. V., Ma, J. E., and Lenton, A. P. (2001) 'Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery', *Journal of Personality and Social Psychology*, 81(5), pp. 828–841.

Blanton, H., and Jaccard, J. (2006) 'Arbitrary metrics in psychology', *American Psychologist*, 61(1), pp. 27–41.

Blanton, H., and Jaccard, J. (2015) 'Not so fast: Ten challenges to importing implicit attitude measures to media psychology', *Media Psychology*, 18(3), pp. 338-369.

Blanton, H. and Jaccard, J. (2017) ‘You can’t assess the forest if you can’t assess the trees: Psychometric challenges to measuring implicit bias in crowds’, *Psychological Inquiry*, 28(4), pp. 249–257.

Bortolotti, L. (2018) ‘Stranger than fiction: Costs and benefits of everyday confabulation’, *Review of Philosophical Psychology*, 9, pp. 227-249.

Bosson, J. K., Swann, W. B., and Pennebaker, J. W. (2000) ‘Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?’, *Journal of Personality and Social Psychology*, 79(4), pp. 631-643.

Brigham, J. C. (1993) ‘College student’s racial attitudes’, *Journal of Applied Social Psychology*, 23(23), pp. 1933-1967.

Brinol P., Petty, R., and McCaslin, M. (2009) ‘Changing attitudes on implicit versus explicit measures: What is the difference?’, in Petty, R. Fazio, R. and P Brinol, P. (eds.) *Attitudes: Insights from the new implicit measures*, New York: Psychology Press, pp. 285-326.

Brochu, P. M., Gawronski, B., and Esses, V. (2011) ‘The integrative prejudice framework and different forms of weight prejudice: An analysis and expansion’, *Group Processes and Intergroup Relations* 14(3), pp. 429-444.

Brownstein M. (2015) ‘Implicit Bias’, *The Stanford Encyclopaedia of Philosophy* (Spring 2015 Edition). Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/implicit-bias/>.

Brownstein, M. (2016). ‘Implicit attitudes, social learning, and moral credibility.’ In Kiverstein, J. (ed) *The Routledge Handbook on Philosophy of the Social Mind.* New York: Routledge, pp. 298-319.

Brownstein, M. (2017) ‘Implicit bias and race’, in Taylor, P. C., Alcoff, L. M., and Anderson, L. (eds.) *The Routledge Companion to Philosophy of Race*. 1st edn. Routledge, pp. 261–276.

Brownstein, M. and Madva, A. (2012) ‘Ethical Automaticity’, *Philosophy of the Social Sciences*, 42(1), pp. 68–98.

Brownstein, M. and Saul, J. (2016) ‘Metaphysics and Epistemology Introduction’, in Brownstein, M. and Saul, J. (eds.) *Implicit Bias and Philosophy*. Oxford: Oxford University Press, pp. 1–27.

Brownstein, M., Madva, A., and Gawronski, B. (2019) ‘What do implicit measures measure?’, *WIREs Cognitive Science,* 10(5), pp. 1-13. https://doi.org/10.1002/wcs.1501.

Brownstein, M., Madva, A., and Gawronski, B. (2020) ‘Understanding Implicit Bias: Putting the Criticism into Perspective’, *Pacific Philosophical Quarterly*. DOI: 10.1111/papq.12302.

Brunel, F. F., Tietje, B. C., and Greenwald, A. G. (2004) ‘Is the Implicit Association Test a valid and valuable measure of implicit consumer social cognition?’, *Journal of Consumer Psychology,* 14(4), pp. 385–404.

Calanchini, J. and Sherman, J. W. (2013) ‘Implicit attitudes reflect associative, non-associative, and non-attitudinal processes: Implicit attitudes reflect non-attitudinal processes’, *Social and Personality Psychology Compass*, 7(9), pp. 654–667.

Cameron, C. D., Brown-Iannuzzi, J. L., and Payne, B. K. (2012) 'Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes', *Personality and Social Psychology Review,* 16(4), pp. 330–350.

Carifio, J. and Perla, R. J. (2007) 'Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes', *Journal of Social Sciences,* 3, pp. 106-116.

Carruthers, P. (2013) 'On knowing your own beliefs: A representationalist account', in Nottelmann, N. (ed.) *New Essays on Belief*. London: Palgrave Macmillan UK, pp. 145–165.

Carruthers, P. (2017) 'The illusion of conscious thought', *Journal of Consciousness Studies,* 24(9–10), pp. 228–252.

Carruthers, P. (2018) 'Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures?', *Review of Philosophical Psychology,* 9(1), pp. 51–72. https://doi.org/10.1007/s13164-017-0354-3.

Chaiken, S., and Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. Guilford Press.

Cherry, K. (2019) 'How does implicit bias influence behaviour: Explanations and impacts of unconscious bias' *Verywellmind.* 8 February Available at: https://www.verywellmind.com/implicit-bias-overview-4178401 (Accessed: 10 February 2019).

Clark, A. (2007) 'Soft selves and ecological control', in Spurrett, D., Ross, D., Kincaid, H., and Stephens, L. (eds.) *Distributed Cognition on the Will*. Cambridge: The MIT Press.

Cooley, E., Payne, K. Loersch C., and Lei, R. (2015) 'Who owns implicit attitudes? Testing a metacognitive perspective', *Personality and Social Psychology Bulletin,* 41(1), pp. 103–115.

Cooley, E. and Payne, K.B. (2016) 'Using groups to measure intergroup prejudice', *Personality and Social Psychology Bulletin,* 43(1), pp. 46-59.

Correll, J., Park, B., Judd, C., and Wittenbrink, B. (2002) 'The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals.', *Journal of Personality and Social Psychology,* 83(6), pp. 1314–1329.

Correll, J., Hudson, S. M., Guillermo, S., and Ma, D. S. (2014) 'The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass,* 8(5), pp. 201-213.

Crenshaw, K. (1989) 'Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics,' *University of Chicago Legal Forum,* pp. 139-167.

Cunningham, W. A., Preacher, K. J. and Banaji, M. R. (2001) 'Implicit attitude measures: Consistency, stability, and convergent validity', *Psychological Science,* 12(2), pp. 163-170.

Dasgupta, N. (2013) 'Implicit attitudes and beliefs adapt to situations', in *Advances in Experimental Social Psychology*. Elsevier, pp. 233–279.

Dasgupta N. and Greenwald, A. (2001) ‘On the malleability of automatic attitudes combatting automatic prejudice with images of admired and disliked individuals’, *Journal of Personality and Social Psychology,* 81(5), pp. 800-814.

Dasgupta N., Desteno D, Williams L. A. and Hunsinger M. (2009) ‘Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice’, *Emotion*, 9(4), pp. 585–591.

De Houwer, J. (2003) ‘A structural analysis of indirect measures of attitudes’, in J. Much, J. and Klauer, K. C. (eds.) *The psychology of evaluation: Affective processes in cognition and emotion*. pp. 219–244.

De Houwer J. (2005). ‘What are implicit measures and indirect measures of attitude? A comment on Spence (2005)’, *Social Psychological Review*, 7(1), pp. 18-20.

De Houwer, J. (2006) ‘What are Implicit Measures and why are we Using them?’ in Wiers, R.W. and Stacy, A.W. (eds.) *Handbook of Implicit Cognition and Addiction.* California: Sage Publishers, pp. 11-28.

De Houwer J. (2009) ‘The propositional approach to associative learning as an alternative for association formation models’, *Learning & Behavior*. 37, pp. 1-20.

De Houwer, J. (2014) ‘A propositional model of implicit evaluation’, *Social and Personality Psychology Compass*, 8(7), pp. 342–353.

De Houwer, J. and Moors, A. (2010) ‘Implicit measures: Similarities and differences.’ in Gawronski B. and Payne, K. B. (eds.) *Handbook of implicit social cognition: Measurement, theory, and applications*. New York: Guilford Press, pp. 176-193.

De Houwer, J., Geldof, T., and De Bruycker, E. (2006) ‘The Implicit Association Test as a general measure of similarity’, *Canadian Journal of Experimental Psychology*, 59(4), pp. 228-239.

De Houwer J., Teige-Macigemba, S., Spruyt, A. and Moors, A. (2009) ‘Implicit measures: A normative analysis and review.’ *Psychological Bulletin*, (135), p. 347-368.

Del Pinal, G. and Spaulding, S. (2018) ‘Conceptual centrality and implicit bias’, *Mind & Language*, 33(1), pp. 95–111.

Dennett, D. C. (1991) ‘Real patterns’, *The Journal of Philosophy*, 88(1), pp. 27-51.

Desteno, D., Dasgupta, N., Bartlett, M. and Cajdric, A. (2004) ‘Prejudice from thin air: The effect of emotion on automatic intergroup attitudes’, *Psychological Science*, 15(5), pp. 319-324.

Devine, P. G. (1989) ‘Stereotypes and prejudice: Their automatic and controlled components’, *Journal of Personality and Social Psychology*, 56(1), pp. 5–18.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E. and Vance, S. L. (2002) ‘The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice’, *Journal of Personality and Social Psychology*, 82, pp. 835-48.

Devine, P. G., Forscher, P. S., Austin, A. J. and Cox, W. T. (2012) ‘Long-term reduction in implicit race bias: A prejudice habit-breaking intervention’, *Journal of Experimental Social Psychology*, 48, pp. 1267-1278.

Devlin, H. (2018) ‘Unconscious bias: what is it and can it be eliminated?’ *The Guardian*. 2 December Available at: https://www.theguardian.com/uk-news/2018/dec/02/unconscious-bias-what-is-it-and-can-it-be-eliminated (Accessed: 22 December, 2018).

Dovidio, J. F. and Gaertner, S. L. (1998) ‘On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism’, in Eberhardt, J. and Fiske S. T. (eds.) *Confronting racism: The problem and the response*. California: Sage Publications, pp. 3-32.

Dovidio, J. F. and Gaertner, S. L. (2004) ‘Aversive Racism’, *Advances in Experimental Social Psychology*, pp. 1–52.

Dovidio, J. F., Kawakami, K. and Gaertner, S. L. (2002) ‘Implicit and explicit prejudice and interracial interaction’, *Journal of Personality and Social Psychology,* 82(1), pp. 62–68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B. and Howard, A. (1997) ‘On the nature of prejudice: Automatic and controlled processes’, *Journal of Experimental Social Psychology,* 33(5), pp. 510–540.

Eagly, A. H., and Chaiken, S. (1993). *The psychology of attitudes.* Fort Worth, TX, Harcourt Brace Jovanovich College Publishers.

Esfeld, M. (1998) ‘Holism and analytic philosophy’, *Mind*, 107(426), pp. 365-380.

Faigman, D., Kang, J., Bennett, M. Carbado D., Casey P., Dasgupta, N., Godsil, R. Greenwald A., Levinson, J. and J. Mnookin (2012). ‘Implicit bias and the courtroom’. *UC Hastings Scholarship Repository,* pp. 1125-1187.

Fazio, R. H. (1990) ‘Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework’, *Advances in Experimental Social Psychology,* 23, pp. 75-109.

Fazio, R. H. (2000) ‘Accessible attitudes as tools for object appraisal: Their costs and benefits’, in Maio, G. R. and Olson, J., M. (eds.) *Why We Evaluate: Functions of Attitudes*. London: Lawrence Erlbaum Associates, pp. 1-36.

Fazio, R. H. (2001) ‘On the automatic activation of associated evaluations’, *An overview Cognition and Emotion.* 15, pp. 115-141.

Fazio, R. H. (2007) ‘Attitudes as object-evaluation associations of varying strength’, *Social Cognition*. 25, pp. 603-637.

Fazio, R. H. and Hilden, L. E. (2001) ‘Emotional reactions to a seemingly prejudiced response: The role of automatically activated racial attitudes and motivation to control prejudiced reactions’, *Personality and Social Psychology Bulletin,* 27, pp. 538-549.

Fazio, R. H. and Olson, M. A. (2003) ‘Implicit measures in social cognition. research: their meaning and use’, *Annual Review of Psychology,* 54, pp. 297-327.

Fazio, R. H. and Olson, M. A. (2007) 'Attitudes: Foundations, functions, and consequences', in *The SAGE Handbook of Social Psychology: Concise Student* Ed. 1. London: SAGE Publications, pp. 123–145.

Fazio, R. H. and Olson, M. A. (2014) 'The MODE model: Attitude-behavior processes as a function of motivation and opportunity', in *Dual-process theories of the social mind*. New York, NY, US: The Guilford Press, pp. 155–171.

Fazio, R. H., Ledbetter, J. E. and Towles-Schwen, T. (2000) 'On the costs of accessible attitudes: detecting that the attitude object has changed', *Journal of Personality and Social Psychology,* 78, pp. 197-210.

Fazio, R. H., Williams, C. J., and Powell, M. C. (2000) 'Measuring associative strength: Category-item associations and their activation from memory', *Political Psychology*, 21, pp. 7-25.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C. and Kardes, F. R. (1986) 'On the automatic activation of attitudes', *Journal of Personality and Social Psychology*. 50, pp. 229-38.

Fazio, R. H. Jackson, J. R., Dunton, B. C. and Williams, C. J. (1995) 'Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?', *Journal of Personality and Social Psychology*, 69(6), pp. 1013–1027.

Fishbein, M. A. (1967) 'Attitude and the prediction of behaviour', in Fishbein, M. (ed.) *Readings in Attitude Theory and Measurement.* New York: Wiley, pp. 477-492.

Frankish, K. (2016) 'Playing double: Implicit bias, dual levels, and self-control', in Brownstein, M. and Saul, J. (eds.) *Implicit Bias and Philosophy: Metaphysics and Epistemology*. Oxford: Oxford University Press, pp. 23-46.

Fricker, M. (2007) *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford. Oxford University Press.

Friedlaender, C. (2018) 'On microaggressions: Cumulative harm and individual responsibility', *Hypatia*, 33(1), pp. 5–21.

Friese, M., Hofmann, W., and Wanke, M. (2008) 'When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour', *Br J Soc Psychol*, 47(3), pp. 387-419.

Galdi S, Arcuri L., and Gawronski B. (2008) 'Automatic mental associations predict future choices of undecided decision-makers', *Science*, 321(5892), pp. 1100–1102.

Gaskin, R. (2008) *The unity of the proposition.* Oxford, Oxford University Press.

Gawronski, B. (2009) 'Ten Frequently Asked Questions About Implicit Measures and Their Frequently Supposed, But Not Entirely Correct Answers', *Canadian Psychology*, 50(3), pp. 141–150.

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

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, pp. 692-731.

Gawronski, B. and Bodenhausen, G. V. (2007) 'Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model', *Social Cognition*, 25(5), pp. 687–717.

Gawronski, B. and Bodenhausen, G. V. (2011) 'The associative-propositional evaluation model: Theory, evidence, and open questions', *Advances in Experimental Social Psychology*, 44, pp. 59-127.

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

Gawronski, B. and Bodenhausen, G. V. (2014b) 'Implicit and explicit evaluation: A brief review of the associative-propositional evaluation model: APE Model', *Social and Personality Psychology Compass*, 8(8), pp. 448–462.

Gawronski, B. and Brannon, S. M. (2017) 'Attitudes and the implicit-explicit dualism', in Albarracin, D. and T., J. B. (eds.) *The Handbook of Attitudes*. 2nd edn. New York: Taylor & Francis, pp. 1–32.

Gawronski, B. and De Houwer, J. (2014) 'Implicit Measures in Social and Personality Psychology', in Reis, H. T. and Judd, C. M. (eds) *Handbook of Research Methods in Social and Personality Psychology*. 2nd edn. New York: Cambridge University Press, pp. 283–310.

Gawronski, B. and Sritharan, R. (2010) 'Formation, change, and contextualization of mental associations', in Gawronski, B. and Payne B. K. (eds) *Handbook of implicit social cognition: Measurement, theory, and applications*. New York. Guilford Press. pp. 216-240.

Gawronski, B. Walthers, E. and Black, H. (2005) 'Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information', *Journal of Experimental Social Psychology*, 41, pp. 618-626.

Gawronski, B., Hofmann, W. and Wilbur, C. J. (2006) 'Are "implicit" attitudes unconscious?', *Consciousness and Cognition*, 15(3), pp. 485–499.

Gawronski, B., LeBel, E. P. and Peters, K. R. (2007) 'What do implicit measures tell us? Scrutinizing the validity of three common assumptions', *Perspectives on Psychological Science*, 2(2), pp. 181–193.

Gawronski, B., Peters, K. R. and Lebel, E. P. (2008) 'What Makes Mental Associations Personal or Extra-Personal? Conceptual Issues in the Methodological Debate about Implicit Attitude Measures', *Social and Personality Psychology Compass*, 2, pp. 1002–1023.

Gawronski, B., Sherman, J. W. and Trope, Y. (2014) 'Two of what? A conceptual analysis of dual process theories', in Serman, J. W., Gawronski, B., and Trope, Y. (eds) *Dual-process theories of the social mind*. New York, pp. 3–19.

Gawronski, B., Brannon, S. M. and Bodenhausen, G. V. (2017) 'The associative-propositional duality in the representation, formation, and expression of attitudes', in Deutsch, R., Gawronski, B. and

Hofmann, W. (eds.) *Reflective and Impulsive Determinants of Human Behavior*. New York: Psychology Press, pp. 103-118.

Gawronski, B., Morrison, M., Phills, C. E. and Galdi, S. (2017) ‘Temporal stability of implicit and explicit measures: A longitudinal analysis’, *Personality and Social Psychology Bulletin*, 43, pp. 300-312.

Gawronski, B., Peters, K. R., Brochu, P. M. and Strack, F. (2008) ‘Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective’, *Personality and Social Psychology Bulletin*, 34, pp. 648-665.

Gendler, T. S. (2008a) ‘Alief and Belief’, *The Journal of Philosophy*, 105(10), pp. 634-663.

Gendler, T. S. (2008b) ‘Alief in Action (and Reaction)’, *Mind & Language*, 23(5), pp. 552–585.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I. and Banaji, M. R. (2007) ‘Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients’, *Journal of General Internal Medicine*, 22(9), 1231–1238.

Greenwald, A. G. and Banaji, M. R. (1995) ‘Implicit social cognition: Attitudes, self-esteem, and stereotypes’, *Psychological Review*, 102(1), pp. 4–27.

Greenwald, A. G. and Banaji, M. R. (2017) The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, 72, pp. 861–871.

Greenwald, A. G. and Nosek, B. A. (2008) ‘Attitudinal dissociation: What does it mean?’ in Petty, R. E., Fazio, R. H. and Brinol, P. (eds) *Attitudes: Insight from the new implicit measures*. NJ: Lawrence Erlbaum, pp. 65-82.

Greenwald, A. G., McGhee, D. E. and Schwartz, J. K. (1998) ‘Measuring individual differences in implicit cognition: The Implicit Association Test’, *Journal of Personality and Social Psychology,* 74, pp. 1464-1480.

Greenwald, A. G., Poehlman, T., Uhlmann, E. and M. Banaji. (2009) ‘Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity’. *Journal of Personality and Social Psychology*, 97(1), pp. 17-41.

Greenwald, A. G., Banaji, M. R., and Nosek, B. A. (2015) ‘Statistically small effects of the Implicit Association Test can have societally large effects’, *Journal of Personality and Social Psychology*, 108(4), pp. 553-561.

Hahn, A. and Gawronski, B. (2016) ‘Implicit Social Cognition’, in Medina, D. P. H. (ed.) *Stevens Handbook of Experimental Psychology*. 4th edn. New York: Wiley, pp. 1–27.

Hahn, A. and Gawronski, B. (2019) ‘Facing one’s implicit biases: From awareness to acknowledgment.’, *Journal of Personality and Social Psychology*, 116(5), pp. 769–794.

Hahn, A., Judd, C. M., Hirsh, H. K. and Blair, I. (2014) ‘Awareness of implicit attitudes’, *Journal of Experimental Psychology: General,* 143(3), pp. 1369-1392.

Haidt J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment, *Psychol Rev.*, 108(4), pp. 814–834.

Haidt, J. (2017) ‘The unwisest idea on campus: Commentary on Lilienfeld (2017)’, *Perspectives on Psychological Science*, 12(1), pp. 176–177.

Hall, D. L. and Payne, B. K. (2010) ‘Unconscious attitudes, unconscious influence, and challenges to self-control’, in Trope, Y., Ochsner, K. and H. R. (eds.) *Self-control in Society, Mind, and Brain*. 4th edn. New York: Oxford University Press, https://doi.org/10.1093/acprof:oso/9780195391381.003.0012.

Heise, D. R. (1970) ‘The semantic differential and attitude research’ in Summers, G. F. (ed.) *Attitude Measurement*. Chicago: Rand McNally, pp. 235-253.

Higgins, E. T. and Bargh, J. A. (1987) ‘Social cognition and social perception’, *Annual Review of Psychology,* 38, pp. 369–425.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H. and Schmitt, M. (2005) ‘A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures’, *Personality and Social Psychology Bulletin,* 31(10), pp. 1369–1385.

Holroyd, J. (2012) ‘Responsibility for implicit bias’, *Journal of Social Philosophy,* 43(3), pp. 274–306.

Holroyd, J. (2015) ‘Implicit bias, awareness and imperfect cognitions’, *Consciousness and Cognition*, 33, pp. 511–523.

Holroyd, J. (2016) ‘What do we want from a model of implicit cognition?’, P*roceedings of the Aristotelian Society,* 116(2), pp. 153–179.

Holroyd, J. and Sweetman, J. (2016). ‘The Heterogeneity of Implicit Bias’ in Brownstein, M and Saul, J. (eds.) *Implicit Bias and Philosophy: Metaphysics and Epistemology* Oxford: Oxford University Press, pp. 81–114.

Holroyd, J. and Kelly, D. (2016) ‘Implicit bias, character, and control’, in Masala, A. and Webber, J. (eds.) *From Personality to Virtue*. Oxford University Press, pp. 106–133

Holroyd, J., Scaife, R. and Stafford, T. (2017) ‘What is implicit bias?’, *Philosophy Compass,* 12(10), pp. 1-18. https://doi.org/10.1111/phc3.12437.

Howell, J. L. and Ratliff, K. A. (2017) ‘Not your average bigot. The better-than-average effect and defensive responding to Implicit Association Test feedback’, *British Journal of Social Psychology*, 56, pp. 125-145.

Howell, J. L., Gaither, S. E., and Ratliff, K. A. (2015) ‘Caught in the middle: Defensive responses to IAT feedback among Whites, Blacks, and biracial Black/Whites’, *Social Psychological and Personality Science,* 6(4), pp. 373–381.

Hu, X., Gawronski, B. and Balas, R. (2017) ‘Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations’, *Personality and Social Psychology Bulletin,* 43(1), pp. 17–32.

Hughes, S., Barnes-Holmes, D. and De Houwer, J. (2011) ‘The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives’, *The Psychological Record*, 61(3), pp. 465–496.

Jacoby, L. L. and Dallas, M. (1981) ‘On the relationship between autobiographical memory and perceptual learning’, *Journal of Experimental Psychology: General*, 110, pp. 306-340.

Kahneman, D. (2003). ‘A perspective on judgment and choice: Mapping bounded rationality’, *American Psychologist*, 58(9), pp. 697–720.

Kang, J. and Lane, K. (2010) ‘Seeing through colorblindness: Implicit bias and the law’, *UCLA Law Review*, pp. 465-520.

Karpinski, A., Steinman, R. B. and Hilton, J. L. (2005) ‘Attitude Importance as a Moderator of the Relationship Between Implicit and Explicit Attitude Measures’, *Personality and Social Psychology Bulletin*, 31(7), pp. 949–962.

Kraus, M. W. and Park, J. W. (2017) ‘Microaggressions as part of the historical context of stigma and prejudice’, preprint, *PsyArXiv*. doi: 10.31234/osf.io/622ke.

Krysan, M. (2000) ‘Prejudice, politics, and public opinion: understanding the sources of racial policy attitudes’, *Annual Review of Sociology*, 26, pp. 135-68.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G. and Banaji, M. R. (2019) ‘Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis’, *American Psychologist*, 74(5), pp. 569–586.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., and Nosek, B. A. (2014, March 24). Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions. *Journal of Experimental Psychology: General.* Advance online publication. http://dx.doi.org/10.1037/a0036260.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., Hu, X., McLean, M. C., Axt, J. R., Asgari, S., Schmidt, K., Rubinstein., R, Marini, M., Rubichi, S., Shin,. J. L., and Nosek, B. A. (2016) ‘Reducing implicit racial preferences: II. Intervention effectiveness across time’, *Journal of Experimental Psychology: General*, 145(8), pp. 1001-1016.

Lane, K. A., Kang, J. and Banaji, M. R. (2007) ‘Implicit Social Cognition and Law’, *Annual Review of Law and Social Science*, 3(1), pp. 427–451.

Lerner, J.S. and P. E. Tetlock (1999) ‘Accounting for the effects of accountability’, *Psychological Bulletin*, 125(2), pp. 255-275.

Levinson, J. D. and Young, D. (2010) ‘Different shades of bias: Skin tone, implicit racial bias, and judgments of ambiguous evidence’, *West Virginia Law Review* 112, pp. 307-350.

Levinson, J. D., Cai, H. and Young, D. (2010) 'Guilty by implicit racial bias: The guilty/not guilty implicit association test', *Ohio State Journal of Criminal Law*, 8(1), pp. 187-208.

Levy, N. (2014) 'Consciousness, implicit attitudes, and moral responsibility', *Noûs*, 48(1), pp. 21-40.

Levy, N. (2015) 'Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements' *Noûs*, 49(4), pp. 800-823.

Levy, N. (2016) 'Implicit Bias and Moral Responsibility: Probing the Data.', Philosophy and Phenomenological Research, 94(1), pp. 3–26.

Levy, N. (2017) 'Am I a Racist? Implicit Bias and the Ascription of Racism', *The Philosophical Quarterly*, 67(268), pp. 534-551.

Lilienfeld, S. O. (2017) 'Microaggressions: Strong Claims, Inadequate Evidence', *Perspectives on Psychological Science*, 12(1), pp. 138–169.

Livingston, R. W. and Brewer, M. B. (2002) 'What are we really priming? Cue-based versus category-based processing of facial stimuli', *Journal of Personality and Social Psychology*, 82(1), pp. 5–18.

Longino, H.E., (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, NJ.

Lundberg, K. B. and Payne, B. K. (2014) 'Decisions among the undecided: Implicit attitudes predict future voting behavior of undecided voters', *PLoS ONE*, 9(1), e85680, https://doi.org/10.1371/journal.pone.0105655.

Machery, E. (2016) 'De-Freuding implicit attitudes', in Brownstein, M and Saul, J. (eds.) *Implicit Bias and Philosophy: Metaphysics and Epistemology* Oxford: Oxford University Press, pp. 104-129.

Machery, E. (2017) 'Do Indirect Measures of Biases Measure Traits or Situations?', *Psychological Inquiry*, 28(4), pp. 288–291.

Madva, A. (2016) 'Why implicit attitudes are (probably) not beliefs', *Synthese*, 193(8), pp. 2659–2684.

Mandelbaum, E. (2013) 'Against alief', *Philosophical Studies*, 165(1), pp. 197–211.

Mandelbaum, E. (2014) 'Thinking is Believing', *Inquiry*, 57(1), pp. 55–96.

Mandelbaum, E. (2016) 'Attitude, inference, association: On the propositional structure of implicit bias', *Noûs*, 50(3), pp. 629–658.

McLaughlin, E. C. (2018) 'Café shut down after protesters enter, chanting 'Starbucks coffee is anti-black!', *CNN*, 17 April. Available at https://edition.cnn.com/2018/04/16/us/philadelphia-police-starbucks-arrest-protests/index.html (Accessed: 12 January 2019).

McConahay, J. B. (1986) 'Modern racism, ambivalence, and the Modern Racism Scale' in Dovidio, J. F. and Gaertner, S. L. (eds.) *Prejudice, discrimination, and racism*. San Diego, CA Academic Press. pp. 91-125.

McConahay, J. B., Hardee, B. B. and Batts, V. (1981) 'Has racism declined in America? It depends on who is asking and what is asked, *The Journal of Conflict Resolution*, 25(4), pp. 563-579.

McGrath, M. and Devin, F. (2018) ‘Propositions’, *The Stanford Encyclopaedia of Philosophy*. Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/propositions/>.

Mendoza, S. A., Gollwitzer, P. M. and Amodio, D. M. (2010) ‘Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions’, *Personality and Social Psychology Bulletin*, 36(4), pp. 512-523.

Mischell, W. and Shoda, Y. (1995) ‘A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure’, *Psychology Review*, 102(2), pp. 246-268.

Mitchell, G. (2017) ‘Measuring Situational Bias or Creating Situational Bias?’, *Psychological Inquiry*, 28(4), pp. 292-296.

Mitchell, J. P., Nosek. B. A. and Banaji, M. R. (2003) ‘Contextual variations in implicit evaluation’, *Journal of Experimental Psychology*: General, 132, pp. 455-469.

Monteith, L. and Pettit, J. (2011) ‘Implicit and explicit stigmatizing attitudes and stereotypes about depression’, *Journal of Social and Clinical Psychology*, 30(5), pp. 484-505.

Monteith, Wookcock, and Gulker (2013) ‘Automaticity and control in stereotyping and prejudice: The revolutionary role of social cognition across three decades of research’, in Carlston, D. E. (ed.). *Oxford Library of Psychology. The Oxford Handbook of Social Cognition*: Oxford University Press. pp. 74-94.

Nelson, T. D. (2009) *Handbook of Prejudice, Stereotyping, and Discrimination*: 2nd Edition. Psychology Press.

Nier, J. A. (2005) ‘How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach’, *Group Processes & Intergroup Relations*, 8(1), pp. 39–52.

Nosek, B. A. (2005) ‘Moderators of the relationship between implicit and explicit evaluation’, *J Exp Psychol Gen*, 134(4), pp. 565–584.

Nosek, B. A. (2007) ‘Implicit – Explicit Relations’, *Current Directions in Psychological* Science, 16(2), pp. 65–69.

Nosek, B. A. and Banaji, M. R. (2001) ‘The Go/No-go association task’, *Social Cognition*, 19(6), pp. 625–666.

Nosek, B. A., Greenwald, A. G., Banaji, M. R. (2007) ‘The Implicit Association Test at age 7: A methodological and conceptual review’, in Bargh, J. A. (ed.) *Automatic Processes in Social Thinking and Behavior*. Psychology Press, pp. 265–292.

Olson, M. A., Kendrick, R. V. and Fazio R. H. (2009) ‘Implicit learning of evaluative vs. non-evaluative covariations: The role of dimension accessibility’, *Journal of Experimental Social Psychology*. 45, pp. 398-403.

Oskamp, S., Harrington, M. J., Edwards, T. C., Sherwood, D. L., Okuda, S. M. and Swanson, D. C. (1991) ‘Factors influencing household recycling behavior’, *Environment and Behavior*, 23, pp. 494-519.

Ostrom, T. M. (1969) 'The relationship between the affective, behavioral, and cognitive components of attitude', *Journal of Experimental Social Psychology*, 5(1), pp. 12–30.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. (2013) 'Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies', *Journal of Personality and Social Psychology*, 105(2), pp. 171-192.

Payne, B. K. (2001) 'Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon', *Journal of Personality and Social Psychology*, 81, pp. 181-192.

Payne, B. K. (2005) 'Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping', *Journal of Personality and Social Psychology,* 89(4), pp. 488–503.

Payne, B. K. (2006) 'Weapon bias: Split second decisions and unintentional stereotyping', Current Directions in Psychological Science, 15(6), pp. 287-291.

Payne, B. K. and Lundberg, K. B. (2014) 'The Affect Misattribution Procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*. 8(12), pp. 672–686.

Payne, B. K. and Gawronski, B. (2010) 'A history of implicit social cognition where is it coming from? Where is it now? Where is it going?', in Gawronski, B. and Payne, B. K. (eds.) *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, pp. 1–15.

Payne, B. K., Burkley, M. A. and Stokes, M. B. (2008) 'Why do implicit and explicit attitude tests diverge? The role of structural fit', *Journal of Personality and Social Psychology*, 94(1), pp. 16–31.

Payne, B. K., Vuletich, H. A. and Lundberg, K. B. (2017a) 'The bias of crowds: How implicit bias bridges personal and systemic prejudice', *Psychological Inquiry*, 28(4), pp. 233–248.

Payne, B. K., Vuletich, H. A. and Lundberg, K. B. (2017b) 'Flipping the script on implicit bias research with the bias of crowds', *Psychological Inquiry*, 28(4), pp. 306–311.

Payne, B.K., Cheng, C. M., Govorun, O., and Stewart, B. (2005) 'An inkblot for attitudes: Affect misattribution as implicit measurement', *Journal of Personality and Social Psychology*, 89, pp. 277-293.

Perugini, M., Richetin, J. and Zogmaister, C. (2010) 'Prediction of behavior', in Gawronski B. Payne, B. K. (eds.) *Handbook of implicit social cognition: Measurement, theory, and applications*. The Guilford Press, p. 255–277.

Pierce, C. (1995) 'Stress analogs of racism and sexism: Terrorism, torture, and disaster', in Willie, C., Rieker, P., Kramer, B. and Brown, B. (eds.), *Mental health, racism, and sexism* Pittsburgh, PA: University of Pittsburgh Press, pp. 277–293.

Plant, A. E. and Peruche, M. B. (2005) 'The consequences of race for police officers' responses to criminal subjects', *Psychological Science* 16, pp. 180-183.

Project Implicit (2019) '*Project implicit social attitudes'* Available at: https://implicit.harvard.edu/implicit/ (Accessed: 01 February 2019).

Quine, W. v. O. (1951) 'Two Dogmas of Empiricism' in Curd, M., Cover, J. A., and Pincock, C. (eds.) *Philosophy of Science: The central issues*. W. W. Norton & Company, London, pp. 250-270.

Quinn, A. and Schlenker, B. R. (2002) 'Can accountability produce independence? Goals as determinants of the impact of accountability on conformity', *Personality and Social Psychology Bulletin,* 28(4), pp. 472-483.

Rees, C. (2016) 'A virtue ethics response to implicit bias' in Brownstein M. and Saul J. (eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. Oxford Press, Oxford, pp. 191-214.

Rivers, A. M., Rees, H. R., Calanchini, J. and Sherman, J. W. (2017) 'Implicit bias reflects the personal and the social', *Psychological Inquiry*, 28(4), pp. 301-305.

Rooth, D. O. (2010) 'Automatic associations and discrimination in hiring: Real world evidence', *Labour Economics*, 17, pp. 523-534.

Rozin P., Millman, L. and Nemeroff, C. (1986) 'Operation of the laws of sympathetic magic in disgust and other domains', *Journal of Personality and Social Science*, 50, pp. 703–712.

Rozin, P., Markith, M., and Ross, B. (1990) 'The sympathetic magical law of similarity, nominal realism, and neglect of negatives in response to negative labels', *Psychological Science*, 1(6), pp. 383-384.

Ruben, D.H. (2003) *Action and its explanation*. 1st ed. Oxford: New York: Clarendon Press; Oxford University Press.

Rudman, L. A., Greenwald, A. G., Mellott, D. S. and Schwartz, J. (1999) 'Measuring the automatic components of prejudice: Flexibility and generality of the implicit association test', *Social Cognition*, 17(4), pp. 437-465.

Rudman, L. A., Ashmore, R. D. and Gary, M. L. (2001) '"Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes', Journal of Personality and Social Psychology, 81(5), pp. 856-868.

Samra, R. (2014) 'A new look at our old attitude problem', *Journal of Social Sciences*, 10(4), pp. 143–149.

Saul, J. (2012) 'Ranking exercises in philosophy and implicit bias', *Journal of Social Philosophy*, 43(3), pp. 256-273.

Saul, J. (2013a) 'Implicit Bias, Stereotype Threat and Women in Philosophy', in Jenkins, F. and Hutchison, K. (eds.) *Women in Philosophy: What needs to Change?* Oxford University Press, pp. 39-60.

Saul, J. (2013b) Skepticism and Implicit Bias, *Disputatio* 5(37), pp. 243-263.

Schacter, D.L., Bowers, J., and Booker, J. (1989) 'Intention, awareness, and implicit memory: The retrieval intentionality criterion', in Lewandowsky, S., Dunn, J. C. and Kirsner K. (eds.). *Implicit memory: Theoretical issues.* Hillssdale, NJ: Erlbaum, pp. 47-65.

Schimmack, U (2019) ‘The implicit association test: A method in search of a construct’, *Perspectives on Psychological Science*, pp. 1-19.

Schwarz, N. (1999) ‘Self-reports: How the questions shape the answers’, *American Psychologist*, 54(2), pp. 95–105.

Schwitzgebel, E. (2010) ‘Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief’, *Pacific Philosophical Quarterly*, 91(4), pp. 531-553.

Sears, D. O. (1988) ‘Symbolic racism’, in Katz, P. A. and Taylor, D. A. (eds.) *Perspectives in social psychology. Eliminating racism: Profiles in controversy.* Plenum Press, (p. 53–84).

Sherman, B. R. (2015) ‘There’s no (testimonial) justice: Why pursuit of a virtue is not the solution to epistemic injustice’, *Social Epistemology*, pp 1-21.

Shoda, T. M., McConnell, A. R. and Rydell, R. J. (2014) ‘Implicit consistency processes in social cognition: Explicit-implicit discrepancies across systems of evaluation: implicit consistency processes’, *Social and Personality Psychology Compass*, 8(3), pp. 135-146.

Simeoni, B. Z. (2005) Testing tests: Determination of the efficacy of prejudice measures. Georgia State University. [Online]. Available at: https://digitalcommons.georgiasouthern.edu/cgi/viewcontent.cgi?article=1434&context=etd (Accessed 23 February 2018).

Sirois, M. (2017) ‘The kind of racism we don’t even know we have’ *Medium.* 21 November. Available at: https://gendercreativelife.com/2017/11/21/the-kind-of-racism-you-dont-even-know-you-have/ (Accessed: 02 March, 2018).

Smith, E. R. and DeCoster, J. (2000) ‘Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems’, *Personality and Social Psychology Review*, 4(2), pp. 108–131.

Spence, J. T., Helmreich, R. L., Stapp, J. (1973) ‘A short version of the attitudes toward women scale’, *Bulletin of the Psychonomic Society*, 2, pp. 219–220.

Spruyt, A. Gast, A. Moors, A. (2011) ‘The sequential priming paradigm: A primer’ in Klauer, K. C., Stahl, C. and Voss, A. (eds.) *Cognitive Methods in Social Psychology.* Guilford Press, pp.48-77.

Stammers, S. (2016) *Awareness, Control and Responsibility for Implicit Bias: The Continuum Thesis.* PhD thesis. King’s College London [Online]. Available at: https://kclpure.kcl.ac.uk/portal/files/57597178/2016_Stammers_Sophie_1239562_ethesis.pdf (Accessed 08 May 2018).

Stammers, S. (2017) ‘A patchier picture still: Biases, beliefs, and overlap on the inferential continuum’, *Pholosophia*, 45, pp. 1829-1850.

Stich, S. (1996) ‘Deconstructing a Deconstruction: A Preview of Coming Attractions’, *Deconstructing the Mind*, pp. 1-60.

Sue, D. W. (2005). Racism and the conspiracy of silence: Presidential address. *The Counseling Psychologist*, *33*(1), pp. 100–114.

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons Inc.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., and Esquilin, M. (2007) 'Racial microaggressions in everyday life: Implications for clinical practice', *American Psychologist*, 62(4), pp. 271–286.

Sullivan-Bissett, E. (2019) 'Biased by Our Imaginings'. *Mind and Language*. 34(5), pp. 627–647.

Swim, J. Aikin, K. Hall, W. (1995) 'Sexism and Racism: Old-fashioned and modern prejudices, *Journal of Personality and Social Psychology*, 68(2), pp. 199-214.

Swim, J. K. and Hyers, L. L. (2009) 'Sexism' in T. D. Nelson (ed.) *Handbook of prejudice, stereotyping, and discrimination.* Psychology Press, pp. 407–430.

Torbio, J. (2018) 'Implicit bias: From social structure to representational format', *Theoria*. 33(1), pp. 41-60.

Tulving, E., Schacter, D. L. and Stark, H. A. (1982) 'Priming effects in word-fragment completion are independent of recognition memory' *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, pp. 336-342.

Uhlmann, E. L. and Cohen, G. L. (2005) 'Redefining Merit to Justify Discrimination', *Psychological Science*, 16(6), pp. 474–480.

Kirwan Institute (2015) *Understanding implicit bias: State of the science.* Available at: http://kirwaninstitute.osu.edu/research/understanding-implicit-bias/ (Accessed: 23 August 2017).

Valian, V. (1998) 'Sex, schemas, and success: What's keeping women back?', *Academe*, 84(5), pp. 50-55.

Valian, V. (2010) 'What works and what doesn't: How to increase the representation of women in academia and business', in Riegraf, B., Aulenbacher, B., Kirsch-Auwarter, E. and Muller, U. (ed.) *Gender Change in Academia*. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 317–328.

Warrington, E.K., and Weiskrants, L. (1974) 'The effect of prior learning on subsequent retention in amnesic patients', *Neuropsychology*. 12, pp. 419-428.

Wilson, T. D., Lindsey, S. and Schooler, T. Y. (2000) 'A model of dual attitudes', *Psychological Review*, 107(1), pp. 101–126.

Yao, V. and Reis-Dennis, S. (no date) '"I Love Women:" The Conceptual Inadequacy of "Implicit Bias"'. Available at: http://peasoup.us/2017/09/love-women-conceptual-inadequacy-implicit-bias-yao-reis-dennis/.

Zajonc, R. B. (1968) 'Attitudinal effects of mere exposure', *Journal of Personality and Social Psychology*, 9(2), pp. 1–27.