The Nature of Attitudes: Profiles of Situation-Specific Evaluative Response Dispositions

Andreas Bunge

Thesis submitted for the degree of PhD in Philosophy

Department of Philosophy
University of Sheffield
December 2017

Abstract

In this thesis, I develop a model of the nature of attitudes, broadly construed as people's evaluative tendencies towards other people qua members of social groups. I set out three desiderata for such a model: it should be conducive to explanations and predictions of people's evaluative responses towards other people (D1), it should provide an appropriate guide to moral character assessment (D2), and it should be a model that all parties that use the notion of an attitude could possibly agree on (because this would simplify knowledge exchange between these parties; D3). According to a model that is prevalent in the contemporary psychological and philosophical literature on prejudice, people's attitudes fall into two classes: implicit and explicit attitudes (both of which are specific kinds of mental states). I show that this account is not well motivated and argue that there is an alternative model of attitudes available that is more in line with desiderata D1, D2, and D3. Building upon an account by Machery (2016), I claim that attitudes are traits of people. As such, attitudes are neither implicit nor explicit, but they are typically grounded in sets of implicit and explicit mental states (e.g., conceptual associations, affects, beliefs, desires). Contra Machery, I argue that these attitudes are not properly characterised in aggregationist terms because this obscures relevant evaluative complexities of attitudes. Instead, these attitudes should be analysed as profiles of situation-specific evaluative response dispositions. This model does justice to the fact that people's evaluative responses are strongly context-dependent. Taking this context dependence into account helps us to explain and predict people's evaluative responses (D1) and to appropriately evaluate people's moral characters (D2). Due to these benefits, the proposed model should appeal to different parties (philosophers, psychologists, and ordinary people) that rely on the notion of an attitude (D3).

Acknowledgements

I am grateful to the Leverhulme Trust for providing funding for me as a PhD student on the “Bias and Blame” research project (RPG-2013-326). Without this generous financial support, this thesis would not have been possible. I am also grateful to the contributors to the “Bias and Blame” project – Jules Holroyd, Robin Scaife, and Tom Stafford – for helpful conversations that greatly informed my work on this thesis. Special thanks to Jules Holroyd, my first supervisor, for the continuous support, the numerous meetings, and in particular the detailed comments on numerous text drafts that played a large role in improving this thesis. I would also like to thank Luca Barlassina, who provided supervision during Jules’ maternity leave, coinciding with the final stage of my PhD. I have been very lucky to be part of two great philosophy graduate communities: for the first part of my PhD at the University of Nottingham and for the second part of my PhD at the University of Sheffield. I would like to thank my fellow graduate students for the friendly and supportive environment that they have provided throughout my studies as well as all the insightful and lively conversations that we had. Finally, I would like to thank Alexander Skulmowski for the proofreading, but first and foremost for his invaluable intellectual and emotional support without which my time as a PhD student would have been a much less pleasurable experience. I am honoured to call you my friend.

Content

Introduction

We frequently refer to people's attitudes in day-to-day conversation. For example, we may say that someone exhibits a negative or a positive attitude towards a particular group of people, say immigrants, or that someone possesses a racist or sexist attitude.[1] Attitude ascriptions of this sort help us to explain and predict people's responses towards other people as well as to convey information about a person's character. It is important to note that attitudes are not only a folk psychological posit but that the notion of an attitude also plays a crucial role in academic psychology. As early as 1935, Gordon Allport noted in his seminal article "Attitudes" that "[n]o other term appears more frequently in the experimental and theoretical literature" (p. 798), and there is no doubt that the prevalence of the attitude notion in the psychological literature has persisted until today. Psychologists construe attitudes broadly as evaluations (evaluative mental states or evaluative tendencies) in regard to an entity (e.g., a social group) that become expressed in cognition, affect, and behaviour (Ajzen, 1988; Eagly & Chaiken, 1993; Fabrigar, MacDonald, & Wegener, 2005). Attitudes are often said to have a valence (i.e., they are positive or negative) and to vary in strength (e.g., Fazio, 2007). In recent years, philosophers have shown an increased interest in this notion of an attitude (Frankish, 2016; Machery, 2016; Webber, 2013, 2016b).[2] Yet, as I will point out shortly, it remains unclear how exactly we should conceive of the nature of attitudes. My goal in this thesis is to remedy this shortcoming.

Although people may have attitudes towards all kinds of entities (e.g., objects, institutions, events, brands, or beliefs), I restrict my investigation in this thesis to attitudes towards social groups (or towards people qua members of social groups). I choose this focus because attitudes towards social groups have significant moral implications and are thus particularly interesting from a philosophical point of view. When I use the term "attitude" in this thesis, I am thus always referring to attitudes towards social groups (or towards people qua members of social groups). Yet, despite this focus, many of the conclusions that I reach in this thesis may equally apply to attitudes towards other kinds of entities.

[1] It shall be mentioned that there are of course also other usages of the term "attitude" in ordinary discourse, such as when we say that someone "has quite an attitude" or when we say that someone "has a bad attitude". I am not concerned with these usages in this thesis.

[2] In this thesis, I am not concerned with any of the technical uses of the term "attitude" as they are prevalent in philosophy, such as in the notions of "intentional attitude", "propositional attitude", or "reactive attitude". See Webber (2013) for an elaboration on the complex relation between what psychologists call "attitude" and the philosophical notions of intentional and propositional attitudes (pp. 1085-1087).

Consider the following case, which showcases how hard it can be to pin down a person's attitude and which I use to motivate the questions that I am concerned with in this thesis. Sarah, who identifies herself as white, condemns racism. She endorses egalitarian values, believes that it is morally reprehensible to treat people differently because of their skin colour, and desires not to discriminate against black people. Many of her deliberate responses in regard to black people fall into line with her anti-racist ideals. She is a physician and actively encourages young black people to study medicine because she is concerned about the underrepresentation of black people in the profession. She has repeatedly participated in rallies against the oppression of black people. When she hears someone making a racist joke, she calls that person out. Yet, Sarah's spontaneous responses towards black people are often at odds with her anti-racist ideals. When speaking to black patients, she tends to keep more spatial distance and to make less eye contact than when speaking to white patients. Moreover, when walking home through a deprived neighbourhood in which crime and violence is rife, she reacts with anxiety when black people are approaching her but stays entirely calm when white people come her way. On a few occasions, she even mistook a harmless object held by a black person for a gun. This has never happened to her with respect to a white person. When made aware of these biases, Sarah feels genuine regret. She realises that her spontaneous reactions in regard to black people are at odds with her egalitarian values. Yet, she has a hard time changing her unintentional responses.[3]

It shall be emphasised that the case of Sarah is not a far-fetched fiction. There is in fact abundant empirical evidence that people who endorse egalitarian values often also exhibit biases of the kind that Sarah exhibits.[4] Studies have revealed, amongst others, that even egalitarian minded people often keep more distance and make less eye contact with black than with white interaction partners (Dotsch & Wigoboldus, 2008; Dovidio et al., 1997), and tend to mistake ambiguous objects in black people's hands for guns when they are prompted to make quick "gun-or-no-gun" decisions in a computer simulation (Correll et al., 2002, 2007).

Cases like Sarah's raise a range of interrelated questions about the nature of attitudes that I want to address in this thesis:

(Q1) How should we individuate attitudes?
(Q2) What mental states underpin attitudes?

[3] The philosophical moral psychology literature is replete with examples that resemble the here presented case of Sarah (e.g., Besser-Jones, 2008; Smith, 2004; Holroyd, Scaife, & Stafford, 2017a)

[4] Biases of this kind are often referred to as "implicit biases". See Bronwstein (2017) and Holroyd and colleagues (2017b) for reviews, and Brownstein and Saul (2016a, 2016b) for an extensive collection of articles on the phenomenon of implicit bias. See also footnote 11 below.

(Q3) What is the ontological status of attitudes?

It seems intuitive to say that Sarah's responses towards black people fall into two classes. On the one hand, she condemns racism and engages in various behaviours that reflect her concern for black people. On the other hand, she exhibits various spontaneous responses that seem to reflect negativity towards black people. This is a common pattern that philosophers and psychologists alike have described as "aversive racism" (e.g., Brownstein & Madva, 2012a; Dovidio & Gaertner, 2000). Note that this intuitive characterisation leaves the question of attitude individuation open (*Q1*). Are we to say that Sarah harbours two conflicting attitudes towards black people (maybe a positive and a negative one)? Or are we to say that only one of these response classes is expressive of Sarah's "real" attitude towards black people? Could we maybe even say that Sarah exhibits a complex attitude towards black people that includes all of her response tendencies towards black people? Note also that the question of attitude individuation is directly linked to the question about the mental states that underpin attitudes (*Q2*). Sarah's attitude(s) towards black people could be based on associations in her memory (e.g., an association between her concept BLACK PERSON and her concept DANGER)[5], on affective dispositions (e.g., her disposition to feel scared of black people), on beliefs of hers (e.g., her belief that it is problematic to treat people differently because of their skin colour), or maybe a cluster of all (or a number of) these states. This again relates directly to the question about the ontological status of attitudes (*Q3*). We could say that for any attitude X of Sarah, X can be identified with an individual mental state (e.g., her association between BLACK PERSON and DANGER or her belief that it is morally reprehensible to treat people differently because of their skin colour). Yet, we may also be inclined to say that her attitude is a complex trait of hers that is based on a variety of different mental states and dispositions (e.g., her association between BLACK PERSON and DANGER, plus her disposition to feel scared of black people, plus her belief that it is morally reprehensible to treat people differently because of their skin colour, etc.). In this thesis, I will argue for just such a trait view of attitudes.

However, before I go about answering the aforementioned questions about the nature of attitudes, I will address in the next section (section I) the question as to what we need the notion of an attitude for (in ordinary discourse, in psychology, and philosophy). This will allow me to derive some desiderata for a model of attitudes. Throughout this thesis, these desiderata will guide my search for answers to questions *Q1*, *Q2*, and *Q3*. In section II of the present introduction, I will then elaborate on the conception of attitudes that is predominant in the contemporary psychology and

[5] Throughout this thesis, I use capital letters when mentioning mental concepts.

philosophy of prejudice (what I call “the standard view”). This is the view that people possess distinct implicit and explicit attitudes. I will review what answers the standard view provides with respect to *Q1*, *Q2*, and *Q3*. In section III of the present introduction, I will provide a brief overview of my main argument against the standard view and in favour of an alternative trait view of attitudes, and in section IV, I will provide an overview of the content of the individual chapters of this thesis.

I. Desiderata for a model of attitudes

Before we address questions *Q1*, *Q2*, and *Q3* about the nature of attitudes, it is worth considering why we need the notion of an attitude at all. In short, there are two broad functions that the notion of an attitude fulfils:

(F1) The notion of an attitude plays a role in explanations and predictions of people’s cognitive, affective, and behavioural responses towards other people.

(F2) The notion of an attitude plays a role in the assessment of people’s moral character.

Let us elaborate on these functions in turn. In both folk and academic psychology, the notion of an attitude plays an explanatory and a predictive role (*F1*). It should be noted that prediction and explanation of people’s responses are tightly linked to each other. If we can predict a certain response of a person (e.g., a person’s aversion of eye contact with black people) by pointing to the fact that the person possesses a certain attitude (e.g., by pointing to the fact that the person possesses a negative attitude towards black people), we can also retrospectively explain that response with reference to the fact that the person possesses that attitude. We may hope that knowing about a person’s attitude towards a social group might help us to predict a vast array of responses of that person towards members of the respective social group. Suppose that someone tells you that Chung, of whom you have no other information, has a negative attitude towards black people. This will certainly lead you to form some expectations about Chung’s responses in regard to black people. You may, for example, expect him to keep above-average distance to black interlocutors (Dotsch & Wigboldus, 2008) or to shortlist disproportionally few people with “black sounding” names when being on a hiring committee (Purkiss et al., 2006). In fact, knowing about Chung’s attitude may not only help you to predict his overt behaviour towards black people but also relevant aspects about his cognitions and affective responses (Dotsch

& Wigboldus, 2008; Dovidio et al., 1997).[6] You may predict that negative stereotypes about black people will come to his mind and that he may feel scared or angry when he encounters or imagines black people. Similarly, you may draw on the fact that Chung has a negative attitude towards black people to explain, retrospectively, his responses towards black people. You may wonder why Chung kept so much distance to the man that he was talking to and come to the conclusion that this was likely because the men was black and Chung has a negative attitude towards black people. Its role in the explanation and prediction of people's cognition, affect, and behaviour is also the reason why the attitude notion is so widely used in academic psychology. Psychologists assume that people's reactions towards other people are, at least partly, driven by some sort of evaluative mental state or disposition, which is referred to as "attitude" (Ajzen, 1988: 1).

Note that if we were told that Sarah has a negative attitude towards black people, we would frequently go wrong in our predictions of her responses towards black people. After all, Sarah exhibits a range of favourable responses concerning black people (she encourages them to study medicine, she participates in anti-racism rallies, etc.). Similarly, if we were simply told that Sarah has a positive attitude towards black people, we would presumably also form wrong predictions. We would, for example, not expect her to keep more distance towards black people than towards white people. But how should we then describe Sarah's attitude towards black people to facilitate optimal predictions? In the next section, I present one suggestion, which I call "the standard view". For now though, I want to stress that the notion of an attitude can only fulfil its explanatory and predictive function if it picks out exactly those features of an individual's psychology that drive that individual's evaluative responses towards the group in question.[7] Ascribing a positive attitude towards black people to Sarah only picks out a subset of those features that drive her responses towards black people (e.g., her belief that it is morally reprehensible to treat people differently because of their skin colour). Similarly, ascribing a negative attitude towards black people to Sarah would direct our attention only to a part of what drives her responses towards black people (e.g., her fear of black gun violence). This brings me to my first desideratum for a model of attitudes:

[6] The fact that attitudes become expressed in cognition, affect, and behaviour has long been recognised by proponents of the so-called tripartite model of attitudes, originally proposed by Rosenberg & Hovland (1960).

[7] I deliberately use the broad notion "features of an individual's psychology" to cover all those entities, such as mental states, mental processes, dispositions, or traits, that may possibly constitute attitudes. Similarly, I have a broad notion of "evaluative response" in mind. This includes all occurring cognitions, affects, and behaviours that express an evaluation. The occurrent thought "black people are dangerous", the occurrent feeling of fear of black people, or the excessive distance that Sarah keeps towards black people all express a negative evaluation and thus count as negative evaluative responses on my account.

(D1) To optimally fulfil its explanatory and predictive function, our notion of a person's attitude towards group X must pick out exactly those features of that person's psychology that drive that person's evaluative responses towards group X.

In ordinary discourse, the attitude notion also fulfils a character evaluative function (*F2*). Note that when we say that someone exhibits a racist, sexist, or homophobic attitude, we convey information that the recipient of this message will use in her moral assessment of the person. For example, if we are told that Chung has a negative attitude towards black people, we may well come to the conclusion that Chung is morally corrupt. Note that our verdict about Chung's character is only justified if we understand Chung's attitude as a feature of his psychological make up for which he is morally evaluable. Sarah's case is complicated by the fact that she endorses egalitarian values and regrets her unintentional biases against black people. One may thus argue that Sarah is an egalitarian and that her discriminatory tendencies are not part of what she *really* stands for (Glasgow, 2016).[8] In short, her problematic biases may not reflect on her moral character. I elaborate on the question as to what kind of dispositions can be said to reflect on a person's moral character in later chapters of this thesis (in particular chapter 2 and chapter 5). For now though, it is important to note that this points us to a second desideratum for a model of attitudes:

(D2) To optimally fulfil its role in character assessment, our notion of a person's attitude towards group X should be sensitive to any difference that there may be between aspects of that person's psychology that can rightly be said to be constitutive of that person's moral character and those aspects that are not part of that person's moral character.

Note that this is not yet to say that there is in fact such a distinction to be made between evaluative tendencies of a person that form part of her moral character and those that are not. My claim is conditional: if there is such a distinction to be made, our model of attitudes should account for this.[9] This is a requirement that has largely been neglected in the psychological literature.

So far, I have mentioned two desiderata for a model of attitudes that can be derived from functions *F1* and *F2* of the attitude concept. These desiderata will guide my evaluation of views concerning what attitudes are, and how we need to individuate

[8] This view can be motivated by Frankfurt's (1971, 1988) account of agency, as I show in chapter 5.

[9] In fact, I argue in chapter 5 that even those evaluative dispositions that the agent does not identify with or feels alienated from can and should be seen as part of her moral character.

them, throughout this thesis. Of course, it could turn out that there is no model of attitudes that would in fact fulfil both desiderata. That is, it could turn out to be the case that any conception of attitudes that would satisfy *D1* does not satisfy *D2* (and vice versa). However, if there is a model that fulfils both desiderata it should be preferred over models that only fulfil one of these.

My goal is to develop a model of the nature of attitudes that is in line with the empirical evidence, that proves useful for psychological and philosophical research on issues such as prejudice, discrimination, sexism, or racism, and that can also guide our day-to-day attitude ascriptions. This is of course an ambitious aim as psychologists, philosophers, and ordinary people (folk psychologists) may possibly have different conceptions of attitudes. Note, for example, that academic psychologists may not necessarily be concerned about the character evaluative role of attitudes, while this is important to philosophers and folk psychologists (*F1*). Yet, I believe that it would be highly beneficial if all these parties could find common ground regarding their understanding of attitudes because this would simplify communication between academic disciplines as well as between academia and the general public. We may state this as a third desideratum:

(D3) To facilitate communication on attitudes between academic disciplines as well as between academia and the wider public, our notion of a person's attitude towards group X should ideally be a notion that psychologists, philosophers, and ordinary people can agree on.

If philosophers and psychologists would use the same attitude notion, this would facilitate cross-disciplinary discourse on important issues such as discrimination. Moreover, if scholars in philosophy and psychology as well as ordinary people would use the same attitude notion, this would simplify knowledge exchange between academia and the wider public. It must be stressed that scholarship on socially pressing issues should aim to inform public discourse. This can only be achieved if scholars communicate their findings or arguments in a way that is widely accessible. It will be easier to inform the general public or policy makers about attitude research if the attitude notion used corresponds, at least roughly, to how ordinary people (folk psychologists) use the term. Of course, it may (sometimes) be the case that ordinary discourse about psychological phenomena is confused, in which case it may actually be advisable to replace folk psychological concepts with scientific ones (P. M. Churchland, 1981; P. S. Churchland, 1986; Stich, 1983).[10] Yet still, if there are different alternative models of attitudes available that are scientifically (and philosophically)

[10] See section 4.5 in chapter 4 for an argument to this effect.

sound, we may as well prefer the model that corresponds best to the folk psychological notion of attitudes in order to facilitate communication between academia and the wider public.

II. The standard view

It has become common in psychology, and also in the philosophy of prejudice, to distinguish between implicit and explicit attitudes (Machery, 2016). On this perspective, which I call the "standard view", Sarah's unintentional biases against black people (e.g., her tendency to keep distance to black people) are based on a negative implicit attitude (or as it is sometimes called an "implicit bias")[11]. Implicit attitudes are commonly understood to operate outside of the person's control (and awareness).[12] As a consequence, they are often at odds with the person's explicitly endorsed beliefs or values.[13] By contrast, Sarah's tendency to condemn racism is reflective of an explicit attitude of hers on this view because this tendency is based on her endorsed beliefs which are subject to control.[14] In fact, what I call the standard view is a cluster of views, which share the central assumption that people possess implicit and explicit attitudes (e.g., Dovidio et al., 1997; Wilson, Lindsey, & Schooler, 2000; Levy, 2014b). These views differ in many details, but there is a substantial agreement among proponents of the standard view as to the nature of attitudes.

With respect to *Q2*, many proponents of the standard view claim that implicit attitudes are based on conceptual or affective associations (e.g., Sarah's association between BLACK PERSON and DANGER or her association between BLACK PERSON

[11] It must be noted that the term "implicit bias" is ambiguous. It can denominate an output, such as a judgment, decision, or behaviour, that is implicitly biased or a mental state (or mental process) that is implicitly biased (Holroyd & Sweetman, 2016: 81-82). Used in this latter way, the term "implicit bias" may well refer to the same entities as the term "implicit attitude". However, in the philosophical literature at least, the term "implicit bias" is typically used for mental states with negative evaluative implications (see for example the articles in Brownstein and Saul, 2016a, 2016b). By contrast, the term "implicit attitude" is more broadly used for mental states that can have a positive or negative valence. In this thesis, I consistently use the term "implicit attitude" and not the term "implicit bias" as I am not exclusively concerned with negative evaluations.

[12] Although some authors have characterised implicit attitudes as unconscious (Greenwald & Banaji, 1995), it is increasingly recognised, even among proponents of the standard view that people can become aware of their so-called implicit attitudes (Levy, 2014b; Wilson et al., 2000; see section 1.2.5 in chapter 1). This is why I put "awareness" here in brackets and why I do not follow Machery (2016) in calling the view that there are distinct implicit and explicit attitudes "the Freudian view" (see chapter 4 in this thesis).

[13] Sometimes a person may possess an implicit attitude which content is perfectly in accordance with the content of her explicit attitude. Such conformity is according to proponents of the standard view a matter of coincidence rather than a matter of control that the subject has over her implicit attitudes.

[14] In chapter 1, I distinguish two kinds of control. People may lack control over the acquisition of an attitude (rational control) or over the activation of an attitude and its influence on behaviour (intentional control).

and a negative affective reaction), whereas explicit attitudes are based on propositional mental states (e.g., Sarah's belief that it is morally reprehensible to treat people different because of their skin colour). With respect to *Q3*, the standard view also provides us with a clear answer: implicit and explicit attitudes are not only based on mental states but are in fact to be identified with mental states (e.g., associative and propositional mental states, respectively). Accordingly, we may say that Sarah's association between BLACK PERSON and DANGER *is* an implicit attitude of hers and that her moral belief *is* an explicit attitude of hers. With respect to *Q1*, however, the answer of proponents of the standard view is not so clear. On the one hand, proponents of the standard view often speak of "dual attitudes" when speaking about evaluative conflicts between explicit and implicit attitudes (Wilson et al., 2000), which may suggest that people have a single implicit and a single explicit attitude towards the respective social group. On the other hand, the claim that attitudes can be identified with individual mental states may suggest that people can in fact have several implicit and several explicit attitudes towards the same social group. I elaborate further on this point in chapter 1, where I present the standard view in more detail. A detailed examination of the implications of the standard view, and the empirical evidence that supposedly supports it, is crucial for assessing its validity.

Here it shall already be mentioned that part of the appeal of the standard view stems from the fact that it seemingly fulfils *D1* (though see next section). As mentioned in the previous section, when we ascribe either a positive or a negative attitude to Sarah, we only pick out a part of what drives her responses towards black people. Yet, by ascribing both a negative implicit and a positive explicit attitude to Sarah we seem to provide a more holistic description of her psychology that helps us explain and predict her responses towards black people. Sarah is on the one hand likely to report that discrimination against black people is wrong, which we can predict on the assumption that she has a positive explicit attitude towards black people. On the other hand, Sarah shows subtle signs of discomfort in the presence of black people, which we can predict on the assumption that she has a negative implicit attitude towards black people.

The standard view may also seem to satisfy *D2* (though see next section). Recall that one may argue that Sarah is a self-identified egalitarian whose unintentional biases do not reflect on her moral character (see last section). By describing Sarah's egalitarian beliefs as explicit attitudes and her spontaneous responses towards black people as expressive of implicit attitudes, the standard view may thus capture accurately the distinction between those aspects of her psychology that form part of her moral character and those aspects that do not reflect on her moral character (Levy, 2014b, 2015, 2017a).

It is unclear in how far the standard view can satisfy desideratum *D3*. As I have mentioned, the standard view is the common conception of attitudes in the psychology and philosophy of prejudice. However, it should be noted that this conception of attitudes conflicts with the folk psychological conception of attitudes. When we ascribe attitudes to people in day-to-day life, we do not seem to pick out individual (implicit or explicit) mental states but seem to highlight general traits of people.

III. A preview of the argument to come

In the following chapters, I will scrutinise the standard view and argue that there is a better model of attitudes available. According to a plausible version of the standard view, implicit attitudes are associative mental states over which agents have only indirect control, while explicit attitudes are propositional mental states that are subject to direct control. Yet, I will argue that this is not the best way to construe attitudes. Important motivations for distinguishing between implicit and explicit attitudes do not hold up to scrutiny. Firstly, the psychometric evidence does not establish that there are indeed two distinct classes of attitudes. Secondly, evidence suggests that to optimally explain and predict people's evaluative responses, we do not actually need to distinguish between implicit and explicit attitudes (Oswald et al., 2013; see desideratum *D1*). Thirdly, it is misguided to assume that the distinction between so-called implicit attitudes and so-called explicit attitudes marks a distinction between mental states that form part of a person's moral character and mental states that do not form part of a person's moral character (see desideratum *D2*). However, the problem with the standard view of attitudes is not only that it is not well motivated. The standard view is also at odds (as already mentioned above) with the folk psychological conception of attitudes (see desideratum *D3*). When we ascribe, for example, a sexist attitude to a person, we do not normally mean to pick out a particular belief or association but rather a general trait of the agent. I will argue that there is in fact a scientifically sound model of attitudes available that is better aligned with the folk psychological conception of attitudes and more conducive to our explanatory/predictive and character evaluative purposes.

Regarding the question about the ontological status of attitudes (*Q3*), I claim that attitudes are traits of people that can be analysed as profiles of situation-specific evaluative response dispositions. Sarah, for example, can be said to possess an aversive racist attitude that consists of two situation-specific response dispositions: (1) the disposition to respond in a favourable manner towards black people in situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments, and (2) the disposition to respond in a negative

manner towards black people in situations in which she does not have sufficient time (e.g., when she has to judge quickly whether a person poses a threat to her) or cognitive resources (e.g., when she is distracted by the conversation with her patients) to reflect on and be guided by her endorsed egalitarian commitments.

In regard to the question about the mental states that underpin attitudes (*Q2*), I will argue that each attitude is grounded in a variety of distinct (implicit and explicit) mental states (see Machery, 2016, for a related view). Sarah's aversive racist attitude, for example, may be based on her belief that it is wrong to treat people differently because of their skin colour, her desire not to discriminate against black people, various associations (such as the association between BLACK PERSON and DANGER), her disposition to feel scared of black people, etc. On my proposed view attitudes are neither implicit nor explicit. The implicit-explicit distinction applies only to the mental states at the psychological basis of the attitude.

Concerning the question about attitude individuation (*Q1*), I will argue that there are different legitimate ways to individuate a person's attitude(s), which depend on our interests and purposes as attitude ascribers. Given my brief description of the case of Sarah, it may be salient that she has an aversive racist attitude as described above. Yet, it should also be noted that my description of Sarah's responses towards black people can only be incomplete. Sarah's evaluative responses towards black people may vary dependent on a myriad of contextual factors that we can hardly all keep track of. I argue that attitude ascribers often need to extract salient or especially noteworthy patterns from a person's more complex mesh of situation-specific response dispositions to give an intelligible account of that person's attitude(s). The process of extracting relevant response patterns is influenced by our interests and purposes as attitude ascribers. As our interests und purposes may differ, we may end up with different ways to individuate attitudes. These different ways to individuate attitudes are all legitimate as long as they track actual dispositions of the agent and thus help us to explain/predict the agent's responses and to convey accurate information about the agent's moral character.

IV. The structure of the thesis

This thesis is structured as follows. In chapter 1, I present those assumptions that motivate the standard view of attitudes and elaborate on how the standard view answers *Q1*, *Q2,* and *Q3.* This allows me to draw some initial conclusions about the extent to which the standard view satisfies the desiderata for a model of attitudes. In the first part of the chapter, I argue that the distinction between implicit and explicit attitudes can possibly be defended with reference to the following features: mental

structure (associative mental structure vs. propositional mental structure), rational control (reason-insensitivity vs. reason-responsiveness), and intentional control (automaticity vs. control). Awareness, by contrast, does not provide a feature that would allow us to distinguish implicit from explicit attitudes as recent findings suggest that people can become aware of their so-called implicit attitudes just as they can become aware of their so-called explicit attitudes (Gawronski, Hofmann, & Wilbur, 2006; Monteith, Voils, & Ashburn-Nardo, 2001; Hahn et al., 2014; Scaife et al., 2016). I also elaborate on what the standard view implies for conceptions of evaluative agency and a person's moral character (see desideratum *D2*). In short, explicit attitudes are generally assumed to form part of a person's moral character, while implicit attitudes do not. In the second part of the chapter, I show that the distinction between implicit and explicit attitudes is also assumed to correspond to two different ways to measure attitudes (i.e., indirect and direct measures of attitudes). However, I argue that divergences between people's responses on indirect and direct measures cannot prove that people possess distinct implicit and explicit attitudes, unless we already adopt a certain account of attitude individuation. Moreover, I discuss evidence that indicates that in order to optimally explain and predict people's evaluative responses towards other people (see desideratum *D1*), we may not actually need to postulate the existence of two distinct classes of attitudes that correspond to what is measured on indirect and direct measures of attitudes (Forscher et al., 2016; Oswald et al., 2013).

In chapter 2, I scrutinise some of the claims that proponents of the standard view have made about implicit attitudes (defined for the purposes of this chapter as those mental states that are measured on indirect measures of attitudes). I present a recent account by Mandelbaum (2016) according to which implicit attitudes are not, as usually assumed by proponents of the standard view, reason-insensitive associative mental states but in fact reason-responsive propositional mental states. I argue that Mandelbaum's argument fails. Even if we grant Mandelbaum that the evidence that he bases his argument on is evidence of propositionally structured implicit attitudes, this does not establish that all or the majority of implicit attitudes are propositionally structured. Moreover, there are alternative explanations available for the effects that Mandelbaum discusses that are consistent with an associative account of implicit attitudes. It follows that proponents of the standard view may be right that implicit attitudes are associative mental states, while explicit attitudes are propositional mental states. However, even on the assumption that implicit attitudes are associative mental states, it is not correct that implicit attitudes are completely outside of the subject's rational or intentional control. I emphasise that associative mental states are, at least to some extent, subject to indirect rational and indirect intentional control. Drawing on an argument by Holroyd & Kelly (2016), I further argue that this implies that implicit

attitudes can in fact form part of people's moral characters. This undermines one important motivation to draw the distinction between explicit and implicit attitudes. That is, the distinction between explicit and implicit attitudes fails to mark a relevant distinction between what belongs to and what does not belong to a person's moral character (see desideratum *D2*). Together with my conclusions from chapter 1, this suggests that the distinction between implicit and explicit attitudes is not well motivated. I also highlight that the standard view's identification of attitudes with individual mental states is out of line with the folk psychological conception of attitudes as traits. This may impede scholars' attempts to inform public discourse with their research (see desideratum *D3*). I thus propose to examine whether there is an alternative model of attitudes available that is better aligned with the folk psychological conception of attitudes as traits while still being scientifically sound.

In chapter 3, I turn to another distinction that is often made in regard to those mental states that are candidate (components of) attitudes: the distinction between stereotypes about and affect towards social groups (henceforth, "social affect"). Many scholars assume that this distinction is not only a conceptual distinction but that these concepts in fact correspond to distinct mental kinds (e.g., Amodio, 2008; Judd, Blair, & Chapleau, 2004; Valian, 2005). On this "two-type model", stereotypes, such as Sarah's association between BLACK PEOPLE and DANGER, can in principle occur independently of affective responses, such as Sarah's fear of black people (and vice versa). Other scholars have replied with a "one-type model" according to which stereotypes inherently possess an affective valence and social affect inherently possesses stereotypic conceptual content (Holroyd & Sweetman, 2016; Madva & Brownstein, 2016). On my proposed view, one-type theorists are right in so far as stereotypes about social groups and affects towards social groups form tight clusters (what Madva & Brownstein, 2016, call "evaluative stereotypes"). I show that the empirical evidence that proponents of the two-type view have brought forward cannot establish that stereotypes and social affect can operate independently of each other. Moreover, I point out that by focusing on the interactions between stereotypes and social affect we can yield better predictions of discriminatory behaviour than by focusing exclusively on either stereotypes or social affect. Yet, I also argue, contra Madva and Brownstein (2016), that the proposed clusters (the evaluative stereotypes) are not unified mental states but are composed of different kinds of mental states (e.g., conceptual mental states and affective mental states) that are causally closely linked to each other. Although this may appeal to some proponents of the two-type model, I also emphasise that the causal interconnectedness between conceptual and affective mental states makes it appropriate to say that stereotypes are affective and that social affect has a conceptual or stereotypic quality (which is a key claim of proponents of the

one-type model). As stereotypes and affect jointly drive people's responses towards other people qua members of social groups, it has to be acknowledged that both form part of people's attitudes (see desideratum *D1*). This provides a further answer to the question of what kind of mental states underpin attitudes (*Q2*).

In chapter 4, I develop my preferred model of attitudes. In short, I argue that attitudes are traits of people that can be analysed as profiles of evaluative response variation across situations (answer to *Q3*). I start out by discussing the recently proposed trait view of attitudes by Machery (2016) according to which attitudes "are broad-track dispositions to behave and cognize (have thoughts, attend, emote, and so on) toward an object […] in a way that reflects some preference" (p. 112). On this account, an attitude is based on a multitude of distinct mental states and processes (such as associative mental states, beliefs, emotions, and self-control processes; answer to *Q2*) and can be characterised in terms of an aggregate strength and valence. I highlight that the view that attitudes are traits is attractive because there are striking similarities in the explanatory, predictive, and character evaluative roles of trait and attitude ascriptions, and because the trait view of attitudes aligns well with the folk psychological understanding of attitudes. Yet, there is an objection that any view that holds that attitudes are traits must address. This is the situationist challenge according to which people's responses are largely determined by aspects of situations that they encounter and not by inner response dispositions of the kind that traits are usually identified with (e.g., Doris, 2002). I present a reply that is open to Machery (2016). As attitudes are characterised in terms of an aggregate strength and valence on his account, one may insist that the situationist argument merely establishes that attitudes are oftentimes relatively weak and not that there are no attitudes conceived as traits at all. However, this reply comes at a price. By characterising attitudes in aggregationist terms, Machery (2016) obscures attitudes' complex structure. His account masks evaluative conflicts and ambivalences, such as when people feel alienated by their own racist dispositions (see the case of Sarah) or exhibit both benevolent and hostile sexist tendencies. Moreover, his account does not do justice to relevant differences in the affective content of attitudes. I hold that my proposed model of attitudes, which describes attitudes as profiles of situation-specific evaluative response dispositions, both fends off the situationist challenge and does justice to the described evaluative complexities of attitudes. Based on Mischel and Shoda's (1995) influential cognitive-affective personality system model, I argue that it is misguided to assume that it speaks against the existence of attitudes understood as traits if people exhibit different evaluative responses towards members of a particular social group in different situations. Quite to the contrary, I take it to be a defining feature of attitudes that they are composed of situation-specific response dispositions. However, as agents usually

exhibit innumerable situation-specific evaluative response dispositions in regard to a single social group, we need to read off the most relevant response patterns if we want to give an intelligible account of a person's attitude. I discuss several ways in which the process of highlighting relevant response patterns (i.e., highlighting profiles of situation-specific response dispositions) is influenced by the attitude ascribers' interests and purposes. As the attitude ascriber's interests and purposes may differ to some extent, there are different legitimate ways to individuate attitudes (answer to *Q1*).

In chapter 5, I present a possible objection against my proposed profile view of attitudes. My account implies that an evaluative response disposition that an agent does not identify with (henceforth "non-endorsed disposition") may nevertheless be partly constitutive of an attitude of that agent. For example, Sarah's disposition to show negative responses towards black people when she does not have sufficient time and cognitive resources to reflect on her endorsed egalitarian commitments may form part of her attitude towards black people, even though she condemns racism and does not want to behave in a negative manner towards black people. Proponents of so-called real self theories may find this implication untenable because they hold that only those dispositions that the agent identifies with or that conform to the agent's considered values and rational judgments constitute the persons "real self" for which she is morally evaluable (e.g., Frankfurt, 1971; Stump, 1988; Velleman, 1992; Watson, 1975). On this view, my model of attitudes may violate desideratum *D2* because it is not appropriately sensitive to the difference between mental states that can rightly be said to be constitutive of a person's moral character and those mental states that are not part of a person's moral character. I reply that the real self perspective is unconvincing for a number of reasons. I show that we in fact routinely take both endorsed and non-endorsed evaluative response disposition into account when we evaluate the moral character of other persons, which is at odds with the real self account. I grant that some people (although not all people) are happy to accept that non-endorsed response dispositions do not reflect on their moral character. Yet, I insist that this is likely the result of a self-serving bias rather than an honest assessment. The real self view allows us to create a positive self-image because we can regard problematic response dispositions that conflict with our values as external to who we really are. However, the fact that the real self view helps us to feel good about ourselves is not a good reason to believe that this view should be adopted. Quite to the contrary, I regard the real self-perspective as problematic because we are less likely to do something against problematic response dispositions that harm others if we believe that these dispositions do not reflect negatively on us as persons. This leads me to a pragmatic argument for the inclusion of non-endorsed response dispositions in our model of attitudes. By including non-endorsed response dispositions in our conception of attitudes, we can

encourage the perception that these dispositions reflect on our moral character and make it thus more likely that we will tackle problematic biases.

Chapter 1: The standard view

1.1 Introduction

In the introduction to this thesis, I mentioned that there is a conception of attitudes that deserves to be called "the standard view" since it is predominant in both the contemporary philosophical and psychological literature on attitudes. This is the view that people possess distinct implicit and explicit attitudes.[15] Following on from my earlier example, Sarah may be said to have a positive explicit attitude towards black people, which is reflected in her favourable deliberate responses in regard to black people, and a negative implicit attitude towards black people, which is reflected in her problematic spontaneous responses in regard to black people.

In this chapter, I will describe in more detail how proponents of the standard view have characterised implicit and explicit attitudes and will elaborate on those psychological measurement procedures that supposedly identify people's implicit and explicit attitudes. This investigation will reveal how proponents of the standard view answer (and in part fail to answer) those questions about the nature of attitudes that were introduced in the introduction to this thesis:

(Q1) How should we individuate attitudes?
(Q2) What mental states underpin attitudes?
(Q3) What is the ontological status of attitudes?

Moreover, my investigation will allow some initial conclusions about the extent to which the standard view satisfies the desiderata for a model of attitudes that were mentioned in the introduction to this thesis:

(D1) To optimally fulfil its explanatory and predictive function, our notion of a person's attitude towards group X must pick out exactly those features of that

[15] See Fazio (1990, 2007) for a somewhat different perspective. Fazio does not distinguish between implicit and explicit attitudes, but what he describes as "attitude" corresponds roughly to what proponents of the standard view would call "implicit attitude". According to Fazio (2007), attitudes are "associations between a given object and a given summary evaluation of the object" (p. 608) that can become "activated automatically from memory" (p. 610). On his view, people's spontaneous evaluative responses are a function of these attitudes (Fazio et al., 1995). People's deliberate evaluative responses, by contrast, are often (but not necessarily) influenced by other mental states and processes (such as a person's moral beliefs or self-presentational motives) besides the attitude. On Fazio's model, these other mental states and processes are not attitudes (or constituents of attitudes).

person's psychology that drive that person's evaluative responses towards group X.

(D2) To optimally fulfil its role in character assessment, our notion of a person's attitude towards group X should be sensitive to any difference that there may be between aspects of that person's psychology that can rightly be said to be constitutive of that person's moral character and those aspects that are not part of that person's moral character.

(D3) To facilitate communication on attitudes between academic disciplines as well as between academia and the wider public, our notion of a person's attitude towards group X should ideally be a notion that psychologists, philosophers, and ordinary people can agree on

This chapter has two parts. The first part (section 1.2) is concerned with the question of how implicit and explicit attitudes are characterised by proponents of the standard view. It will become clear that the standard view has strongly been influenced by dual-process models of cognition in psychology. In section 1.2.1, I will show that it is common among proponents of the standard view to identify explicit attitudes with propositional mental states, whereas implicit attitudes are commonly identified with associative mental states. In section 1.2.2, I will argue that this provides clear answers to the question of what mental states underpin attitudes (*Q2*) and the question as to the ontological status of attitudes (*Q3*). However, proponents of the standard view have largely neglected the question of attitude individuation (*Q1*). In section 1.2.3, I will elaborate on how the alleged distinction between (associative) implicit attitudes and (propositional) explicit attitudes relates to philosophical accounts of evaluative agency, rational control, and moral character. In section 1.2.4, I will show that other scholars have linked the distinction between implicit and explicit attitudes to a distinction between automatic and controlled attitude activation. In section 1.2.5, I will then elaborate on the claim that implicit and explicit attitudes are distinguished by the fact that the former are unconscious, while the latter are consciously accessible. The upshot will be that the distinction between implicit and explicit attitudes can possibly be defended by reference to mental structure, rational control (reason responsiveness), and/or intentional control, but that consciousness does not provide a reasonable criterion to draw this distinction.

The second part of this chapter (section 1.3) is about the psychological measurement procedures that are supposed to reveal implicit and explicit attitudes. In section 1.3.1, I will present some examples of direct measures of attitudes that are supposed to reveal explicit attitudes and of indirect measures of attitudes that are supposed to reveal implicit attitudes. In section 1.3.2, I will argue that dissociations

between people's results on indirect and direct measures of attitudes do not provide proof for the claim that people possess distinct implicit and explicit attitudes unless we already adopt a certain account of attitude individuation. In section 1.3.3, I will discuss recent meta-analyses that indicate that both indirect and direct measures of attitudes are relatively poor predictors of people's evaluative responses towards other people. I take these results to suggest that we may not actually need to postulate the existence of distinct classes of implicit and explicit attitudes (a measured on indirect and direct measures of attitudes) in order to optimally explain and predict people's evaluative responses.

1.2 The nature of attitudes on the standard view

The standard view of attitudes holds that there are two distinct kinds of attitudes. This view has strongly been influenced by dual-process models of cognition (Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Strack & Deutsch, 2004). Although dual-process models differ in many details, there is a substantial overlap in the features that individual scholars ascribe to the two alleged classes of processes (that may or may not be claimed to operate in two different types of cognitive system). One type of process (often said to operate in "system 1") is described as associative, automatic, unconscious, effortless, independent of attentional resources, fast, and impulsive. The other type of process (often said to operate in "system 2") is typically described as propositional, rule-based, controlled, conscious, effortful, attention demanding, slow, and reflective (Frankish & Evans, 2009; Kahneman, 2012; Sloman, 1996; Smith & DeCoster, 2000).[16] Processes of the former kind are often described as "implicit", while processes of the latter kind are commonly referred to as "explicit". Influenced by this general framework of cognition, it has become common, both in social psychology and in the philosophy of prejudice, to distinguish between implicit and explicit attitudes. It has been claimed that different kinds of mental states – mental states that allow for associative and rule-based processing – underlie implicit and explicit attitudes respectively (see section 1.2.1 and 1.2.2), that implicit attitudes are insensitive to reasons while explicit attitudes are reason-responsive (see section 1.2.3), that implicit attitudes operate in an automatic mode while explicit attitudes are controlled (see

[16] The terms "system 1" and "system 2" were introduced by Stanovich (1999). This terminology highlights the assumption that people's minds are divided into two distinct cognitive systems that give rise to two distinct kinds of cognitive processes. See for example Frankish and Evans (2009), table 1.1, or Sloman (1996), table 1, for summaries of the properties that are ascribed to the two systems. It shall be noted, however, that we may be able to distinguish two different kinds of cognitive processes without these processes issuing from distinct cognitive systems (e.g., Gawronski & Bodenhausen, 2006). That is, dual-process theories are not necessarily dual-system theories.

section 1.2.4), and that implicit attitudes are unconscious while explicit attitudes are conscious (see section 1.2.5).

1.2.1 Implicit attitudes as associative and explicit attitudes as propositional mental states

As mentioned above, on predominant dual-process views in psychology there is a distinction to be made between associative and rule-based (or propositional) processes. It is often claimed that the distinction between implicit and explicit attitudes maps onto this distinction: implicit attitudes are assumed to operate in an associative manner, while explicit attitudes are taken to operate in a rule-based (or propositional) fashion (Gawronski & Bodenhausen, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004).[17] As I will show in this section, this suggests an answer to the question as to what kind of mental states underpin attitudes (*Q2*). In the next section, I will then elaborate on what this implies for the ontological status of attitudes (*Q3*) and the question of attitude individuation (*Q1*).

We can distinguish between association as a mental process and association as a mental state. The mental process is the time-dependent spreading of activation from one mental representation to associated mental representations. Associative processes presuppose the existence of associative networks of representations – what I call associative mental states. Implicit attitudes are understood to be associative mental states of this sort. That is, they are commonly understood to be underpinned by networks of associatively linked mental representations (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011).[18] Commonly, the relevant representations are supposed to be concepts. Sarah from our example above may associate the concept BLACK PERSON with the concept DANGER. It is also often assumed that representations of positive or negative affective valence take part in these associations (De Houwer, 2014: 343; Mandelbaum, 2016: 630). For example, the concept BLACK PERSON may be associatively linked to negative affect in Sarah (and to the concept

[17] See also De Houwer (2014), Levy (2015), Mandelbaum (2016), and Stammers (2016: chapter 4) for characterisations of the difference between associative and propositional mental states and processes in evaluation.

[18] Hughes and colleagues (2011) conclude their review of dominant models of implicit attitudes in psychology with the statement that these models share "the pre-analytic belief that implicit attitudes should be understood largely in terms of the formation, activation, and change of associations between mental representations" (p. 471). The perspective that implicit attitudes are associative mental structures is certainly the standard view in the psychological literature. To name just a few examples, Gawronski and Bodenhausen (2011) claim that "implicit evaluations are the behavioral outcome of associative processes" (p. 1), Rydell and McConnell (2006) understand implicit attitudes as the products of "a slow learning, associative system of reasoning" (p. 1006), and Amodio and Devine (2006) argue that implicit evaluation and implicit stereotyping are based on affective and semantic associations, respectively.

DANGER).[19] Saying that representations are associatively linked implies that if one representation becomes activated (e.g., the concept BLACK PERSON), its activation spreads over to those representations to which it is linked (e.g., to the concept DANGER and to negative affect; Collins & Loftus, 1975). Associative links can be unidirectional (e.g., when the activation of BLACK PERSON spreads over to DANGER *but not* vice versa) or bidirectional (e.g., when the activation of BLACK PERSON spreads over to DANGER *and* vice versa; Cox & Devine, 2015).[20] Activation spreading is assumed to happen fast, effortless, without reflective control and independent of attentional resources – which is why it is often attributed to system 1 (see also section 1.2.4 on the automaticity-control contrast). The connection strength between the representations determines how strongly the activation of one representation will affect the activation of connected representations. We may distinguish between occurrent and dispositional associative mental states. An association is occurrent when the representations that constitute the association are momentarily co-activated. However, even when the association is not currently activated, we may say that the association is present in a dispositional sense. By this I mean that due to the link between the representations, the representations have the propensity to become co-activated. For example, when we say that Sarah associates BLACK PERSON with DANGER, we do not normally mean that this association is currently activated but that it will become activated when she encounters or imagines a black person.

It is generally assumed that associative mental states can only be changed over multiple experiences (De Houwer, 2014: 342; Mandelbaum, 2016: 632-635; Stammers, 2016: 103-104).[21] Encountering a black person who is clearly not dangerous will not do much to weaken Sarah's association between BLACK PERSON and DANGER, but if Sarah repeatedly encounters black people who clearly pose no threat to her and if she is not confronted with any negative representations of black people (e.g., in the media or in conversation with other people) for a while, her association may weaken over time. The assumption that implicit attitudes have associative structure is often appealed to in order to explain why it is so difficult to change people's implicit attitudes. Sarah's tendency to keep more distance to black patients than to white patients persists despite

[19] In chapter 3, I will elaborate on the question of how conceptual associations (such as Sarah's association between BLACK PERSON and DANGER) are related to affect (the anxiety that Sarah experiences). I will argue that these kinds of mental states are causally tightly linked and jointly contribute to people's attitudes.

[20] Henceforth, when I simply speak of „association", I mean associative mental states as they are characterised here.

[21] It should be noted, however, that associations can presumably be *acquired* by a single experience. A clear case in point is taste aversion (Mandelbaum, 2016: 633-634). If you get seriously sick after eating a tomato, you will acquire a strong association between the taste of tomatoes and the feeling of sickness. Note that once acquired, it will take several positive experiences with tomatoes to get rid of your distaste for tomatoes. Acquiring associations is generally much easier than getting rid of them.

her belief that this tendency is problematic. This can be explained by the fact that Sarah's implicit attitude is associative and thus can only be changed by changing certain external contingencies (Levy, 2014a: 99-100).

Just like associative mental states, propositional mental states are understood to be composed of abstract mental representations. Yet, unlike associations, propositional mental structures possess a language-like syntax and a truth value. If one associates BLACK PERSON with VIOLENCE, this does not imply any particular relation between these two concepts. By contrast, a propositional structure containing BLACK PERSON and VIOLENCE, such as a belief with the content "black persons are violent", specifies the relation between these concepts.[22] Violence is here described as an attribute of black persons (see also De Houwer, 2014: 344-345, and Stammers, 2016: 99-100). The content of propositional mental states is assumed to be compositional. That is, their content is a function of the content of the constituent concepts and the syntax that links up these concepts (Margolis & Laurence, 2007: 562). Due to their internal structure propositional mental states can feature in inferential transitions. For example, if one holds the belief (or as we may want to say the attitude) "all black persons are violent" and the belief "John is a black person", it is rational to infer that "John is violent". We may call such inferences over propositional mental states "propositional processes".

Unlike associations, propositional mental states can be (but are not necessarily) eradicated by single experiences (De Houwer, 2014: 344; Mandelbaum, 2016: 635-636; Stammers, 2016: 103-104). For example, someone who holds the belief that all Germans are industrious may possibly (although not necessarily) eradicate this belief upon encountering a lazy German person. By contrast, such an experience will do little to change an established association between Germans and industriousness. It is thus assumed that implicit attitudes qua being associative are difficult to change by one-off interventions, while explicit attitudes are much more amenable to such interventions due to their propositional structure (Levy, 2014a: 99-100). I will return to this point in section 1.2.3, in which I discuss the implications of the association-proposition distinction for philosophical accounts of evaluative agency, rational control, and a person's moral character. Before I turn to that, it is worth examining what answers to

[22] I am assuming here a representationalist account of believing, according to which a person's beliefs can be identified with particular mental representations (that have the right internal structure or that stand in appropriate causal relations with other mental states). By contrast, proponents of non-representationalist accounts of believing (dispositionalists, interpretivists, and also some functionalists), would deny that particular representational structures instantiate beliefs (Schwitzgebel, 2015). Most proponents of these alternative accounts of believing would grant that there are mental states with propositional structure but deny that any of these states underwrites believing (or desiring). I invite adherents of these accounts to substitute any reference to "belief" in what follows with "propositionally structured mental state".

the questions concerning attitude individuation (*Q1*) and the ontological status of attitudes (*Q3*) the standard view provides.

1.2.2 The ontological status of attitudes and attitude individuation on the standard view

It has to be stressed that proponents of the standard view describe implicit and explicit attitudes not only in a way that suggests that they are *based* on (associative and propositional) mental states. Rather implicit and explicit attitudes are characterised in a way that suggests that they are in fact *to be identified* with (associative and propositional) mental states. This is apparent from the fact that proponents of the standard view describe attitudes as entities that can occur, be activated, be introspected and be retrieved (e.g., Levy, 2014b; Wilson, Lindsey & Schooler, 2000; see also Machery, 2016: 107-108). That is, they characterise attitudes as having a comparable ontological status to mental states such as beliefs, desires, intentions and emotions. The standard view thus provides not only an answer to the question as to what kind of mental states underpin attitudes (*Q2*) but also to the question as to the ontological status of attitudes (*Q3*). Note that in principle a psychological construct could be based on mental state(s), without being a mental state. For example, it has been claimed that attitudes are traits, which are based on a variety of different mental states, without being identical to them (Machery, 2016).[23] This is not what proponents of the standard view have in mind. For them attitudes are clearly to be identified with mental states. Levy (2015), for example, posits that "[i]mplicit attitudes are *mental states* that appear sometimes to cause agents to act in ways that conflict with their considered beliefs" (p. 800, my emphasis).

Despite providing clear answers to *Q2* (the question as to what kind of mental states underpin attitudes) and *Q3* (the question of the ontological status of attitudes), proponents of the standard view have largely neglected the question of attitude individuation (*Q1*). This is problematic because if it remains unspecified how we should individuate attitudes, we do not have a way to decide how many attitudes a given individual has (or can have) towards a given social group. In particular, we cannot say whether people can have attitudes that conflict with each other if we do not know how to individuate attitudes. In what follows, I elaborate on what the standard view entails with respect to *Q1*.

The emphasis that is often put on the claim that people possess "dual attitudes" towards a particular social group seems to reflect the assumption that people

[23] In fact, I will defend a version of this view in chapter 4.

(normally) have *exactly* one implicit and one explicit attitude towards that group. Consider for example this passage from Wilson and colleagues (2000):

> We propose that people can have dual attitudes, which are different evaluations of the same attitude object, one of which is an automatic, implicit attitude and the other of which is an explicit attitude. (p. 102)

This passage suggests the existence of *just* two attitudes, one implicit and one explicit, towards the attitude object.[24] However, this simple perspective on attitude individuation is in conflict with the claim that attitudes are mental states, as can be seen when we consider once more the case of Sarah. With what would we identify Sarah's implicit and her explicit attitude if she really had just one of each? To be sure, we may say that her implicit attitude towards black people is an association between the concept BLACK PERSON and the concept DANGER and that her explicit attitude is an anti-racist belief, such as the belief that black people do not pose a threat to her. However, this is an overly simplistic picture of Sarah's psychology. It is implausible to assume that Sarah associates only one attribute with black people. It is much more likely that she will associate black people with a range of stereotypical attributes, some of which may even have a positive valence (e.g., the concept MUSICALITY). Moreover, she will likely hold a range of different beliefs about black people that are expressive of an evaluation of black people (e.g., the belief that it is wrong to treat black people any different to white people, the belief that black people pose no threat to her, etc.). It would be unduly arbitrary to identify her implicit attitude with any particular associative link (e.g., the link between BLACK PERSON and DANGER) and to pick out one particular belief as her explicit attitude.

In response, one may suggest that implicit and explicit attitudes are mental states with a complex structure. One could say that her explicit attitude is the entirety of her evaluative beliefs about black persons. However, note that although each of Sarah's beliefs is clearly a mental state, the entirety of her beliefs about black people is certainly not a mental state. This is because her beliefs about black people can be tokened independently of each other. In some situations her behaviour may be guided by her belief that black people pose no threat to her (e.g., when she encourages young black people to study medicine) and in other situations her belief that it is wrong to treat black people any different to white people may be tokened (e.g., when she notices her inclination to keep more spatial distance and to make less eye contact with black patients than with white patients). In short, a set of beliefs fails to constitute a mental

[24] Wilson and colleagues (2000) acknowledge at the end of their article the possibility of multiple implicit and explicit attitudes, each of which is tied to a particular context. Nonetheless, their initial description of the dual attitude account is representative of how many scholars speak about the implicit-explicit distinction.

state because we should expect that components of a mental state always co-occur. Identifying Sarah's explicit attitude with the entirety of her endorsed evaluative beliefs about black people is thus incompatible with identifying her explicit attitude with a mental state. If attitudes are to be identified with mental states (as proponents of the standard view claim), we need to acknowledge that Sarah has at least as many explicit attitudes towards black people as she has evaluative beliefs with regard to black people.[25]

Similarly, we have to acknowledge that Sarah has multiple *implicit* attitudes towards black people if we want to maintain that implicit attitudes are to be identified with mental states. Relying on research on object representation, Gawronski & Bodenhausen (2011) note that "the same attitude object may activate different patterns of associations in memory depending on the particular context in which the object is encountered" (p. 62). For example, Sarah's association between BLACK PERSON and DANGER may become activated whenever she is approached by a black person on the street but her association between BLACK PERSON and MUSICALITY may be tokened whenever she encounters black people at concerts. The fact that Sarah's association between BLACK PERSON and DANGER and her association between BLACK PERSON and MUSICALITY can, in principle, be tokened independently of each other, indicates that they are separate mental states.

To sum up, proponents of the standard view often speak about attitudes in a way that suggests that people (usually) have *only* one implicit and one explicit attitude towards a given social group. However, as I have shown above, this is incompatible with the claim that attitudes are mental states because people harbour a multitude of relevant mental states with respect to a given social group. If implicit attitudes are associative mental states and if explicit attitudes are propositional mental states, people will likely have a multiplicity of implicit and explicit attitudes towards a given group. As the claim that attitudes are mental states (entities that can occur, be activated, be introspected, and be retrieved) is essential to the standard view, proponents of the standard view should accordingly acknowledge that people can have multiple implicit and explicit attitudes.

A motivation for the claim that people possess dual-attitudes may be that this accounts for the conflict that people often experience between an endorsed commitment (i.e., an explicit attitude) and an automatic association (i.e., an implicit attitude). However, it needs to be stressed that conflicts of this sort can also be accounted for by an account on which people possess multiple implicit and multiple

[25] I say "at least" because explicit attitudes can arguably not only be identified with beliefs but also with propositional mental states of other kinds, such as desires. For example, Sarah may have the desire to interact with black people in the same way as she would interact with white people.

explicit attitudes. On such an account, a particular implicit attitude may be in conflict with one or several explicit attitudes (but maybe not all explicit attitudes). Conversely, one particular explicit attitude may be in conflict with one or several implicit attitudes (but maybe not all implicit attitudes).[26] Attitudinal conflict is thus compatible with a multiple attitude account and does not carry any weight in favour of a dual-attitude account. I therefore suggest that the most defensible version of the standard view implies that people can possess several implicit and several explicit attitudes towards a particular social group.

It shall already be mentioned that this view of attitudes is at odds with how the attitude notion is used in day-to-day discourse. For example, when we say that someone has a negative attitude towards immigrants, we do not seem to refer to an individual mental state (e.g., a belief or an association) of the agent. Rather we want to express that the agent is generally disposed to respond in a negative way towards immigrants. In short, we refer to a generic trait of the agent. The fact that the predominant view of attitudes in the psychological and philosophical literature on attitudes (the standard view) is so far detached from this folk psychological understanding of attitudes is worrying. As mentioned in the introduction to the thesis, scholars in philosophy and in psychology will find it difficult to inform public discourse on such important issues such as discrimination if their notion of an attitude does not correspond, at least roughly, to how ordinary people use the term. However, before I explore whether there is an alternative model of attitudes available that better corresponds to the folk psychological notion of attitudes (while still being of use to psychologists and philosophers; see desideratum *D3* of a model of attitudes), I will continue with my review of the standard view. We first need to understand the implications of the standard view before we can examine whether there is a more appropriate model of attitudes available.

1.2.3 Implications of the association-proposition distinction for evaluative agency, rational control, and a person's moral character

In the introduction to this thesis, I have mentioned that the notion of an attitude should be sensitive to the difference between aspects of an individual's psychology that can rightly be said to be constitutive of that person's character and those aspects that are not part of her character (if there is indeed such a difference; desideratum *D2*). As I will show in this section, by distinguishing between associative implicit attitudes and

[26] Also, there may of course be conflicts among the implicit attitudes and among the explicit attitudes of an agent. Yet, one should assume that conflicts between different explicit attitudes will normally be resolved fairly quickly by the person. See next section for Levy's (2014b) notion of the "unification of the person" (p. 35).

propositional explicit attitudes the standard view may possibly fulfil this criterion. To show this, I will elaborate on how associative and propositional mental states are usually understood to relate to an agent, and link this to the notion of rational control.

It has been argued that (certain) propositional mental states, such as beliefs, are agential mental states, while associative mental states do not have the right structure to be attributed to an agent (e.g., Gendler, 2008a, 2008b; Levy, 2014a).[27, 28] The standard view thus has implications for how we conceive of moral agency. Levy is presumably the philosopher who has argued most extensively for the relevance of the implicit-explicit distinction on the basis of an account of agency (Levy, 2011, 2014a, 2014b, 2015, 2017a). According to Levy (2014b), being an agent crucially involves the capacity to pursue projects over time, which in turn presupposes what he calls "the unification of the person":

> [A]gency depends upon unification of the person; an agent has a relatively consistent set of beliefs and desires, and is able to ensure that she acts upon those beliefs and desires. This unification depends upon explicit attitudes, I propose, because only such attitudes can cause broad and integrated behaviors. Explicit attitudes are employed by the agent to impose unity effortfully, by being taken as premises in reasoning, and through their role in coordinating plans and projects. Pursuit of plans and projects requires rule-based processing, not associative. (p. 35)

Levy argues here that rule-based (and thus norm-driven) reasoning over propositions is necessary to achieve consistency among one's mental states and in one's conduct. Contradictions among one's propositionally structured beliefs and desires can be resolved by reasoning, leading to an integrated set of mental states, which forms the basis for coherent behaviour over time. By contrast, associations cannot play this integrative role in a person's agency according to Levy (2014b). As associations are not able to feature in inferences, they cannot be brought into line with other mental states by reasoning. They simply track whatever contingencies between stimuli are present in a person's environment, irrespective of what the person believes or desires. That is, rather than contributing to an agent's projects and plans, associations often prevent the realisation of agential behaviour according to Levy.

Levy (2014a: 99-100) links his conception of agency to the notion of rational control. With reference to Gendler (2008b), he argues that implicit attitudes are not

[27] See Stammers (2016: chapter 2) for an extensive review of the literature on this, what she calls, "substantial distinction view".

[28] Gendler (2008a, 2008b) calls these associative mental states "aliefs" to contrast them with beliefs. Aliefs have according to Gendler representational, affective, and behavioural components. Gendler (2008a) mentions incidentally that the representational component of an alief may represent state of affairs "perhaps propositionally, perhaps nonpropositionally, perhaps conceptually, perhaps nonconceptually" (p. 643). Although she raises the possibility that the representational component of an alief may have propositional structure in this statement, she insists that aliefs are not reason-responsive and not subject to intentional control. While I focus on Levy's model at this stage, I will get back to Gendler's model in chapter 3.

responsive to reasons. By this he means that their acquisition and change is not a function of what the agent takes to be facts that justify their acquisition and change but rather of mere regularities in a person's environment. Whereas an agent's beliefs and desires (which may form the basis of her explicit attitudes) can be updated in accordance with what the agent judges to be good reasons, implicit attitudes are not subject to such rational modification due to their associative structure. Levy (2014a) notes accordingly that implicit attitudes do not belong "to the class of judgment-dependent attitudes" (p. 99). This implies that implicit attitudes can be acquired and persist, even though the agent judges their content to be factually wrong or to be morally problematic. In a later article, Levy (2015) admits that implicit attitudes may have some propositional structure but maintains that "[t]hey do not feature often enough and broadly enough in the kinds of normatively respectable inferential transitions that characterize beliefs" (p. 816).[29] He argues that they form a *sui generis* class of mental states which he calls "patchy endorsements". As patchy endorsements lack the inferential promiscuity of beliefs (i.e., the ability to interact in a normatively appropriate way with any other propositional state), they are not sufficiently reason-responsive (i.e., judgment-dependent) to contribute to the unification of the person (see also Levy, 2017a).

Levy's (2014a, 2015, 2017a) view that implicit attitudes often compromise agential behaviour because we lack rational control over them is widely shared among philosophers (Gendler, 2008a, 2008b; Glasgow, 2016; Zheng, 2016). On this view, there is a fundamental difference between non-agential implicit attitudes and agential explicit attitudes, which is marked by the reason-insensitivity of the former and the reason-responsiveness of the latter attitudes. This distinction may have important implications for our assessment of a person's moral character (and, on some accounts, also for moral responsibility).[30] Consider again the case of Sarah. Sarah's tendency to keep more spatial distance to black than to white interlocutors may plausibly not be subject to rational control. After all, this tendency persists despite the fact that Sarah believes it to be wrong to treat people differently because of their skin colour. Accordingly, proponents of the standard view could say that her tendency to keep

[29] See next chapter for an extensive discussion of the propositional account of implicit attitudes.

[30] Watson (2004), for example, contrasts two forms of moral responsibility: Responsibility as attributability and responsibility as accountability (see Fischer & Tognazzini, 2011, and Zheng, 2016, for related distinctions). Whether someone is responsible for a response in the attributability sense depends on the relation between the response and the agent: individuals are responsible for those thoughts and behaviours that are attributable to them as reflections of their agency. The distinction between associative and propositional mental states is thus arguably relevant for this form of moral responsibility. By contrast, responsibility in the accountability sense depends on the relation of the agent to her moral community. The agent is accountable for her conduct if others have justifiable expectations of how she should behave. Scholars are divided upon the question as to whether accountability for an action necessarily implies attributability of that action (see Zheng, 2016: footnote 3).

excessive distance from black interlocutors does not tell us anything about her moral character. On this view, we should judge Sarah rather by her endorsed beliefs, which are subject to rational control.

To sum up, I showed in this section that some proponents of the standard view claim that implicit attitudes are not attributable to an agent, while explicit attitudes are agential mental states. This is because implicit attitudes are allegedly not (sufficiently) responsive to reasons and can thus not contribute to the unification of a person. Explicit attitudes, by contrast, are reason-responsive (i.e., under the subject's rational control) due to their propositional structure. The standard view's distinction between implicit and explicit attitudes may thus nicely capture the distinction between those evaluative mental states that are not part of a person's moral character (reason-insensitive mental states) and those mental states that are reflective of a person's moral character (reason-responsive mental states). If this is correct, the standard view fulfils the second desideratum of a model of attitudes (*D2*). That is, it is sensitive to the difference between aspects of an individual's psychology that can rightly be said to be constitutive of that person's character and those aspects that are not part of her character. However, in the next chapter (see especially section 2.3), I will argue that so-called implicit attitudes can reflect on a person's moral character after all. This is because implicit attitudes are subject to *indirect* rational control (and *indirect* intentional control).

1.2.4 Implicit and explicit attitudes as automatic and controlled mental states

In the last section, I have reviewed Levy's (2014a, 2015) view, which distinguishes implicit and explicit attitudes by reference to rational control. According to this, explicit attitudes are acquired and changed in accordance with what the agent considers to be good reasons, but implicit attitudes are insensitive to such reasons. However, it must be noted that other scholars refer to another form of control that allegedly enables us to distinguish implicit from explicit attitudes (e.g., Devine, 1989; Rydell & McConnell, 2006; Wilson et al., 2000). This second notion of control has been inspired by psychological research on automatic and controlled processing (e.g., Shiffrin & Schneider, 1977). In short, the idea is that controlled processes require substantial attentional resources (i.e., are non-efficient) and are voluntarily initiated and sustained (i.e., they require an intention), while automatic processes occur without the subject's intention and attentional focus (i.e., they are efficient) and are difficult to suppress

(Gawronski & Payne, 2010).[31, 32] We may refer to this as a difference in "intentional control".

A classic example of a task in which automatic and controlled processes compete is the so called Stroop task (Stroop, 1935; MacLeod, 1997). In a typical Stroop experiment participants are asked to name as quickly as possible the colour of the ink of a colour word, whilst there is a mismatch between the ink colour and the meaning of the word (e.g., the word "green" printed in red ink). It has been found that participants respond much more slowly in this task than in a control task in which they are asked to name the colours of solid squares (Stroop, 1935, experiment 2). That is, when colour and meaning of word are incompatible, participants find it hard not to read out the word instead of naming the colour. This finding is commonly taken to show that reading is an automatic process that interferes with the colour naming task. While the colour naming requires the participant's attention, reading of the words proceeds without attention. Participants must selectively attend to the colour of the words in order to be able to successfully complete the task, whereas reading does not demand (many) attentional resources and proceeds without the subject's intention to read the words (in fact, it proceeds despite the subject's intention *not* to read the word).

Building upon this and related research on control and automaticity in cognitive psychology, social psychologists began to develop and to test accounts of automatic and controlled attitude (and stereotype) activation (Devine, 1989; Dovidio, Evans, & Tylor, 1986; Fazio et al., 1995).[33] The key idea is that some evaluations of a social group (usually understood as associations) are activated automatically from memory (and influence behaviour in an automatic manner) whenever a member of the social group is present (implicit attitudes), while other evaluations, such as those implied by endorsed beliefs about the social group, must be intentionally retrieved by the subject (explicit attitudes; Wilson et al., 2000). It is widely assumed that those evaluations (or

[31] The relation between the intentionality of a process and the attentional resources that the process requires is complex. On the one hand, it is arguably true that highly attention demanding processes require an intention to be sustained. On the other hand, It must be noted that processes that are initiated (and guided) by an intention do not necessarily require (many) attentional resources. For example, tooth brushing is clearly an intentional activity (those mental processes that guide my tooth brushing are intentional) but does not require (much of) my attention. I can focus my attention on a demanding task (e.g., adding up numbers) while I brush my teeth without necessarily compromising my tooth brushing performance. Thus, when psychologists claim that controlled processes require attentional resources and are voluntary initiated, this should not be taken to imply that these features necessarily go together. The claim seems rather to be that a *prototypical* controlled process combines these features. Similarly, we can say that a prototypical automatic process occurs without the subject's intention and her attentional focus, without committing us to the claim that unintentionality of a processes and low attentional requirements of a process always go together.

[32] Bargh (1994) uses the term "automatic" more broadly to denote processes that are non-intended, outside of awareness, non-controllable, or efficient and emphasises that these features, although often being linked to each other, do not necessarily co-occur.

[33] See chapter 3 for an elaboration on how stereotypes relate to attitudes.

stereotypes) that are activated automatically (e.g., Sarah's association between BLACK PERSON and DANGER) have been learned over repeated experiences. Devine (1989), for example, argued that culturally prevalent negative stereotypes of black people become ingrained in memory due to numerous encounters with representations of them throughout a person's lifetime. She showed that even people with explicitly non-prejudiced commitments are influenced by these automatically activated stereotypes when judging the hostility of an ambiguously described black person (Devine, 1989, study 2). As with the controlled inhibition of the automatic response on the Stroop task, overriding the influence of automatically activated evaluations with a different evaluative response is supposed to require effort and substantial attentional resources. Sarah, for example, may need to focus on her anti-racist commitments in order to be able to override her unintentional habitual response to keep excessive spatial distance from black people.

It can be argued that Sarah's inclination to keep excessive spatial distance from black people is not reflective of her moral character because it is an unintentional response that she can hardly suppress. Reflective of Sarah's moral character may instead be her anti-racist commitments because Sarah wants them to drive her responses. Insofar as these assumptions are correct, it seems that the standard view fulfils the second desideratum of a model of attitudes *(D2):* it is sensitive to the difference between aspects of an individual's psychology that are not part of her character (her automatic mental states) and those aspect that can rightly be said to be constitutive of her character (her controlled mental states). However, in the next chapter, I will argue that implicit attitudes can form part of a person's moral character because agents can take *indirect* intentional control (and *indirect* rational control) of their implicit attitudes.

For now though, it is important to note that there are two notions of control that can potentially be used to distinguish implicit from explicit attitudes. Firstly, we may ask whether we have or lack rational control over *the acquisition and change of attitudes* (see section 1.2.3). This comes down to the question as to whether the respective attitudes are reason-responsive. Implicit attitudes are assumed to be insensitive to what the agent's considers to be good reasons while explicit attitudes are said to be reason-responsive (and thus under the subject's rational control). Secondly, we can ask whether *the activation of attitudes and their influence on behaviour* is automatic or controlled (discussed in the present section). This comes down to the question as to whether the occurrence of the attitude and its influence on behaviour is intentional and requires the person's attentional focus (we may call this "intentional control"). Implicit attitudes are often said to become activated and influence behaviour automatically, while the retrieval and operation of explicit attitudes requires both the subject's

intention and attentional focus (Devine, 1989; Fazio et al., 1995). Claims about the automaticity-control distinction (about intentional control) can be made independently of claims about rational control (and vice versa). If one chooses to base the distinction between implicit and explicit attitudes on the automaticity-control distinction, one does not need to commit to any claims about reason-responsiveness (e.g., De Houwer, 2014; Fazio, 2001; Wilson et al., 2000). Conversely, if one chooses to base the distinction between implicit and explicit attitudes on a difference in reason-responsiveness, one does not need to commit to any claims about the automaticity-control distinction.[34] That is, proponents of the standard view can base the distinction between implicit and explicit attitudes on either or both types of control.

1.2.5 Implicit and explicit attitudes as unconscious and conscious mental states

Besides differences in mental structures, rational control, and intentional control, differences in introspective awareness have traditionally been taken to be characteristic of the difference between implicit and explicit attitudes (Rydell et al., 2006; Rydell & McConnell, 2006; Greenwald & Banaji, 1995). My argument in this section proceeds as follows. I will first present the traditional view that differences in introspective awareness are characteristic of the implicit-explicit difference. I will then show that it is increasingly acknowledged, even among proponents of the standard view, that awareness does in fact not provide a criterion by reference to which a distinction between implicit and explicit attitudes could be drawn.

Greenwald and Banaji's (1995) have been influential in forming the view that implicit attitudes are unconscious. They characterise implicit attitudes as follows:

> *Implicit attitudes* are introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects. (Greenwald & Banaji, 1995: 8)

Introspection is commonly defined in the philosophical and psychological literature as some kind of unmediated access to (or perception of) one's own mental states (Borgoni, 2015). Greenwald and Banaji's (1995) claim that implicit attitudes are "introspectively unidentified (or inaccurately identified) traces of past experience" (p. 8) can thus be understood to imply that people lack direct access to the contents of those

[34] Yet, to the best of my knowledge those authors who distinguish between reason-responsive explicit attitudes and reason-insensitive implicit attitudes do as a matter of fact also acknowledge that implicit attitudes operate in an automatic mode, while explicit attitudes operate in a controlled mode (Gendler, 2008a, 2008b; Levy, 2014a, 2015).

memory states that form their implicit attitudes.[35] According to this claim, implicit attitude research is concerned with the influence of memories that are inaccessible to the agent on evaluative responses (favourable or unfavourable feeling, thought, or action). For example, the finding that people are more likely to hire white job candidates with ambiguous qualifications than black job candidates with the same qualifications has been attributed to fact that people "unconsciously harbor negative feelings and beliefs about blacks" (Dovidio & Gaertner, 2000: 315). The distinction between unconscious implicit attitudes (often also called "implicit biases") and conscious explicit attitudes has also been taken up by a range of philosophers (e.g., Kelly & Roedder, 2008; Levy, 2011, 2013; Saul, 2013; Washington & Kelly, 2016). Some of these philosophers have suggested that people are not blameworthy for problematic implicit attitudes and behaviour resulting from these if they are unaware of having these attitudes (Saul, 2013; Levy, 2011, 2013). According to Levy (2013), for example, the agent's moral responsibility for an action depends on whether the action is attributable to the agent. He argues that "only actions settled upon by conscious deliberation are deeply attributable to agents, because only such actions express the agent's evaluative stance" (p. 211). According to this view, actions that are driven by unconscious implicit attitudes do not express the agent's evaluative stance.[36] Such a way of distinguishing between unconscious implicit and conscious explicit attitudes may be in line with desideratum *D2* of a model of attitudes: it draws a line between aspects of an individual's psychology that can be said to be constitutive of that person's character (conscious explicit attitudes) and those aspects that are not part of her character (unconscious implicit attitudes).

However, in more recent years scholars in both philosophy and psychology have become sceptical of the claim that so-called implicit attitudes are in fact unconscious (e.g., Gawronski, Hofmann, & Wilbur 2006; Holroyd, 2015, 2016; Levy, 2014b; Stammers, 2016). This is because empirical evidence has accumulated to show that people can, at least sometimes, become aware of the content of those mental states that are commonly referred to as "implicit attitudes" (Gawronski, Hofmann, & Wilbur, 2006; Monteith, Voils, & Ashburn-Nardo, 2001; Hahn et al., 2014; Scaife et al., 2016). Especially striking in this regard are the results by Hahn and colleagues (2014). They conducted several studies in which participants were asked to predict their result on a

[35] I use the term "content" here and in what follows in a colloquial sense. In particular, I do not want to imply that the mental state itself has propositional content. Sarah may be said to be aware of the content of her conceptual association between BLACK PERSON and DANGER when she comes to the realisation that she associates black people with danger.

[36] Yet, as I will show below, Levy has adopted more recently the view that people can become aware of their implicit attitudes in the same way as they can become aware of their explicit attitudes (Levy, 2014b). According to this, reason-insensitivity and lack of inferential promiscuity (but not awareness) distinguish implicit from explicit attitudes.

psychological test of implicit attitudes (i.e., the implicit association test described below in section 1.3.1) before actually taking part in the test. For example, in one study (study 2), participants were required to indicate on a scale the relative negativity or positivity of their "true attitude" towards black people (and white people) that they expected the test to reveal. Afterwards, they took part in the implicit association test. Strikingly, participants' estimates of their "true attitudes" were largely in line with the attitudes that the measure of implicit attitudes revealed. This result was found across a variety of testing conditions. For example, it was replicated when participants had no previous experience with the implicit association test and received only minimal information about the test before making their prediction (study 4). Moreover, participant's estimate of their own implicit attitude was shown to be a better predictor of their result on the implicit association test than their estimate of the implicit attitude of an average person (study 3). The researchers take this to show that participants have "unique insight into their own implicit responses" that extends beyond their knowledge about other people's attitudes (p. 1380).

One may of course wonder what this "unique insight" amounts to and whether it differs in any relevant way from how we come to know about our explicit attitudes. Hahn and colleagues mention two possible routes by which their participants might have acquired knowledge about the content of their implicit attitudes: firstly, participants might have considered how they have responded to the target group on past encounters, which might have helped them to infer the contents of their attitude towards the target group. We may call this indirect or inferential awareness. Secondly, participants may have experienced a certain negative or positive gut feeling when thinking about the target group, which they then reported as their implicit attitude. Assuming that this gut feeling is constitutive (part) of the attitude, this may qualify as direct (or introspective) awareness.

One could perhaps argue that people have only inferential awareness of implicit attitudes (i.e., people need to rely on evidence to infer the content of their implicit attitudes), while they have direct introspective access to their explicit attitudes. However, this claim is problematic for at least three reasons. Firstly, the current empirical evidence concerning people's awareness of implicit attitudes is compatible with the possibility that people have direct access to their implicit attitudes (Gawronski, Hofmann, & Wilbur, 2006).[37] Secondly, it is notoriously difficult to pin down what introspective awareness is (Holroyd, 2015). While introspection is commonly defined as

[37] In fact, Hahn and colleagues (2014) argue that participants in their studies had most likely direct, and not only inferential, access to the content of their implicit attitudes. However, the empirical support that they provide for this claim is contestable. I will therefore content myself with the claim that it remains an open question as to how exactly people become aware of the content of so-called implicit attitudes. As I will argue below, the same is true for our awareness of so-called explicit attitudes.

unmediated access to (or perception of) one's own mental states, it has recently been argued that relying on evidence to infer the content of one's mental states is in fact in line with the ordinary notion of introspective access (Borgoni, 2015). Given this latter conception of introspection, the participants in Hahn and colleagues' (2014) studies had introspective access to their so called implicit attitudes even if they inferred the content of these attitudes from memories of their past behaviour towards the target group. Thirdly, and perhaps most importantly, even with respect to so called explicit attitudes it is anything but clear how we come to be aware of their contents. In fact, there is an ongoing dispute over whether people have direct (introspective) access to the content of their own propositional mental states. Carruthers (2009a), for example, argues that we use the same mindreading capacity that we use to infer other people's mental states to learn about our own beliefs and desires. According to this view, our access to our own propositional mental states is never direct (or introspective) but always mediated by certain perceptual states. This is not the place to discuss Carruthers' account of mindreading. All that I want to show is that even for mental states that are commonly assumed to be explicit (i.e., our agential propositional attitudes) it is an open question whether we have direct introspective awareness of them. It is thus anything but clear whether awareness is a factor on which the distinction between implicit and explicit attitudes can be based. On the contrary, leading proponents of the standard view (both in philosophy and psychology) have acknowledged that there is no relevant difference in the awareness that people have of so-called implicit and explicit attitudes (Levy, 2014b; Wilson et al., 2000). Levy (2014b), for example, relies on Carruthers' (2009a) account of mindreading and argues that people can become aware of the contents of their implicit attitudes in the same way as they can become aware of the contents of their explicit attitudes.[38]

To conclude, people can become aware of the contents of their alleged implicit attitudes and it is not clear whether the manner in which they become aware of their implicit attitudes differs in any way from how they become aware of their alleged explicit attitudes. This shows that if one wants to defend the distinction between implicit and explicit attitudes, basing this distinction on a difference in awareness does not seem to be promising.

[38] It may of course be the case that it is more difficult to become aware of the content of some mental states (e.g., conceptual associations) than of the content of other mental states (e.g., beliefs). However, to support the distinction between implicit and explicit attitudes on the basis of a difference in awareness one would have to argue that the content of so-called implicit attitudes is per se more difficult to access than the content of so-called explicit attitudes. This is a claim that is by far more difficult to establish. Why should we for example assume that the contents of propositional mental states are per se more easily to access than the contents of associative mental states?

1.2.6 Summary

In the foregoing part of this chapter, I showed that proponents of the standard view have defended the distinction between implicit and explicit attitudes with reference to a range of different features: mental structure, rational control, intentional control, and awareness. In the last section, I showed that it is increasingly acknowledged, even among proponents of the standard view, that awareness does not provide a criterion by reference to which a distinction between implicit and explicit attitudes can be drawn. This leaves us with the criteria of mental structure, rational control, and intentional control as potential difference makers. Implicit attitudes are frequently claimed to be associatively structured, while explicit attitudes are claimed to be propositionally structured. It is also usually assumed that this difference in mental structure goes together with a difference in reason-responsiveness (i.e., with a difference in rational control). According to this, implicit attitudes do not respond to what the subject regards to be good reasons (i.e., are not subject to rational control), while explicit attitudes are reason-responsive (i.e., are subject to rational control). Independently of this distinction, other scholars have emphasised that implicit attitudes operate in an automatic mode (i.e., without the subject's intention or attentional focus), while explicit attitudes operate in a controlled mode (i.e., guided by the subject's intention and reliant on the subject's attentional resources). In principle, one can support the claim that there are distinct implicit and explicit attitudes merely by reference to the distinction between automatic and controlled processing without committing oneself to any particular assumption about mental structure or reason-responsiveness (e.g., De Houwer, 2014; Fazio, 2001; Wilson et al., 2000). Similarly, one could defend the distinction between implicit and explicit attitudes by reference to mental structure (and the reason-responsiveness implied by this) without committing oneself to the control-automaticity distinction. I do thus not claim that the standard view implies a conjunction of claims about mental structure, rational control, *and* intentional control. Rather I take the standard view to be the claim that there are distinct implicit and explicit attitudes, and this claim can be supported by any of the above mentioned features.

1.3 Psychological measures of attitudes

In the foregoing, I showed that on the standard view certain differences in mental structure, rational control, and/or intentional control (and on some accounts also awareness) characterise the distinction between implicit and explicit attitudes. This view receives some of its support from common sense. For example, it seems plausible that Sarah's inclination to keep excessive spatial distance from black

interlocutors is driven by an association that Sarah cannot eliminate by mere reasoning (e.g., her association between BLACK PERSON and DANGER) and that becomes activated without Sarah's intention. At the same time, Sarah possesses certain anti-racist beliefs that are reason-responsive and that play a role in Sarah's intentional behaviour.

Yet, it must be stressed that proponents of the standard view do not only rely on common sense or intuition when defending their account. Crucially, they also claim that psychological measurements of people's evaluative dispositions support their view that there is a distinction to be made between implicit and explicit attitudes. This is a claim that deserves philosophical scrutiny. Our model of attitudes should certainly be informed by the psychological data, but we also have to keep in mind that our interpretation of the data relies on and deploys certain views about cognitions and their structure in the first place. This implies that we should not uncritically rely on common interpretations of measurement outcomes that suggest particular answers to the questions about attitude individuation, the mental states that underpin attitudes, and the ontology of attitudes (see questions *Q1*, *Q2*, and *Q3* in the introduction to this thesis).

In what follows, I will first elaborate on the difference between direct measures of attitudes, which supposedly measure explicit attitudes, and indirect measures, which supposedly measure implicit attitudes, and will provide some examples of each of these classes of measurement techniques (section 1.3.1). This is important because throughout the rest of this thesis I will take for granted that the reader is familiar with these different measures. Moreover, this review of measurement techniques will provide the reader with a clearer sense of the research that has inspired the distinction between implicit and explicit attitudes. In section 1.3.2, I will then argue that dissociations between people's results on indirect and direct measures of attitudes do not prove that there is a distinction to be made between implicit and explicit attitudes, unless we presuppose a particular account of attitude individuation. We should therefore first settle the issue of attitude individuation before we interpret the measurement data in terms of attitudes. I will also point to the often neglected fact that the statistical dissociation between people's results on different indirect measures of attitudes is at least as pronounced as the statistical dissociation between people's results on indirect and direct measures of attitudes. This may indicate that different measures (including different indirect measures) tap into different mental states or combinations thereof – a claim that I will further elaborate on in the following chapters. Lastly, in section 1.3.3, I will point to evidence that indicates that the distinction between implicit attitudes (as measured on indirect measures of attitudes) and explicit

attitudes (as measured on direct measures) is not actually crucial for the prediction of people's spontaneous vs. deliberate evaluative responses.

1.3.1 Direct and indirect measures of attitudes

Many of the claims about attitudes discussed in the previous sections have been informed by attitude measurement research in psychology. Broadly speaking, there are two classes of attitude measurement techniques: while on some measures people are directly prompted to express their attitudes (e.g., on semantic differentials or feeling thermometers as described below), other tests involve people in tasks that are supposed to allow for conclusions about their attitudes without directly asking them for their attitudes (e.g., the affective priming task and the implicit association test described below). I will follow De Houwer (2006) in referring to these as direct and indirect measures, respectively. It should be noted that in much of the literature these different techniques are referred to as explicit and implicit measures instead. However, this nomenclature is problematic because it may tacitly suggest that these measurement procedures assess explicit and implicit attitudes, respectively. When characterising measurement techniques, we should not confound the features of the measurement procedure itself (whether people are directly asked for their attitudes or not) with those properties of the constructs that they supposedly measure (De Houwer, 2006). Although it is commonly assumed that by directly asking participants for their attitudes they will express evaluations that exhibit those features of explicit attitudes that were reviewed in the first part of this chapter (propositional structure, reason-responsive, intentionally controlled), we should not make this part of the definition of the measurement technique. Similarly, we should refrain from defining indirect measurement techniques in terms of those evaluations that they are supposed to measure (associative, reason-insensitive, automatic evaluations). After all, it might turn out that they, at least sometimes, tap into constructs that differ from those that they are commonly supposed to measure. It may for example be the case that the Implicit Association Test (described below) does not necessarily measure associations (a possibility that I will explore in the next chapter). I will therefore use the terms "direct" and "indirect" for the measurement procedures and the terms "explicit" and "implicit", as characterised in the first part of this chapter, for those entities that are *assumed* to be measured by these procedures.

Giving some examples of direct and indirect assessment techniques will provide an insight into their differences. Questionnaires for the direct assessment of attitudes often include, amongst others, semantic differentials or feeling thermometers. On semantic differentials participants are asked to indicate how strongly they think that

certain attributes, such as “pleasant”, “aggressive”, and “friendly”, apply to a social group (e.g., Eagly & Mladinic, 1989; Payne, Burkley, & Stokes, 2008). On a feeling thermometer participants are simply required to indicate on a scale (e.g., ranging from 0 to 100) how coolly or warmly they feel towards the social group in question (e.g., Payne, Burkley, & Stokes, 2008; Hahn et al., 2014). When asked in these direct ways to report on their attitude(s), people can deliberate on which response(s) they should give and will thus engage in and report on propositional thought processes. As a result, their responses are assumed to be intentional (i.e., non-automatic) and subject to rational control. In short, direct measures are assumed to tap into explicit attitudes as they have been characterised in the first part of this chapter. However, it should also be clear that people’s responses on these measures do not necessarily reflect their sincere evaluation of the group. They may adjust their responses in order to present themselves to the researcher in a way that they deem desirable (that is, for the most part, less prejudiced than they really are; Krumpal, 2013). Moreover, people may simply be mistaken about their attitude towards the group in question (see section 1.3.2 for further elaboration on these possibilities).

These possible confounds motivated researchers to develop techniques to assess attitudes indirectly. The underlying idea is that people’s automatic, non-deliberate, evaluations of a social group are expressed in their performance on tasks that require them to respond as quickly as possible to certain representations of that social group. Many indirect measures of attitudes belong to the category of priming measures. Examples include the affective priming task (Fazio et al., 1986), the semantic priming task (Wittenbrink, Judd, & Park, 1997), and the affect misattribution procedure (Payne et al., 2005). These measures rely on the principle that the presentation of a stimulus (the prime) systematically affects participants’ subsequent reaction to another stimulus, which allows researchers to draw conclusions about how the prime is evaluated by the participants. The affective priming task (also often referred to as “evaluative priming task”) is one of the most popular indirect measures of attitudes (Fazio et al., 1986, 1995). In a classic experiment, Fazio and colleagues (1995) presented participants with picture primes of black and white individuals on a computer screen. Each prime was followed by an adjective, which had to be categorized as positive or negative as quickly as possible by a key press. For white participants, pictures of white individuals speeded up responses to positive words (e.g., attractive, likable, wonderful) as compared to negative words (e.g., annoying, disgusting, offensive) and for black participants the opposite pattern of facilitation was detected. Interestingly, these facilitation effects occur even when the prime is presented so swiftly that participants are not able to consciously perceive them (Wittenbrink, Judd, & Park, 1997, 2001). The results of affective priming studies are usually explained by an automatic activation of an

evaluation (i.e., the implicit attitude) when (subconsciously) perceiving the prime. It is assumed that the activated evaluation creates a processing advantage for evaluatively congruent target words, presumably due to activation spreading across associatively linked mental representations (Fazio, 2001; Fazio & Olson, 2003; see section 1.2.1).[39] For example, if the perception of a black person elicits a negative evaluation one will be quicker at identifying a negative word as negative and slower at identifying a positive word as positive. This process is assumed to be automatic as the subject does not pay attention to the valence of the prime and does not intend to be influenced by the valence of the prime (see section 1.2.4) Moreover, one can speculate that the mental states that are measured on the affective priming task are beyond rational control due to their associative structure (see section 1.2.3).[40]

Probably the most frequently used indirect measure in the context of attitude research is however the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). This technique is assumed to measure differential associations of two categories of interest (e.g., black persons and white persons) with two attribute dimensions (e.g., positivity and negativity). In a typical race IAT, participants have to categorise person stimuli as either belonging to the category of white or black people and words as either belonging to the class of positive or negative attributes. For example, participants may be instructed to respond by pressing the “a” key on a keyboard whenever a face of a white person (or a name that is typically linked to white people) appears on the computer screen in front of them and to press the “l” key whenever a face of a black person (or a name that is typically linked to black people) is displayed. Moreover, they are asked to use the same keys to categorise attribute words as positive or negative. For example, they may be instructed to press the “a” key whenever a positive attribute word is displayed and the “l” key whenever a negative attribute word is shown. Face stimuli and attribute words appear on the screen in alternating order. Half-way through the experiment the response mapping is swapped. That is, if participant’s were previously required to respond with “a” to white faces and positive attribute words and to respond with “l” to black faces and negative attribute words, they are now asked to respond with “a” to black faces and positive attribute words, and to respond with “l” to white faces and negative attribute words. Participants’

[39] It must be noted that although the affective priming task is generally interpreted as a measure of implicit attitudes, Fazio does not in fact distinguish between implicit and explicit attitudes (e.g., Fazio, 1990, 2007; Fazio et al., 1995; Fazio & Olson, 2003; see also footnote 15 above). According to Fazio (2007), attitudes are “associations between a given object and a given summary evaluation of the object” (p. 608). Responses on indirect measures, such as the affective priming task, are reflective of these attitudes. Responses on direct measures, by contrast, are just fallible reports of an attitude that are subject to many influences (e.g., the person’s self-presentational concerns) beyond the influence of the attitude itself.

[40] However, it shall be noted that psychologists emphasise the feature of automaticity and rarely speak about lack of rational control when describing what is measured on the affective priming task (or on other indirect measures).

reaction times (i.e., the time it takes them to press the key after onset of the respective stimulus) are recorded during the entire experiment. In a range of studies that employed this paradigm, it has been shown that white participants tend to respond faster when the response for black people and negative attributes (and the response for white people and positive attributes) is paired as compared to when the response for black people and positive attributes (and the response for white people and negative attributes) is paired (e.g., Dasgupta & Greenwald, 2001; Greenwald et al., 1998; Nosek et al., 2007).[41] This is taken to show that white people tend to associate white people with positive attributes and black people with negative attributes. The underlying assumption is that when the same reaction (e.g., pressing the "l" key) is required in response to two classes of stimuli (e.g., black person stimuli and negative attribute stimuli), the reaction will be quicker if there are pre-established associations between these stimuli. As the IAT is presumably tapping into associative mental states, we can speculate in accordance with what has been said in section 1.2.3 that the IAT taps into mental states that are beyond the person's rational control. Moreover, it is generally assumed that the IAT effect is driven by automatic processing (see section 1.2.4). Note that the participants focus their attention on the categorisation task and intend to respond as accurately and fast as possible. Although they do not intend to reveal their evaluations of the target groups, these evaluations clearly influence their performance on the IAT.

1.3.2 Interpreting dissociations between scores on different attitude measures

The finding that the outcomes of indirect and direct attitude measurements frequently diverge has often been interpreted as evidence for the claim that these measures tap into different kinds of attitudes (e.g., Dovidio et al., 1997; Wilson et al., 2000). The social psychological literature is replete with reports of studies in which participants' openly expressed evaluation of a social group diverged substantially from the evaluation that was indirectly assessed (e.g., Dovidio et al., 1997; Fazio et al., 1995; Greenwald et al., 1998; Rudman & Killianski, 2000). Greenwald and colleagues (1998: experiment 3), for example, report that while participants showed on average a significant bias against black people (and in favour of white people) on race IATs, a semantic differential measure indicated that the average participant had no racial preference whatsoever.[42] Note that this is exactly the pattern that we would expect Sarah, from the example with which we started, to show if asked to take part in a race

[41] There is mixed evidence as to how black people perform on this kind of race IAT (see Morin, 2015; Nosek et al., 2007).

[42] A feeling thermometer measure indicated some bias against black people, but this bias was only half the magnitude of the average bias against black people shown on the IAT.

IAT and to complete the semantic differential measure. Assuming that Sarah harbours associations that link black people to attributes such as danger or violence, it is likely that she will show a bias against black people on the IAT. However, as she firmly believes in egalitarianism, she will most likely ascribe as many positive and negative attributes to black people as she ascribes to white people on the semantic differential measure (on which she can control her responses). Greenwald and colleagues (1998) report an average statistical correlation of those direct measures used in their study (feeling thermometer, semantic differential, modern racism scale, diversity index, and discrimination index) with their race IAT measures of 0.14, which is quite low given that a value of 0 would indicate an absence of correlation, while a value of 1 would indicate a perfect correlation.[43]

It has to be emphasised that there are different possible reasons for why a person's reported attitude towards a social group may be misaligned with the evaluation that an indirect measure reveals. Firstly, the person may express her endorsed evaluation of the social group (say, a positive evaluation), which deviates from evaluations that indirect measures reveal (say, a negative evaluation). The person may even be aware of the fact that she harbours evaluative tendencies that conflict with her endorsed commitments. Yet, when asked for her attitude, she expresses her endorsed evaluation because she sincerely believes that this is her *real* attitude towards the group. A second possibility is that a person tries to base her attitude assessment on the content of relevant mental states that come to her mind (including mental states that she would not endorse on reflection) but that she cannot (fully) access the content of those mental states that show up on indirect measures of attitudes. A third possibility is that a person is insincere. She may be aware of the content of those evaluative mental states that drive her performance on indirect measures (and endorse this as her attitude) but report a different attitude in order to present herself in a way that she believes other people (e.g., the researcher) will approve of. The phenomenon that people underreport evaluations that others are likely to disapprove of is an example of what social scientists and psychologist call "social desirability bias" (Krumpal, 2013). In this third case, the dissociation between what the person reports when directly asked for her attitude and the evaluation that the person shows on an indirect measures of attitudes can certainly not be taken as evidence for a dissociation between attitudes. This is because what she reports on the direct measure does not reflect a sincere assessment of her attitude at all.

[43] See Taylor (1990) for a comprehensible guide to the interpretation of the correlation statistic. According to him, correlations of 0.35 or smaller are typically said to be "low or weak", correlations ranging from 0.36 to 0.67 are typically considered to be "modest or moderate", and correlations ranging from 0.68 to 1.0 are generally interpreted as "high or strong" (p. 37).

However, it must be stressed that even if (most) people give sincere answers on direct measures of attitudes (such as in the first two cases described above), a divergence from indirectly measured evaluations is not a proof for the existence of distinct sets of attitudes (see Fazio, 2007, and Fazio & Olson, 2003, for related arguments). The lack of correlation presumably indicates that indirect and direct measures tap into different (sets of) mental states, but the mere fact that certain states are dissociated does not tell us that they constitute distinct attitudes, unless we already know how attitudes are individuated (see question *Q1* that was introduced in the introduction to this thesis). To be sure, as proponents of the standard view identify attitudes with specific mental states, the fact that people's evaluations on indirect and direct measures of attitudes are often dissociated may indicate the existence of two different kinds of attitudes. But it is not clear why we should identify attitudes with specific mental states in the first place. I have already mentioned the possibility that attitudes may have a different ontological status altogether (see question *Q3* that was introduced in the introduction to this thesis). It may be that attitudes are complex traits of people that are based on a variety of different mental states that may vary in their evaluative implications (Machery, 2016). On this view, indirect and direct measures of attitudes would tap into different parts of the psychological basis of an attitude but not into different attitudes. In fact, I will defend a version of the trait view of attitudes in chapter 4 of this thesis. For now though, it suffices to emphasise that in order to interpret measurement results in terms of attitudes, we need already to have an account of the nature of attitudes. That is, some philosophical groundwork needs to be done in the first place.

So far I have stressed that if there is indeed a substantial statistical dissociation between people's scores on indirect and direct measures of attitudes, this does not constitute a proof for the existence of distinct implicit and explicit attitudes (and thus for the standard view of attitudes). However, it shall also be mentioned that it is not even clear in how far results on indirect and direct measures of attitudes are indeed dissociated. In fact, there is evidence that the size of statistical correlation between indirectly and directly assessed evaluations varies widely across different attitude objects and across different studies (Hofmann et al., 2005; Nosek, 2005, 2007). Hofmann and colleagues (2005), who conducted a meta-analyses of 125 studies in which they found a mean correlation of 0.24 between IAT results and direct measures, conclude that directly and indirectly assessed evaluations are evidently not "completely dissociated and that correlations between the two are [not] purely random" (p. 1382). One may of course reply that at least for socially sensitive issues (e.g., gender, race, and sexual orientation) there is a clear dissociation between indirectly and directly assessed evaluations (Fazio & Olson, 2003). However, even with regard to this domain

the evidence is not unequivocal. There are a number of studies that found substantial correlations between the outcomes of indirect and direct measures of prejudice (e.g., Wittenbrink et al., 1997; Kawakami, Dion, & Dovidio, 1998; McConnell & Liebold, 2001). Wittenbrink and colleagues (1997), for example, found a substantial correlation of 0.41 between people's implicit prejudice scores on a racial priming measure and people's scores on a questionnaire measure of explicit racial prejudice (the modern racism scale).[44]

Curiously enough, there is abundant evidence that the correlations between the outcomes of different indirect measures of attitudes are of a similar magnitude as the correlations between the outcomes of indirect and direct measures of attitudes (Bar-Anan & Nosek, 2014; Fazio & Olson, 2003; Sherman et al., 2003).[45] In a mass data analysis, Bar-Anan & Nosek (2014) found a mean correlation of 0.36 between participants' results on IATs and their results on six different indirect measures of attitudes. The correlation of participant's results on IATs with their results on various direct measures of attitudes was only somewhat lower (0.27). It must be noted that some of the indirect measures that Bar-Anan and Nosek (2014) included were modified versions of the IAT (e.g., they included a shorter version of the IAT). The correlations between people's IAT results and structurally more dissimilar indirect measures were lower. For example, the mean correlation of participants' scores on the race IAT and participants' scores on a race affective priming task was just 0.29.[46] This indicates that the relatively weak correlations that are (sometimes) found between the results on direct and indirect measures of attitudes do not have any special status. After all, similar dissociations are found between the results on different indirect measures of attitudes. It would be dubious to defend the standard view of attitudes by reference to the presumed dissociation between direct and indirect measures while neglecting the dissociations between different indirect measures of attitudes.

The claim that results on different measures of attitudes are often dissociated is consistent with the claim that people harbour a multitude of evaluative mental states in regard to a particular social group (see section 1.2.2). Different measures (including

[44] Moreover, Hofmann and colleagues (2005) could not find an influence of the social sensitivity of a given topic on the correlations between IAT results and the results of direct measures in their meta-analysis.

[45] By contrast, there is evidence that correlations between different direct measures of attitudes (e.g. between semantic differential and feeling thermometer measures of attitudes) are more considerable. Greenwald and colleagues (1998), for example, report an average correlation of 0.5 between five different direct measures of attitudes (p. 1475).

[46] Somewhat more anecdotally, Fazio and Olson (2003) mention that IAT measures and priming measures have repeatedly failed to correspond in their own lab, with the correlation coefficients being close to zero. Relatedly, Wittenbrink and colleagues (2001) found that two different priming tasks, involving evaluative and conceptual judgments respectively, revealed largely dissociated bias scores. They conclude that "these results suggest that automatic responses are not as invariant as it is sometimes posited" (p. 244).

different indirect measures) may often tap into different mental states (or combinations thereof), which explains why results on different measures often do not correlate well with each other (Machery, 2016: 116-117). This raises a range of questions about the notion of an attitude, which I seek to answer in the upcoming chapters. First of all, we may ask what kind of mental states different indirect measures of attitudes tap into. This relates to the question as to what kind of mental states underpin attitudes (*Q2*). Secondly, we need to take up the attitude individuation question (*Q1*). Are we to claim that different measures tap into different (implicit or explicit) attitudes or is only one of the various measures tapping into the person's actual (implicit or explicit) attitude? Or, alternatively, is a person's attitude a complex set of representations that is tapped into by both direct and indirect measures? This leads us seamlessly to the question of the ontological status of attitudes (*Q3*). Are attitudes mental states or are they traits that are based on a variety of mental states without being identical to them? We need to look beyond the attitude measurement data to answer these questions. Here, the desiderata for a model of attitudes that I have mentioned in the introduction to this thesis come into play. When the attitude measurement data allows for different attitude conceptualisations, we should adopt the model that is most conducive to the explanation and prediction of people's responses towards other people (*D1*), to the moral assessment of the attitude holder's character (*D2*), and to the communication between philosophers, psychologists, and the wider public on attitudes (*D3*). In the next section, I will be concerned with desideratum *D1*, and assess in how far indirect and direct attitude measurement results are predictive of people's evaluative responses.

1.3.3 Predictive validity of indirect and direct measures of attitudes

Assuming that attitudes, properly understood, are predictive of people's responses towards other people and assuming that indirect and direct measures of attitudes indeed access attitudes, we should expect that results on these measures are reasonably good predictors of people's responses towards other people. That is, we should expect in line with desideratum *D1* that indirect and direct measures of attitudes tap into those aspects of people's psychology that drive their responses towards other people. In particular, we would expect that indirect measures that supposedly tap into automatically operating, reason-insensitive implicit attitudes are predictive of unintentional responses and that direct measures that allegedly tap into intentionally controlled, reason-responsive explicit attitudes are predictive of deliberate responses. In line with these expectations, early studies suggested that that indirectly assessed evaluations are good predictors of spontaneous non-deliberate responses towards others and that directly assessed evaluations are good predictors of deliberate

responses (Dovidio et al., 1997; Dovidio, Kawakami, & Gaertner, 2002; Fazio et al., 1995). Dovidio and colleagues (1997) report, for example, that indirectly assessed attitudes of white participants towards black people were predictive of participants' amount of eye contact with a black interviewer and their rate of eye-blinking in the interview situation (i.e., spontaneous behaviour; see experiment 3). Moreover, directly assessed attitudes of white people towards black people predicted their verbal evaluation of the black interviewer (i.e., deliberate behaviour). In particular, higher levels of implicit racial bias against black people (as measured on a race priming task) were associated with less visual contact with the black interviewer and higher rates of eye-blinking, while higher levels of explicit bias (as measured with questionnaires) were linked to more negative verbal evaluations of the black interviewer.

However, recent meta-analyses raise doubts about the predictive validity of both indirect and direct measures of attitudes (Oswald et al., 2013; Forscher et al., 2016). Oswald and colleagues (2013) conducted a meta-analysis of 46 studies in which they found an overall correlation of just 0.1 between people's scores on direct measures of racial attitudes and different kinds of responses towards black people (e.g., microbehaviour, expressed policy preferences, expressed person perception). Recall that a value 0 would indicate an absence of correlation while a value of 1 would indicate a perfect correlation. As people may not always report their attitudes sincerely on direct measures, one might hope to find a stronger correlation between results on indirect measures and the examined classes of responses. However, Oswald and colleagues (2013) found an overall correlation between people's scores on race IATs with the examined responses that was not considerably higher than the correlation between direct measures and these responses (0.15 vs. 0.1). Even when particular categories of responses were analysed separately, there was no considerable difference between the predictive utility of the IAT and the predictive utility of direct measures detectable. Based on the above mentioned results by Dovidio and colleagues (1997: experiment 3), one would have expected that indirect measures are better predictors of microbehaviour (a category that includes nonverbal behaviour such as eye-blinking) than direct measures of attitudes. Instead, Oswald and colleagues found in their meta-analysis that neither direct measures nor the IAT were reasonable predictors of microbehaviour (the correlation coefficients were 0.02 and 0.07, respectively). Moreover, one might have expected that direct measures would outperform the IAT when it comes to the prediction of person perceptions (a category that included verbal judgments about others). However, Oswald and colleagues found comparable low correlations between the IAT and person perceptions and between direct measures and person perceptions (0.13 and 0.11, respectively). A limitation of Oswald and colleagues' meta-analysis is that it includes only the IAT as indirect

measure. However, a recent meta-analysis by Forscher and colleagues (2016) suggests that low predictive validity is not just a shortcoming of the IAT but of indirect measures in general. They analysed 426 studies that featured a range of different indirect measures of stereotypes and attitudes and found a mean correlation of these measures with different response indices of 0.11. This is arguably in the same ballpark as the correlation of 0.15 between race IATs and different sorts of discriminatory responses towards black people found by Oswald and colleagues (2013).[47]

This suggests that studies in which indirect and direct measures have been shown to be reasonable predictors of people's evaluative responses towards other people, such as the one by Dovidio and colleagues (1997), are evidently the exception rather than the rule. However, as I will argue in later chapters, this does not necessarily mean that indirect and direct measures are of no use when it comes to the prediction of evaluative responses. I will argue that the relatively low predictive validity that has been revealed in the meta-analyses may just be due to the fact that researchers often have not used the appropriate measure for a given task (see section 3.3.2) and that results on indirect and direct measures are only predictive of highly context-specific responses rather than broad classes of responses (see for example section 4.5.2).

This being said, Oswald and colleagues (2013) findings shed some doubt on the claim that indirect and direct measures of attitudes tap into different kinds of attitudes. If these measures tapped into different kinds of attitudes, we would expect them to be predictive of different kinds of responses. However, for none of the response categories that were examined in Oswald and colleagues (2013) study, there was a considerable difference between the predictive success of the IAT and the predictive success of direct measures detectable. In particular, if proponents of the standard view were right that indirect measures (such as the IAT) tap into automatic, reason-insensitive mental states, while direct measures tap into controlled, reason-responsive mental states, we should expect that indirect measures (such as the IAT) are better predictors of unintentional responses and that direct measures are better predictors of intentional responses. This is not what Oswald and colleagues (2013) found. Results on the IAT did not predict spontaneous, unintentional responses (such as microbehaviour) any better than results on direct measures and results on direct measures of attitudes did not predict deliberate, controlled behaviour (such as verbal judgments about others) any better than results on indirect measures. This suggests that we may not actually need to postulate two different kinds of attitudes (as measured

[47] It shall be mentioned that while Oswald and colleagues (2013) restricted their analysis to discriminatory behaviour, Forscher and colleagues' (2016) meta-analysis included a broader range of responses (e.g., alcohol-related behaviours, p. 11). Moreover, while Oswald and colleagues (2013) focussed their analysis on the predictive validity of the IAT, the main purpose of Forscher and colleagues' (2016) meta-analysis was an assessment of the effectiveness of different bias intervention strategies.

by indirect and direct measures of attitudes) in order to optimally explain and predict people's evaluative responses (see desideratum *D1*).

1.4 Conclusion

In this chapter, I have elaborated on how attitudes are construed on what I take to be the predominant account of attitudes in the psychological and philosophical literature (the standard view), and on how these attitudes are measured. According to the standard view, people possess implicit and explicit attitudes. Implicit attitudes are usually identified with associative mental states, while explicit attitude are commonly identified with propositional mental states (section 1.2.1). Due to the identification of attitudes with individual mental states, the standard view implies that people can possess multiple implicit and explicit attitudes (section 1.2.2). I pointed out that this view is at odds with the folk psychological conception of attitudes and thus out of line with desideratum *D3* of a model of attitudes.

The alleged associative structure of implicit attitudes is generally understood to imply that implicit attitudes are reason-insensitive (section 1.2.3). According to this assumption, the acquisition and change of implicit attitudes is not a function of what the subject acknowledges to be good reasons for their acquisition and change but rather of mere regularities in that subject's environment (i.e., implicit attitudes are not subject to rational control). By contrast, explicit attitudes are assumed to be reason-responsive (i.e., subject to rational control) due to their propositional structure. That is, explicit attitudes can be acquired and changed in accordance with what the subject deems to be good reasons for such an acquisition or change. Other proponents of the standard view base the distinction between implicit and explicit attitudes on a difference between automaticity and control (i.e., a difference in intentional control; section 1.2.4). According to this, implicit attitudes can be activated and influence behaviour without the subject's intent and without requiring attentional resources (i.e., automatically), whereas the retrieval of explicit attitudes and their influence on behaviour is intentional and requires the subject's attention. Yet other scholars have referred to awareness as a criterion to distinguish between implicit and explicit attitudes (section 1.2.5). However, I showed that recent empirical evidence indicates that people can become aware of those mental states that are usually described as implicit attitudes (i.e., those mental states that are accessed on indirect measures of attitudes) and that how they become aware of these mental states is not necessarily any different to how they become aware of their so-called explicit attitudes. If one wants to defend the distinction between implicit and explicit attitudes, awareness thus does not seem to be the right criterion. Tying the distinction between implicit and explicit attitudes to the criteria of mental

structure, rational control, and/or intentional control, by contrast, seems more promising.

A model of attitudes that distinguishes between explicit and implicit attitudes on the basis of rational and/or intentional control may seem to fulfil desideratum *D2* for a model of attitudes (as it has been mentioned in the introduction to this thesis). It can be argued that mental states that are subject to rational control and/or intentional control are part of the agent's moral character, while mental states that are not subject to these kinds of control do not form part of the agent's moral character. Hence, the standard view seems to be sensitive to the difference between aspects of a person's psychology that are and that are not constitutive of that person's moral character. However, in the next chapter I will show that those mental states that are commonly described as implicit attitudes (i.e., those mental states that are measured on indirect measures of attitudes) are, at least to some extent, subject to *indirect* rational control (i.e., implicit attitudes are indirectly reason-responsive) and *indirect* intentional control (see section 2.3). I will argue that this suffices to establish that so-called implicit attitudes can in fact reflect on a person's moral character.

In the second part of the present chapter, I elaborated on the evidence for the standard view that is allegedly provided by attitude measurement data. I gave some examples of direct and indirect measures that supposedly access explicit and implicit attitudes as they have been characterised in the first part of the chapter (section 1.3.1). Then I discussed how we should interpret dissociations between scores on different attitude measures (section 1.3.2). I argued that even when people give sincere answers on direct measures of attitudes, divergences between people's responses on indirect and direct measures cannot proof that people possess distinct implicit and explicit attitudes, unless we already adopt a certain account of attitude individuation. Moreover, I stressed that the dissociation between the outcomes of different indirect measures of attitudes is at least as strong as the dissociation between the outcomes of indirect and direct measures of attitudes. This shows that the dissociations between indirect and direct measures of attitudes deserve no special status when we are theorising about the nature of attitudes. Lastly, I examined in how far results on indirect and direct measures of attitudes are predictive of people's evaluative responses (section 1.3.3). I pointed out that the results of recent meta-analyses indicate that both indirect and direct measures of attitudes have a relatively low predictive validity. Most remarkably, the evidence suggests that there is no difference in the relative predictive success of indirect measures and direct measures of attitudes across different domains of evaluative responses (including unintentional and intentional responses). This suggest that the distinction between implicit attitudes (as measured on indirect

measures of attitudes) and explicit attitudes (as measured on direct measures) may not actually be crucial for the prediction of people's evaluative responses.

Chapter 2: Scrutinising the standard view of attitudes

2.1 Introduction

In the last chapter, I presented the standard view of attitudes, which holds that people possess two distinct classes of attitudes, implicit and explicit attitudes, which can be measured on indirect and direct measures of attitudes, respectively. Implicit attitudes are often claimed to have an associative structure, while explicit attitudes are supposed to have a propositional structure. Also, implicit attitudes are often assumed to be outside of the agent's rational and intentional control, while explicit attitudes are assumed to be subject to these kinds of control. In this chapter, I will assume for the sake of the argument that indirect measures of attitudes assess implicit attitudes, and examine whether these mental states are indeed associative and not subject to rational and intentional control.

Accordingly, when I speak of "implicit attitudes" in this chapter, I refer loosely to those mental states (irrespective of their structure) that are measured on indirect measures of attitudes and that may drive people's spontaneous responses towards other people qua members of social groups. Moreover, when I speak about "implicit evaluative responses" in this chapter, I refer loosely to these spontaneous (cognitive, affective, and behavioural) responses that are assumed to be the function of implicit attitudes.[48] Implicit evaluative responses include those responses that people typically exhibit on indirect measures of attitudes (e.g., responding more swiftly to negatively valenced words when being primed by a picture of a black person) but also spontaneous responses that people may show in day-to-day life (e.g., avoiding eye contact with black persons). My tentative use of the term "implicit attitude" notwithstanding, I will reach the conclusion that the standard view of attitudes, which distinguishes between implicit and explicit attitudes, is not the best available model of attitudes (see section 2.4).

This chapter has two main parts. In the first part (section 2.2), I will examine whether implicit attitudes are indeed associative mental states. I will review a recent account by Mandelbaum (2016), according to which implicit attitudes are in fact propositional mental states. I will give a detailed account of the argument that Mandelbaum (2016) provides for his propositional account of implicit attitudes and

[48] As mentioned in an earlier footnote (footnote 11), the term "implicit bias" as it is used in the literature is ambiguous because it may denote implicit attitudes or implicit evaluative responses (Holroyd & Sweetman, 2016).

present some of the empirical evidence that he discusses in support of his argument (section 2.2.1). I will argue that Mandelbaum fails to establish that implicit attitudes are not associative but propositional mental states (section 2.2.2). Even if the data that Mandelbaum discusses provides evidence for the propositional structure of implicit attitudes, this does not establish that all or the majority of implicit attitudes are structured propositionally. More importantly, even for the effects that Mandelbaum reviews, there are alternative explanations available that are consistent with an associative account of implicit attitudes. I will emphasise that the currently available evidence does not allow for a definite answer to the question as to how so-called implicit attitudes are structured

In the second part of this chapter (section 2.3), I will examine whether implicit attitudes indeed fail to be subject to rational and intentional control. I will argue that even on the assumption that implicit attitudes are associative mental states people can indirectly control their implicit attitudes. Associative mental states are, at least to some extent, subject to *indirect rational control*: people can structure their external environment and their internal propositional thought in order to modify their associations in accordance with their considered values (section 2.3.1). Moreover, associative mental states are, at least to some extent, subject to *indirect intentional control*: by directly controlling what they think, people can influence which associations become activated in a given situation (section 2.3.2). Drawing on Holroyd & Kelly (2016), I will argue that the fact that associations can be indirectly controlled in these ways suggests that it is misguided to assume, as some proponents of the standard view do (e.g., Levy, 2014a, 2015; Glasgow, 2016), that implicit attitudes cannot reflect on people's moral characters (section 2.3.3).

I will conclude this chapter by examining what the foregoing arguments imply for the notion of an attitude (section 2.4). I concede that it may be possible to distinguish associative, indirectly controlled, implicit attitudes from propositional, directly controlled, explicit attitudes. Yet, I will call into question whether this is in fact the best way to conceptualise attitudes. In particular, I will motivate the view (to be developed in the following chapters) that attitudes are better conceived of as complex traits, each typically based on a variety of implicit and explicit mental states.

2.2 Mental structure

As I have shown in the previous chapter, many proponents of the standard view hold that implicit attitudes are associative mental states (Gawronski & Bodenhausen, 2006; Gendler, 2008a, 2008b; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Due to their associative structure, implicit attitudes are assumed to change in accordance with

changes in external contingencies (e.g., changes in cultural stereotypes) but not in accordance with what the subject takes to be good reasons (e.g., the subject's belief that racism is wrong). By contrast, explicit attitudes are assumed to be reason-responsive (i.e., under the subject's rational control) because they have propositional structure (see section 1.2.3 in the previous chapter). In recent years, a range of scholars have called the assumption that implicit attitudes are associative mental states into question. They argue (or raise the possibility) that implicit attitudes (or "implicit biases" as they are sometimes called) are propositional mental states or, on some accounts, even fully fledged beliefs (De Houwer, 2014; Frankish, 2016; Hughes, Barnes-Holmes, & De Houwer, 2011; Levy, 2015; Mandelbaum, 2016; Webber, 2016a). The arguments that these authors provide differ considerably, but they all agree that implicit attitudes are (at least sometimes and to some extent) responsive to reasons, which should not be expected if they were associative mental states but which is predicted on a propositional account of implicit attitudes. Nevertheless, the proponents of propositional accounts of implicit attitudes maintain the claim that people possess distinct implicit and explicit attitudes, which can be identified on indirect and direct measures of attitudes, respectively.[49] In what follows, I will focus on Mandelbaum's (2016) account of implicit attitudes because it takes centre place in recent discussions on the structure of implicit attitudes.

2.2.1 Mandelbaum's argument for the propositional structure of implicit attitudes

Mandelbaum's (2016) account of implicit attitudes can be divided into a negative and a positive claim. The negative claim, which he describes as his main claim, is that implicit attitudes are not associative. The positive claim, for which he argues more tentatively, is that they are structured beliefs.[50] In what follows, I will elaborate on both of these claims in turn (sections 2.2.1.1 and 2.2.1.2). Moreover, I will elaborate on the empirical evidence that Mandelbaum refers to in support of his argument (sections 2.2.1.3 and 2.2.1.4).

[49] They mention different reasons for assuming that there are distinct implicit and explicit attitudes despite the fact that both direct and indirect measures tap into propositional mental states. In particular, they refer to differences in automaticity (De Houwer, 2014), consciousness (Mandelbaum, 2016), or the degree of rational control (Levy, 2015) to draw the implicit-explicit distinction.

[50] Although Mandelbaum (2016) initially claims that the main purpose of his paper is to argue for the negative claim, and that he will not argue extensively for the positive claim (p. 637), a considerable extent of his article turns out to be devoted to the defense of the positive claim.

2.2.1.1 Mandelbaum's negative claim

Mandelbaum's (2016) negative claim is based on assumptions about how associative mental structures can be modified:

> [I]f you want to break apart an associative structure your options are limited; you can extinguish it by presenting one of the relata without the other or you can countercondition it, by changing the valence of the relata. Those are the only routes to modulating an associative structure. In other words, if rational argumentation (or any logical or evidential intervention) can be used to modulate an implicit attitude, then that implicit attitude does not have associative structure. (p. 635)

Mandelbaum's claim that associative mental structures can only be broken apart by extinguishing and counterconditioning procedures (but not by rational argumentation) resonates with what I said in the previous chapter about associative mental states. I argued that associative mental states are sensitive to external contingencies and change only over repeated experiences. The terms "extinction" and "counterconditioning" are used by learning psychologists to denote the two ways that are available to change associations (Lieberman, 2012: chapter 2). In extinction, a contingency that previously held between two stimuli is removed. If the stimuli now occur independently of each other, the association between the mental representations of these stimuli will weaken over time. If someone, for example, associates women with motherhood (because this is a prevalent cultural stereotype), and we want to eliminate that association, it might help to introduce that person to women that are not mothers. Counterconditioning, by contrast, implies that a previous contingency is replaced with another contingency. For example, if someone associates women with supportive professional roles, and we want to break up this association, it might help to present that person repeatedly with counterstereotypic examples of women in leadership positions (Dasgupta & Asgari, 2004). The important point to take from this is that both extinction and counterconditioning change associations incrementally over repeated experiences. If someone associates women with supportive roles, encountering one woman in a leadership position or being made aware on one occasion of the fact that there are women in leadership positions will not do much to change the association. This is why Mandelbaum assumes that a substantial modification of an implicit attitude by one-shot rational argumentation speaks against the assumption that the implicit attitude is associatively structured. Argumentation is regarded to be just not the right kind of intervention to modify associations in any substantial way. If the person in the above example ceases to associate women with supportive roles after having been presented with a single argument to the effect that there are women in leadership positions, this modification is not due to extinction or counter-conditioning, and this speaks according to Mandelbaum against the associative structure of that attitude.

Some words are in order to explain the meanings of the terms “logical intervention” and “evidential intervention” that appear in the above quote from Mandelbaum (2016). Unfortunately, Mandelbaum does not make explicit what he means by these terms. From the examples he gives one can infer that an evidential intervention involves the onetime presentation of information pertaining to an attitude object. For example, in one study that Mandelbaum refers to participants were presented with a statement informing them about whether a target person is liked or disliked by another person (Gawronski, Walther, & Blank, 2005). An evidential intervention stands in direct contrast to counterconditioning or extinction that always require repeated presentations of information.[51] A logical intervention involves highlighting a logical relation in which the attitude object stands. For example, in another study that Mandelbaum mentions, participants were told that a group to which they had previously developed an attitude is equivalent to another group (Gregg, Seibt, & Banaji, 2006). Thus, a logical intervention is also a onetime presentation of information pertaining to the attitude object, where the information is about a logical relation. I therefore suggest viewing logical interventions as a subclass of evidential interventions. Henceforth, when I speak of “evidential interventions”, I take this to include logical interventions.

Mandelbaum’s argument against the associative structure of implicit attitudes can be reconstructed in the form of a *modus ponens*:

(*P1*) If a mental structure can significantly be changed by a single argument or an evidential intervention, it is *not* an associative structure.

(*P2*) Implicit attitudes can significantly be changed by single arguments or evidential interventions.[52]

(*C*) Hence, implicit attitudes are *not* associative structures.

[51] One may also speculate that another difference between evidential interventions and counterconditiong/extinction is that these latter interventions do not involve the presentation of propositional information, whereas evidential interventions do. However, this does not seem to be right to me. Firstly, although all examples Mandelbaum provides for evidential interventions seem to involve the presentation of propositional information, I suspect that he does not take this to be a necessary feature of evidential interventions. All he needs to argue for is that the intervention has some impact on propositional mental states of the individual and this can be achieved by presenting a simple non-propositional stimulus (e.g., a picture of a person that contradicts a common stereotype). Moreover, it seems possible to use propositional information for counterconditioning/extinction. One might countercondition the association between MOTHERHOOD and WOMEN by repeatedly reading the propositional statement “many women are childless” (see also section 2.2.2.2 on how propositional thought may modify associations).

[52] The inclusion of the word “significant” is crucial here. As mentioned above, associative structures change incrementally. Each individual trial in an extinction or counterconditioning procedure contributes to the overall change of association. Each individual trial is a onetime presentation of information (i.e., information on what stimuli are paired with each other) and can thus be seen as an evidential intervention. This leads us to the conclusion that evidential interventions *can* in principle change associations. However, the influence of each individual trial should barely be noticeable. Evidence that an evidential intervention changes people’s implicit attitudes in a statistically significant manner can thus be seen as evidence that implicit attitudes are not associative (see section 2.2.1.3 and 2.2.1.4 for such evidence).

P1 follows from what I have said above about how one can modify associations. Mandelbaum refers to evidence from psychological experiments to support *P2*. I will review some of this evidence in sections 2.2.1.3 and 2.2.1.4. Before I do that, however, it is worth elaborating on Mandelbaum's positive claim. That is, the alternative view that implicit attitudes are propositional structures.

2.2.1.2 Mandelbaum's positive claim

Mandelbaum (2016) argues that implicit attitudes are unconscious beliefs. He describes these mental states as "honest-to-god propositionally structured mental representations that we bear the belief relation to" (p. 635). He remains agnostic about whether these, what he calls, "structured beliefs" are necessarily unconscious or just usually unconscious. Recall that I have pointed out in the last chapter that so-called implicit attitudes are not necessarily unconscious and that it is still hotly debated in what sense we are conscious of our agential propositional attitudes (see section 1.2.5). If there is in fact no difference between implicit and explicit attitudes in terms of consciousness, this poses a challenge for Mandelbaum because he would need to find a different criterion to distinguish between implicit and explicit attitudes.

For now though, it is important to review what other properties Mandelbaum ascribes to structured beliefs. He insists that structured beliefs differ, due to their propositional structure, in important ways from associations. Most importantly, they can respond to reasons and take part in inferences. Consider that someone holds the (unconscious) belief that all blonde women are stupid. Such a belief can, at least in principle, be modified by an evidential intervention. One may for example point out to this person that his friend Sue has blonde hair and is not stupid, which may lead him to (unconsciously) infer that not all blonde women are stupid. That is, the evidential intervention provides a reason for the subject to update his belief. To use a term from the last chapter, beliefs are under the subject's rational control (see section 1.2.3). That is, they (usually) change in accordance with what the subject accepts to be a good argument or good evidence. Accordingly, Mandelbaum argues that if an argument or an evidential intervention changes an implicit attitude, this suggests that the implicit attitude is a structured belief. Combined with his negative claim, Mandelbaum's argument can be interpreted as follows:

(*P1*) If a mental structure can significantly be changed by a single argument or an evidential intervention, it is *not* an associative structure but a structured belief.

(*P2*) Implicit attitudes can significantly be changed by single arguments or evidential interventions.

(*C*) Hence, implicit attitudes are *not* associative structures but structured beliefs.

This argument does of course *not* imply that structured beliefs *always* change in accordance with good arguments or evidential interventions. Obviously, people sometimes stick to their beliefs despite contrary evidence. This implies that instances in which implicit attitudes do not respond to arguments or evidential interventions do not necessarily disprove the structured belief account.

A problem with the structured belief account of implicit attitudes is that it supposes a notion of belief that many scholars would reject. This is because it allows for beliefs that the subject would not endorse or indeed reject on reflection. Mandelbaum (2014) defends a Spinozan view of belief fixation (see also Gilbert, 1991). On this account, people automatically believe every proposition that they entertain. According to him, the automatic acceptance of a proposition is a subpersonal process that can, but must not, be followed by a reflective endorsement or rejection of the belief on person-level. By contrast, on the more standard Cartesian view of belief fixation, propositions can be entertained mentally without assenting to them (e.g., Fodor, 1983: 102). On this account, a belief is formed upon accepting or rejecting the respective proposition in a second step (in the case of rejection the negation of the proposition is believed). This account implies that implicit attitudes that the subject has not endorsed cannot count as beliefs.

Other scholars would refute Mandelbaum's (2014) conception of believing on different grounds. For example, it has been argued that propositional mental states only qualify as beliefs if they "feature often enough and broadly enough in the kinds of normatively respectable inferential transitions that characterize beliefs" and that this is not the case for implicit attitudes (Levy, 2015: 816). Moreover, dispositionalists, interpretivists, and some functionalists would deny that beliefs are ever instantiated by particular representational structures (Schwitzgebel, 2015). This is not the place to argue for a specific account of belief. In fact, Mandelbaum (2016) anticipates that many of his readers will not accept his account of belief and offers them to view propositionally structured thought rather than structured belief as the alternative to the associative account of implicit attitudes (p. 636).[53] Accordingly, I will present Mandelbaum's argument as an argument for the propositional structure of implicit attitudes and leave it open whether these propositional structures qualify as beliefs. Framed in this less radical way various proponents of propositional accounts of implicit attitudes, who do not endorse the claim that implicit attitudes are fully-fledged beliefs, will find common ground with Mandelbaum (e.g., De Houwer, 2014; Hughes, Barnes-

[53] See however Borgoni (2015), Frankish (2016), and Webber (2016a) for accounts in support of the view that implicit attitudes are beliefs.

Holmes, & De Houwer, 2011; Levy, 2015). The modified argument, which combines Mandlebaum's negative and positive claims, reads as follows:

(*P1*) If a mental structure can significantly be changed by a single argument or an evidential intervention, it is *not* an associative structure but a propositional structure.

(*P2*) Implicit attitudes can significantly be changed by single arguments or evidential interventions.

(*C*) Hence, implicit attitudes are *not* associative structures but propositional structures.

In the next two sections, I will present two (sets of) studies that Mandelbaum refers to in order to support premise *P2*.[54]

2.2.1.3 Argument strength

One of the experiments that according to Mandelbaum (2016) support the view that implicit attitudes are propositional mental states has been conducted by Briñol and colleagues (2008, 2009).[55] I will present the experiment as Mandelbaum presents it in his paper to make clear how he uses it to support his argument. However, it shall already be noted that Mandelbaum fails to mention a crucial aspect of the experiment, which leads him to interpret the data in a different way than the authors of the study (see section 2.2.2.2). Briñol and colleagues (2008) presented participants with either strong or weak arguments for hiring African-American professors and assessed

[54] Mandelbaum reviews in total four different (sets of) studies (Briñol, Petty, & McCaslin, 2009; Gawronski, Walther, & Blank, 2005; Gregg, Seibt, & Banaji, 2006; Sechrist & Stangor, 2001). However, I believe that the two (sets of) studies that I present in what follows provide the strongest support for premise *P2* because they are arguably about the *change* of already existent attitudes (Briñol, Petty, & McCaslin, 2009; Sechrist & Stangor, 2001). The other two (sets of) studies that Mandelbaum mentions are primarily concerned with the *acquisition* of attitudes towards previously unknown persons or fictitious tribes (Gawronski, Walther, & Blank, 2005; Gregg, Seibt, & Banaji, 2006; though see studies 3 and 4 in Gregg, Seibt, & Banaji, 2006). This is surprising because studies about attitude acquisition do not actually support Mandelbaum's argument as stated above. Early in his 2016 article, Mandelbaum makes clear that his argument is about attitude change and not about attitude acquisition. While he argues that associations can only be *changed* by conditioning procedures, he emphasises that associations are not necessarily *acquired* through conditioning. According to him, they can for example be acquired trough "one-shot learning" (p. 633). This implies that if an evidential intervention, which is a one-shot intervention, leads to the *acquisition* of an attitude, this is compatible with an associative account of attitudes. It is thus surprising that later in his article Mandelbaum tries to support his argument also with reference to evidence that certain one-shot interventions lead to the *acquisition* of attitudes. Here, I will be concerned with Mandelbaum's argument as he initially presents it: as an argument about the modification of already existent attitudes.

[55] The results are reported in Briñol and colleagues (2009), who reference an unpublished working paper by Briñol and colleagues (2008) as the original source.

subsequently how this manipulation affected their race IAT scores. For instance, one of the strong arguments consisted in the claim “that because the number and quality of professors would increase with this program (without any tuition increase), the number of students per class could be reduced by 25%” (Briñol, Petty, & McCaslin, 2009: 294). In comparison, one of the weak arguments stated “that implementing the program would allow the university to take part in a national trend, and that with the new professors, current professors might have more free time to themselves” (ibid).

After the manipulation, participants in the strong argument condition exhibited on average a more positive evaluation of black people on the IAT than participants who had been presented with weak arguments. This supports supposedly premise *P2* of Mandelbaum’s argument because it shows that a single argument can affect implicit attitudes (assuming that the IAT taps into implicit attitudes). According to Mandelbaum, the observed effect is incompatible with an associative account of implicit attitudes. He notes that those features of the arguments that might modify valenced associations by way of conditioning were controlled across both conditions. For example, it was ensured that strong and weak arguments contained the same number of references to “African-American professors” and that in both conditions the hiring of African-American professors was linked to positive attributes (e.g., “better quality” and “more free time”). Accordingly, if implicit attitudes were associative and thus malleable by conditioning procedures, we should expect the strong and weak arguments to be equally effective (or non-effective) in modifying participants’ attitudes. The fact that there was actually a difference detectable between the effect of strong and the effect of weak arguments supports, according to Mandelbaum, the view that implicit attitudes have a propositional structure. He argues that propositional structures update with reasoning and inference and that the strength of an argument or the weight of evidence contained in a persuasive message can influence reasoning and inference. In short, whereas the strength of an argument does not affect associative mental states, it is exactly the sort of thing that can affect propositional mental states (see premise *P1* of Mandelbaum’s argument). Although Mandelbaum does not use the term, we may say that the participants in Briñol and colleagues (2008) experiment exerted, at least to a certain extent, rational control over their implicit attitudes. That is, their attitudes were reason-responsive.

2.2.1.4 Evidential adjustment to peer attitudes

Another study that Mandelbaum (2016) refers to provides evidence for the influence of peer opinions on implicit attitudes (Sechrist & Stangor, 2001). Participants in this study completed first a questionnaire designed to assess attitudes towards African-

Americans (The Pro-Black Scale; Katz & Hass, 1988) and subsequently were made believe that their responses on the questionnaire had been compared to other students' responses. The experimenters allocated the participants to either a low-prejudice or a high-prejudice group based on their score on the questionnaire. Half of the participants in both the high-prejudice and the low-prejudice group were informed that 81% of the students of their university agreed with their judgments expressed on the questionnaire (high-consensus condition), whereas the other half of the participants was told that 19% of their student peers agreed with their judgments (low-consensus condition). In Mandelbaum's terms, this feedback constitutes the evidential intervention. After this feedback was given, each participant was asked to wait in the hallway for the next part of the study to begin. There, an African American confederate of the experimenter was sitting on a chair at the end of a row of chairs. The dependent measure tracked how many chairs apart from the African American the participant would choose to sit. As predicted, the participants in the high-prejudice group sat further away from the African-American than the participants in the low-prejudice group. Most interestingly, however, high consensus feedback increased the extent of the attitude-behaviour relationship. That is, participants in the high-prejudice group who had received the feedback that a large majority of their peers agree with their prejudiced views sat even farther away from the African American, as compared with high-prejudiced participants who had received low-consensus feedback. Similarly, low-prejudiced participants who learned that their beliefs are largely shared sat even closer to the black person in comparison to low-prejudiced participants in the low-consensus condition.

According to Mandelbaum, these results indicate that evidential interventions (i.e., consensus feedback) changed participant's implicit attitudes (see premise *P2* of his argument). Mandelbaum claims that such a pronounced influence of a single evidential intervention on implicit attitudes is not consistent with an associative account of implicit attitudes (see premise *P1* of his argument).[56] However, this effect is according to him

[56] Mandelbaum (2016) also claims that an associative account of implicit attitudes makes predictions that are opposite to the findings observed. However, his remarks on this are rather speculative. In particular, he argues that differences in seating distance to stigmatized individuals are usually explained by differences in negative affect experienced towards these individuals. In accordance with an associative account one can say that the seating distance is determined by an association between a representation of the stigmatized group and a negative valence. Mandelbaum claims that this account cannot explain the particular findings of Sechrist and Stangor's study. Relying on dissonance theory, he argues that it should feel good for the highly-prejudiced person to get positive feedback in the high-consensus condition (Elliot & Devine, 1994). He expects this reaction to inhibit the fear response elicited by the African American, which in turn should counteract the tendency to keep distance to him. Similarly, the participant who finds out that his highly prejudiced views are not shared should experience more negative affect and accordingly sit farther away from the African American. This leads Mandelbaum to claim that an associative account makes the wrong predictions: it predicts that those highly-prejudiced participants in the high-consensus condition would sit closer to the

consistent with, and in fact predicted by, a propositional account of implicit attitudes. He argues that "subjects [were] adjusting the strength of their implicit attitudes in virtue of what they took to be germane evidence, the opinions of their peers" (p. 642). In other words, participants took their peers' opinions (unconsciously) as reason to adjust their implicit attitudes.

One may object that participants in the experiment did not actually update their implicit attitudes but that high-consensus and low-consensus feedback motivated participants temporarily to exhibit control over their behaviour (Madva, 2016: 2676-2677). That is, rather than changing attitudes the experimental manipulation may have brought to the fore people's motivation to control the effects of their attitudes on behaviour. However, it shall be noted that Sechrist and Stangor (2001) were able to replicate their findings in a second study, in which a more straightforward indirect measure of attitudes was used that rules out the influence of behavioural control. Participants in this experiment were first prompted to express stereotypic views about African-Americans. They then received high-consensus or low-consensus feedback, consisting in the information that 81% or 19% of their peers agree with their stereotypic views.[57] Subsequently, they participated in a so-called lexical decision task. This indirect measure requires participants to decide for several strings of letters whether they form a meaningful word or a meaningless non-word. Some of the words denoted traits that are commonly perceived as stereotypic of black people (e.g., uneducated, violent). Importantly, the presentation of each of the letter strings was preceded by the presentation of a priming stimulus that was displayed subliminally (i.e., so swiftly that participants could not consciously perceive it). It was crucial for the experiment that on some trials the word "black" served as prime, while on other trials control words such as "chair" were presented. The researchers found that participants in the high-consensus condition responded faster to stereotype words when they were primed with the word "black" than when they were primed with control words such as "chair". By contrast, for participants in the low-consensus condition it made no difference whether they were primed with the word "black" or a control stimulus. Unlike in the case of seating distance, it can be ruled out that participants in this second experiment intentionally modulated their responses in accordance with the previously received feedback. This is because the participants did not consciously perceive the primes on the lexical decision task and the task was presented to them as visual word recognition study. It is thus unlikely that they saw the lexical decision task as connected to the

African-American than those in the low-consensus condition, although the findings were exactly the opposite.

[57] One may object that Sechrist and Stangor's (2001) second experiment reveals something about implicit stereotypes but not about implicit attitudes. I would like to object that stereotypes are central components of attitudes. I will further support this claim in the next chapter.

feedback that they had previously received. Hence, the results from this second study can be regarded as support for the view that the evidential intervention did not just affect people's motivation to control responses but genuinely affected people's implicit attitudes. This therefore supports premise *P2* of Mandelbaum's argument, namely, that evidential interventions can change implicit attitudes (see, however, section 2.2.2.3 below for an alternative interpretation).

2.2.2 Evaluation of Mandelbaum's propositional model

In what follows, I will argue that Mandelbaum fails to establish that (all) implicit attitudes are propositional mental states. My argument is threefold. Firstly, I will emphasise that even if Mandelbaum convinced us that the observed effects of arguments and evidential interventions on implicit evaluative responses were the function of changes in propositional mental states, this does not establish that in all or the majority of circumstances propositional mental states drive implicit evaluative responses (see section 2.2.2.1). My second and third argument will establish that even in the very cases that Mandelbaum discusses there are alternative explanations for the effects of arguments and evidential interventions on implicit evaluative responses that are compatible with an associative account of implicit attitudes. I will argue that arguments and evidential interventions may trigger propositional thought processes that in turn create or modify those associations that influence people's implicit evaluative responses (section 2.2.2.2; see also Madva, 2016, for a related argument). Moreover, I will argue that arguments and evidential interventions may activate already existent alternative associations that can influence people's implicit evaluative responses (section 2.2.2.3).

2.2.2.1 Insufficiency of the evidence

Let us for the time being assume that Mandelbaum succeeds in convincing us that people's implicit evaluative responses were driven by propositional mental states in the cases reviewed by him. Note that this can of course not yet establish that implicit evaluative responses are always (or in the majority of cases) the function of propositional mental states. In particular, it remains a possibility that people's implicit evaluative responses are frequently driven by associative mental states but that when people are presented with arguments or evidence of a certain kind (unconscious) propositional processes are activated and affect these responses (either alone, or in addition to the associative processes). In other words, the evidence that Mandelbaum provides leaves open the extent to which propositional mental states play a role in

implicit evaluative responses.[58] Although Mandelbaum's argument is clearly presented as an argument against the claim that implicit attitudes are associative, some of what Mandelbaum says suggests that he may be open to the possibility that associative mental states can at least under some circumstances contribute to implicit evaluative responses. He grants that propositional mental states are not the only mental states in the mind's unconscious and that there is no reason to deny the existence of associations. However, he hesitates to add that "such associations do far less causal work than is often supposed, especially in the implicit bias literature" (p. 636). Mandelbaum's suspicion seems to be that if associative mental states play a role in implicit evaluative responses, their role is negligible.

Unfortunately, Mandelbaum does not provide any argument to support this suspicion. The evidence that he discusses cannot, at any rate, establish that the contribution of associative mental states to implicit evaluative responses is negligible. It remains a possibility that some of those mental states that drive implicit evaluative responses are associative, while others are propositional. If we identify implicit attitudes with those mental states that drive implicit evaluative responses, it may thus in fact be the case that some implicit attitudes are structured associatively, while others are structured propositionally. That is, implicit attitudes may be a heterogeneous class of mental states (see Holroyd & Sweetman, 2016, for a related argument). In this case it may in fact be misleading to refer to "implicit attitudes" as if we were speaking about a specific psychological kind. As Holroyd and colleagues (2017b) point out, we may come to the conclusion that implicit attitudes (or implicit biases) are not a psychological kind if it turned out that "there is no unified phenomenon, with any distinctive set of characteristics, that underpins the behavioural responses found on indirect measures" (p. 13). Accordingly, one may want to eliminate the notion of an implicit attitude from our explanations and instead refer to "implicit associations" and "implicit propositional mental states". However, it shall be emphasised that the currently available evidence does not establish that we have to reach this conclusion.

[58] To be fair, Mandelbaum is not the only author who discusses evidence in favour of a propositional account of implicit attitudes. See De Houwer (2014) for further evidence. Yet still, the combined evidence can hardly establish that implicit evaluative responses are always (or in the majority of cases) driven by propositional mental states. De Houwer's (2014) conclusion is relatively carefully phrased anyway. He acknowledges that he cannot provide a knock-down argument against the associative account of implicit attitudes but advises researchers to seriously consider the possibility that implicit attitudes are propositional because this may lead research into new fruitful directions (see also Hughes, Barnes-Holmes, & De Houwer, 2011).

2.2.2.2 Modification of associations through propositional thought

Above, I have assumed for the sake of the argument that Mandelbaum (2016) succeeds to show that *in the cases described* implicit evaluative responses were driven by propositional mental states (i.e., propositional implicit attitudes). In what follows, I will call this assumption into question. There is an alternative explanation for the effects of arguments and evidential interventions on implicit evaluative responses, and this alternative explanation is compatible with an associative account of implicit attitudes. In short, although arguments and evidential interventions may not have a significant *direct* influence on associations, they may have an *indirect* one that is mediated by propositional thought (see Madva, 2016, for a related argument).

My argument will draw on the widely accepted hypothesis that propositional processes can create and modify associations (Gawronski & Bodenhausen, 2006, 2011; Madva, 2016; Strack & Deutsch, 2004). Mandelbaum himself acknowledges parenthetically that "Structured Beliefs can create associations through the mere continued activation of the constituents of the beliefs" (Mandelbaum, 2016: endnote 13). One of his examples is the belief "dogs sleep on tables". Plausibly, if a person frequently tokens this belief (e.g., by thinking about dogs sleeping on tables), a mental association between the concepts DOG and the concept TABLE will be established. Once this connection has been established, an activation of DOG will spread over to TABLE (and vice versa). Mandelbaum refers to associations that have been created in this way as "piggybacking associations". Mandelbaum also stresses that *evaluative* associations can by created by propositional thought: if DOG has a positive and TABLE no definite valence, repeatedly thinking "dogs sleep on tables" will lead the concept TABLE to acquire a positive valence. This can be seen as an instance of evaluative conditioning. Plausibly, conditioning by propositional thought works very much like conditioning by virtue of co-occurrences of external stimuli. The only difference is that in the case of, what we may call, propositional conditioning co-occurrences of concepts in thought create associations rather than co-occurrences between external stimuli. Although this is not mentioned by Mandelbaum, propositional thought can arguably not only establish associations but also play a role in the modification of associations (Gawronski & Bodenhausen, 2006, 2011; Madva, 2016).[59] For example, if I stop thinking "dogs sleep on tables", this will gradually lead to the extinction of my association between DOG and TABLE if this association is not reinforced by any other means. Similarly, if TABLE is linked to positive valence, repeatedly thinking "Faeces

[59] It shall be mentioned that the relation between associations and propositional thought is in fact bidirectional on Gawronski & Bodenhausen's (2006) influential model. Propositional thought can establish, modify as well as activate associations, and reversely activated associations can feed into propositional thought processes.

are on the table" may well countercondition this association due to the negative valence of FAECES. Curiously enough, Mandelbaum does not discuss the possibility that associations that have been created or modified by propositional thought may drive implicit evaluative responses. This is surprising because such associations may well have played a causal role in the very experiments that he refers to.

To demonstrate this, it is worth having a closer look at the argument strength experiment by Briñol and colleagues (2008, 2009) that was presented in section 2.2.1.3.[60] When describing the experiment, Mandelbaum fails to mention a second factor that the researchers manipulated. Briñol and colleagues (2008) did not only vary the strength of the arguments for hiring African-American professors but also the extent to which the participants would think about the arguments. Previous research had indicated that people elaborate more strongly on a message the more they regard the message as personally relevant to them and the more they view it as their personal responsibility to think about the message (Petty & Cacioppo, 1979; Petty, Harkins, & Williams, 1980). Based on these findings, Briñol and colleagues (2008) created a high-elaboration and a low-elaboration condition in their study. In the high-elaboration condition, participants were told that the policy to hire more African-American professors would possibly be realised at their own university in the upcoming academic year (creation of personal relevance) and that only a few people would assess the arguments (high personal responsibility). By contrast, participants in the low-elaboration condition were told that the policy would be realised at a remote university in 10 years (no personal relevance) and that a large group would assess the arguments (low personal responsibility). Importantly, the effect of argument strength on participants' race IAT scores that Mandelbaum emphasises has only been found in the high-elaboration condition – namely, the condition in which the participants thought more about the arguments and propositions involved in them. In this condition, participants' implicit evaluations of black people (as measured on the race IAT) were more positive when they had received strong arguments for hiring African-American professors than when they had read weak arguments. No such effect was present in the low-elaboration condition. This suggests that the extent to which participants thought about the arguments determined whether argument strength had an effect on implicit evaluation. When participants thought extensively about the arguments because they saw personal relevance in them and felt the responsibility to think thoroughly about them, argument strength had an impact on implicit evaluations, but when they did not give much thought to the arguments, argument strength had no effect.

[60] See Madva (2016: 2678-2679) for a related discussion of Briñol and colleagues' results.

Briñol and colleagues' (2009) explanation of this effect draws on the above mentioned assumption that repeated thought can create associations:

> [W]e argue that the effect of argument quality obtained under high elaboration on automatic evaluations is due to the fact that the strong message led to many favorable thoughts associated with the integration program and Blacks, whereas the weak message led to many unfavorable thoughts associated with the integration program and Blacks. We speculate that, at least in this persuasion paradigm, the generation of each (negative) thought provides people with the opportunity to rehearse a favourable (unfavourable) evaluation of Blacks, and it is the rehearsal of the evaluation allowed by the thoughts (*not the thoughts directly*) that are responsible for the effects on the implicit measure. *Thus, the automatic change might involve just getting the link between the attitude object and good (bad) rehearsed by each favorable (unfavorable) thought.* (p. 295, my emphasis)

Briñol and colleagues (2009) argue here that it is not "the thoughts directly" but the "rehearsal of the evaluation allowed by the thoughts" that resulted in the effect that was found in the high-elaboration condition (p. 295). What they imply is conditioning by propositional thought: positive thoughts about black people strengthen an association between black people and positive valence and negative thoughts have the opposite effect. The more positive (or negative) thoughts a person entertains about black people (i.e., the more the positive or negative evaluation is rehearsed), the stronger the association between black people and positive (or negative) valence may become (Madva, 2016: 2678-2779). This can explain why the strength of the arguments only had a substantial effect in the high-elaboration condition: as participants in the high-elaboration condition thought more extensively about the strong and weak arguments than the participants in the low-elaboration condition, their associations changed more substantially as a result of that (propositional) thinking.[61] Note that the thought processes that the participants engaged in may well have been unconscious. There is no reason to assume that thought processes must be conscious to create or change associations.

Note also that the results found in the study by Sechrist and Stangor (2001; see section 2.2.1.4) can possibly be explained along the same lines. Let us consider their second experiment. When participants were presented with information that their peers largely agree with their stereotypic views about black people (high-consensus group), this may have boosted their confidence in these views which led to more negative thinking about black people. This, in turn, may have strengthened negatively valenced associations in regard to black people. By contrast, those participants who learned that their stereotypic views are not shared by their peers may have lost the confidence in

[61] In a second experiment by Briñol and colleagues (2008), participants read either strong or a weak arguments in favour of including more vegetables in their diet. After having read the arguments, they were required to note down their thoughts in regard to the proposal. Strikingly, in the high-elaboration condition (but not in the low-elaboration condition), participants' evaluation of vegetables on an IAT was mediated by the valence of the thoughts that they had listed.

their beliefs (low-consensus group), which may have led them to entertain positive thoughts (or less negative thoughts) about black people. This, in turn, may have strengthened positively valenced associations (or weakened negatively valenced associations). These effects may explain why people in the high-consensus group showed a bias on the lexical decision task (i.e., responding faster to stereotype words than to control words when primed with the word “black”), while people in the low-consensus group showed no such bias.

To sum up, the evidence that Mandelbaum (2016) provides is compatible with the assumption that implicit attitudes are associative. Argumentation and evidential interventions presumably trigger propositional thought processes and these may lead to the modification of associations that cause implicit evaluative responses. This explanation undermines premise *P1* of Mandelbaum’s argument. That is, if a mental structure can significantly be changed by a single argument or an evidential intervention, this *does not necessarily* speak against the claim that this mental structure is associative.

2.2.2.3 Activation of associations through propositional thought

Although the explanation provided in the previous section is possible, it may not yet be the best explanation for the effects reviewed by Mandelbaum (2016). In particular, one may wonder whether the time between experimental manipulation and attitude measurement in the presented studies was long enough to actually lead to a significant *modification of associations* (even if participants thought intensely about the arguments and peer opinions provided). Rather the arguments and peer opinions provided may simply have led to the *selective activation of already existing associations*. This is my second alternative explanation for the data discussed by Mandelbaum (2016).

Recall that I have argued in the last chapter (section 1.2.2) that if we identify implicit attitudes with individual associative mental states (as proponents of the standard view seem to do), we need to acknowledge that people likely harbour a multitude of implicit attitudes in regard to a social group. As an example, I have mentioned that Sarah’s negatively valenced association between BLACK PERSON and DANGER may become tokened independently of her positively valenced association between BLACK PERSON and MUSICALITY (and vice versa). Which of these associations (or we may say implicit attitudes) become tokened may depend on situational influences. Crucially, the fact that associations can become activated selectively may also explain the findings that were observed in the studies discussed by Mandelbaum (2016).

Let us again consider the study by Briñol and colleagues (2008, 2009). As mentioned above, they found that when participants thought extensively about the arguments presented to them (high-elaboration condition), argument strength influenced their implicit evaluations of black people on an IAT. Participants who had received strong arguments for hiring African-American professors showed more positive evaluations of black people on a race IAT than participants who had read weak arguments for hiring African-American professors. Note that strong arguments for hiring African-American professors may have triggered positive thoughts about black people, leading to the temporary activation of positively valenced associations, while weak arguments for hiring African-Americans may have triggered negative thoughts about black people, leading to the temporary activation of negatively valenced associations. The difference in the content and valence of the activated associations may have caused the difference in participants' evaluative responses on the IAT that was conducted shortly after the experimental manipulation.

A similar explanation can be given for the findings observed by Sechrist and Stangor (2001: study 2). When participants received the feedback that a majority of their peers agrees with their stereotypic views about African-Americans (high-consensus group), negative thoughts about African-Americans may have been reinforced, which led in turn to an increased activation of negatively valenced associations. By contrast, when participants were informed that a majority of their peers disagreed with their stereotypic views (low-consensus group), this may have led to them to think temporarily in more positive terms about African-Americans, which in turn may have led to the increased activation of positive associations. The difference in the content and valence of the temporarily activated associations may have caused the difference in responding that was observed on the subsequently conducted lexical decision task.

The here presented alternative explanation implies that premise *P2* of Mandelbaum's argument may actually be wrong. That is, the evidence does not establish that implicit attitudes can be significantly changed by single arguments or evidential interventions. This is because the evidence is also compatible with the idea that single arguments or evidential interventions selectively activated pre-existing associations (i.e., implicit attitudes) rather than "rewired" associations. On this alternative view, arguments or peer opinions that people are confronted with can be seen as situational or contextual factors that temporarily influence people's evaluative responses without actually changing people's attitudes (Madva, 2016: 2676-2677). If this explanation is correct, the effects of arguments and evidential interventions on people's implicit evaluative responses should only be short-lived. This is an assumption that can be tested but that has not been examined in the studies that Mandelbaum

reports. In these studies, indirect attitude tests were conducted shortly after the experimental manipulation. Hence, the data is consistent with the assumption that the observed effects were just due to transient attitude activation rather than lasting attitude change. At this point, it shall already be noted that the situation-specificity of evaluative responses takes centre stage in the model of attitudes that I will propose in chapter 4.

2.2.2.4 Preliminary conclusion

Mandelbaum (2016) fails to establish that implicit attitudes, defined for the purpose of this chapter as those mental states that drive people's spontaneous evaluative responses such as those on indirect measures of attitudes, are not associatively but propositionally structured. Even if we take his interpretation of the empirical data at face value, he cannot establish that *all* so-called implicit attitudes are non-associative (but propositional; see section 2.2.2.1). Moreover, there are alternative interpretations of the data available that are in fact compatible with the associative structure of so-called implicit attitudes (section 2.2.2.2 and section 2.2.2.3; see also Madva, 2016).

The foregoing does not provide a knock-down argument against a propositional account of implicit attitudes. All that I have shown is that the associative account of implicit attitudes remains a viable option. In fact, I grant that the currently available evidence can be accounted for on a propositional, an associative, or a "heterogeneous" model of implicit attitudes. On this latter account, what we describe as "implicit attitudes" (those mental states that drive people's implicit evaluative responses) are a heterogeneous class of mental states that may include both associative and propositional mental states (see Holroyd & Sweetman, 2016, for a related argument). If this is the case, we may want to eliminate the notion of an "implicit attitude" (Holroyd, Scaife, & Stafford, 2017b). One may hope that future experiments will provide more decisive data that will help us to choose between these different models.[62]

2.3 Control

In the foregoing, I elaborated on the mental structure of so-called implicit attitudes. I concluded that although the matter is far from settled, the available data is compatible with the claim that implicit attitudes are associative. Now, I will turn to the question of what sort of control we can exert over implicit attitudes. In the last chapter, I have shown that it is often claimed that implicit attitudes are not subject to rational and/or not subject to intentional control. In what follows, I will argue that these claims are not quite

[62] See Madva (2016: 2679) for some suggestions of experiments that may settle the issue.

right even if we assume that implicit attitudes are associative. In section 2.3.1, I will argue that the idea that propositional thought can modify associative mental states implies that we can take, at least to some extent, *indirect rational control* over our implicit attitudes. In section 2.3.2, I will argue that the fact that propositional thought can activate associations implies that we have, at least to some extent, *indirect intentional control* over our implicit attitudes. In section 2.3.3, I will show that what I refer to as "indirect rational control" and "indirect intentional control" can be regarded as subtypes of "ecological control" (Clark, 2007; Holroyd & Kelly, 2016). Based on Holroyd & Kelly (2016), I will argue that the fact that we have ecological control in regard to our implicit attitudes undermines the claim, made by some proponents of the standard view (e.g., Levy, 2014a, 2015; Glasgow, 2016), that implicit attitudes cannot form part of people's moral characters.

2.3.1 Indirect rational control

As I showed in the last chapter (section 1.2.3), some authors have argued that explicit attitudes are acquired and changed in accordance with what the agent considers to be good reasons (i.e., they are under the agent's rational control), while implicit attitudes are insensitive to such reasons (i.e., they are not under the agent's rational control; Gendler, 2008a, 2008b; Levy, 2014a). In what follows, I will stress that this picture is not quite right even if we accept that implicit attitudes are associatively structured.

Recall that I argued above that propositional thought can play a role in the modification of associations (section 2.2.2.2). Note that if propositional mental states are reason-responsive and if associations change in accordance with propositional mental states, associations can be said to be *indirectly* reason-responsive. If we change our patterns of propositional thought because we see good reasons for that, this may, at least in the long term, lead to changes in our associative mental states. A person who associates men more strongly with leadership abilities than women may deliberately think more often about the leadership abilities of women in order to modify this association. To be sure, the prospects of such an intervention are limited because a given association (e.g., the association of men rather than women with leadership) may be reinforced by contingencies in a person's external environment (e.g., by the fact that men are actually more often to be found in leadership positions than women). However, it should also be noted that we can, to some extent at least, deliberately modify our external environment in such a way to bring about a desired change in our associations (Holroyd & Kelly, 2016: 121-122). One could, for example, put up photos of famous female leaders in one's office to strengthen one's association between women and leadership, or try to diversify one's circle of friends in order to combat one's

negative associations concerning people with backgrounds other than one's own (e.g., different ethnic or socio-economic background).

In a nutshell, if we see good reason to change our associations, we can indirectly achieve this by restructuring our external environment and by restructuring our internal thought patterns. Note that these strategies may well reinforce each other. Restructuring our external environment may lead to corresponding changes in our thought patterns (e.g., seeing photos of female leaders in our office may trigger thoughts about the leadership abilities of women) and changes in our thought patterns may motivate us to restructure our environment (e.g., thoughts about the leadership abilities of women may make it more likely that one will support women applying for leadership positions in one's organisation).

2.3.2 Indirect intentional control

Let us now turn from rational control to intentional control. In the last chapter, I pointed out that some proponents of the standard view have argued that the distinction between implicit and explicit attitudes corresponds to a distinction between automatic and intentionally controlled mental states (Devine, 1989; Rydell & McConnell, 2006; Wilson, Lindsey, & Schooler, 2000; see section 1.2.4). In particular, it is claimed that the activation of implicit attitudes, and their influence on behaviour, may proceed without the subject's intention and attentional focus. While it is certainly right that associative mental states can be activated automatically, it is important to acknowledge that they can also be activated, and indeed be suppressed, in an indirectly controlled manner. This is a direct result of the fact that propositional thinking can activate associative mental states (mentioned in section 2.2.2.3). Sarah, for example, could activate stereotypic associations in regard to black people by entertaining negative thoughts about black people, if she wanted to. Note that Sarah may be aware of the fact that she harbours these associations and thus be able to activate these associations intentionally.[63] More importantly for present purposes, Sarah may be able to inhibit the activation of associations that she feels alienated by. For example, by deliberately thinking "black people are peaceful", she may trigger the activation of associations between her black people concept and positive attributes, and suppress the activation of negative associations (e.g., her association between BLACK PERSON and VIOLENCE) and her fear reaction.[64, 65]

[63] See section 1.2.5 in the previous chapter for evidence that people can become aware of their so-called implicit attitudes.

[64] See also Carruthers (2009b), who argues that although emotions are issued in system 1 (the automatic system) they can be under the subject's intentional control (p. 124).

Evidence that supports the view that people can indirectly suppress the activation of so-called implicit attitudes comes from the literature on implementation intentions (e.g., Stewart & Payne, 2008; Mendoza, Gollwitzer, & Amodio, 2010; see also Holroyd & Kelly, 2016: 122). An implementation intention is a person's resolution to respond in a specified way whenever a particular situation obtains. For example, Stewart & Payne (2008) instructed their participants to think the word "safe" whenever they would see a black face on the screen in front of them. Strikingly, these participants did not exhibit a bias against black people on a weapon identification task that was found amongst participants in a control condition, who were instructed to think stereotype irrelevant words ("accurate" or "quick") in response to the presentation of black faces (see experiments 1 and 2).[66] That is, while participants in the control condition were more likely to identify a briefly presented object mistakenly as gun (rather than as tool) when they had been primed with the picture of a black face than when they had been primed with the picture of a white face, participants in the "think safe" condition were equally likely to mistakenly identify an object as gun in the black face and the white face condition. Presumably, thinking "safe" in response to black people neutralised the influence of negative black stereotypes (such as "black people are violent"), and negative affect (anxiety) that is linked to these stereotypes, on people's performance on the weapon identification task.[67]

Note that there are two possible mechanisms that could lead to this result. Firstly, thinking "safe" in response to the presentation of black faces may incline people to overrule the output of automatic stereotyping and affective processes when deciding how to respond.[68] Secondly, thinking "safe" in response to the presentation of black faces may inhibit the activation of problematic stereotypes and negative affect in the first place. Note that only the latter mechanism can count as indirect intentional control over so-called implicit attitudes themselves. Using a statistical procedure that allows dissociating different processes that contribute to people's performance on the weapon identification task (a process dissociation analysis), Stewart & Payne (2008) found evidence that the latter mechanism (and not the former) drove their participants' responses. This conclusion was further supported by a reaction time analysis. Participants in the think "safe" condition did not respond more slowly than participants

[65] See chapter 3 for an in-depth analysis of how mental stereotypes, such as the association between BLACK PEOPLE and VIOLENCE, relate to affective responses, such as fear (and vice versa).

[66] The words "accurate" and "quick" were chosen because participants in all groups were instructed to respond accurately and quickly.

[67] See Correll and colleagues (2002, study 3) for evidence that the so-called shooter bias, which is similar to the weapon identification bias, is indeed the function of common black stereotypes.

[68] Note that this would be similar to people's intentional effort to name the colour of a colour word on the stroop task and not to read out the colour word (see section 1.2.4 in the previous chapter).

in the control condition, which indicates that they did not deliberate more extensively about how to respond. The evidence thus suggests that implementation intentions allow people to suppress the activation of so-called implicit attitudes and not just to inhibit the influence of these mental states, once activated, on behaviour. In short, people have indirect intentional control over their implicit attitudes. The control is indirect in the sense that people can directly control what they think (e.g., the word "safe") and this in turn influences which associations become activated.

2.3.3 Ecological control and moral character

What I have described in the above two sections as "indirect rational control" and as "indirect intentional control" can be seen as subtypes of what Holroyd & Kelly (2016) call "ecological control", drawing on Clark (2007). Holroyd & Kelly (2016) define ecological control as follows:

> Ecological control is the structuring of one's environment and cognitive habits such that autonomous processes and subsystems can effectively fulfil one's person-level goals. (p. 130)

In the case of implicit attitudes, the relevant autonomous processes may be associative processes and the relevant subsystem may be an associative system. The person-level goal could be the goal not to associate a given social group with negative attributes (where "associate" can be understood in a dispositional or an occurrent sense). In the last two sections, we have seen examples of how a person may achieve this goal by structuring her environment (e.g., by putting up photos of famous female leaders in her office) or her cognitive habits (e.g., by restructuring her propositional thought or by internalising implementation intentions). In short, we have seen examples of how people can take ecological control of their implicit attitudes.

Importantly, Holroyd and Kelly (2016) argue that the fact that people can exert ecological control over their implicit attitudes (or "implicit biases" as they call them) suggests that implicit attitudes can be "proper targets of character-based evaluation" (p. 123; see also Holroyd, 2012). Since ecological control endows agents with the ability to modify implicit attitudes, the presence of problematic implicit attitudes that fuel unfair treatments of other people may reflect negatively on the agent. As Holroyd and Kelly (2016) put it, "there is a real sense in which whether or not [implicit attitudes] influence an individual's behaviour is very much a reflection of that person's character" (p. 126).

Holroyd & Kelly (2016) admit that whether implicit attitudes are in fact appropriate targets of character-based evaluation may additionally depend on certain epistemic conditions, such as whether the agent is aware of her implicit attitudes and aware of

the fact that she could take ecological control of them (pp. 126-127). Yet, recall that I have argued in section 1.2.5 that much speaks in fact for the claim that people can become aware of their implicit attitudes just as they can become aware of their explicit attitudes. Moreover, I would like to emphasise that knowledge about ecological control mechanisms is actually very widespread (even though most people are of course not familiar with the term "ecological control"). It is common knowledge that the environment that we expose ourselves to and our own thought patterns can indirectly shape our associations, automatic processes, or feelings. In fact, Holroyd and Kelly (2016) emphasise that ecological control does not only play a role in regard to implicit attitudes but "underlies a vast swathe of human behaviour and problem-solving" (p. 123). To name just a few examples, people may rearrange files in their office (structuring of environment) to increase their productivity (person-level goal; Clark, 2007), they may form the intention to remove distracting items, such as their mobile phone, from their desk (structuring of environment) to better be able to sustain concentration (person-level goal), or they may remind themselves of positive life events (structuring of cognitive habit) to improve their mood (person-level goal). These examples show that people are familiar with, and indeed routinely employ, ecological control mechanisms.

Denying that those mental states and processes that can only be ecologically controlled (and not be directly controlled) can be subject to character-based evaluation would imply that many mundane cognitive states and processes are not an appropriate target of such an evaluation. I agree with Holroyd & Kelly (2016) that this would be an untenable conclusion. A person's productivity (or unproductivity), for example, may reflect on her character even if the person can only "ecologically control" those mechanisms that determine her productivity. Similarly, we should acknowledge that a person's implicit prejudice can reflect on that person's character even if the person can only "ecologically control" her implicit attitudes. This suggests that it is misguided to assume, as some proponents of the standard view have done (Levy, 2014a, 2015; Glasgow, 2016), that the distinction between implicit and explicit attitudes corresponds to a distinction between mental states that do not reflect on a person's moral character and mental states that do reflect on a person's moral character (see desideratum *D2* of a model of attitudes mentioned in the introduction to this thesis).

However, this may not yet convince everyone. In particular, one may want to reply that if a person does not identify with a particular implicit attitude, or if that implicit attitude does not conform to the person's considered values and rational judgments, the implicit attitude does not form part of the person's moral character, irrespective of whether the person could in principle take ecological control of the implicit attitude. I will deal with this argument in chapter 5 of this thesis, where I will corroborate the view that

even response dispositions that a person does not identify with can reflect on that person's moral character.

2.4 Conclusion

In this chapter, I scrutinised some core assumptions of the standard view of attitudes as I have presented it in the last chapter. Proponents of the standard view distinguish between implicit and explicit attitudes. Implicit attitudes are typically understood to be associative mental states, while explicit attitudes are regarded as propositional mental states. However, recently some authors have argued (or speculated) that implicit attitudes, too, are propositional mental states, or even fully fledged beliefs (De Houwer, 2014; Frankish, 2016; Hughes, Barnes-Holmes, & De Houwer, 2011; Levy, 2015; Mandelbaum, 2016; Webber, 2016a). I focussed in this chapter on the argument put forward by Mandelbaum (2016). I argued that Mandelbaum fails to establish that implicit attitudes (defined for the purpose of this chapter as those mental states that drive people's spontaneous evaluative responses) are not associative mental states but propositional mental states. Even if we grant that in the studies that Mandelbaum describes propositional mental states drove people's implicit evaluative responses, this does not establish that propositional mental states are always or in the majority of cases the driving force behind people's implicit evaluative responses. Moreover, I showed that even for the very effects that Mandelbaum describes there are alternative explanations available that are consistent with an associative account of implicit attitudes. Accordingly, proponents of the standard view could be right about the claim that implicit attitudes are associative after all.

I also scrutinised the assumptions that implicit attitudes are not subject to rational control and not subject to intentional control. I argued that these assumptions are not quite right even if we assume that implicit attitudes are associative mental states. People can structure their external environment and their internal propositional thought *to modify* associations (i.e., they can take indirect rational control of their associations) and they *can suppress and activate* associations by controlling their thoughts (i.e., they can take indirect intentional control of their associations). These indirect forms of control can be analysed as forms of ecological control (Clark, 2007; Holroyd & Kelly, 2016).

This leaves us with the possibility that implicit attitudes are associative mental states that can only indirectly (or ecologically) be controlled by the agent, while explicit attitudes are propositional mental states that can be directly controlled by the agent.[69]

[69] Note, however, that some authors argue that even paradigmatic examples of explicit attitudes, such as beliefs, are not (always) subject to direct forms of control (e.g., Hieronymi,

However, even if we grant that this is a plausible interpretation of the view that there are distinct implicit and explicit attitudes (i.e., the standard view), the question remains whether this is in fact the best way to conceive of attitudes. In particular, it remains unclear why we should assume the existence of implicit and explicit attitudes, identified with individual mental states, rather than just the existence of implicit and explicit mental states (that may ground attitudes in one way or another).

Note that important motivations for distinguishing implicit and explicit attitudes do not hold up to scrutiny. Firstly, I challenged the assumption that the distinction between implicit and explicit attitudes, as it is usually drawn, corresponds to a distinction between mental states that form part of a person's moral character and mental states that do not form part of a person's moral character (see desideratum *D2* of a model of attitudes). This is because people can exert ecological control over their implicit attitudes (even if these are associative mental states; see section 2.3.3). Accordingly, implicit attitudes may reflect on a person's moral character. This undermines an important motivation to distinguish between implicit and explicit attitudes. In fact, drawing a distinction between implicit and explicit attitudes may create confusion because it may wrongly suggest that people can only be evaluated for their so-called explicit attitudes. Secondly, I argued in the last chapter that results of attitude measurements on indirect and direct measures fail to motivate a distinction between implicit and explicit attitudes. The mere fact that results on indirect and direct measures of attitudes are sometimes dissociated does not establish that there are two different kinds of attitudes (see section 1.3.2). Thirdly, I showed in the last chapter that according to Oswald and colleagues (2013) results on IATs (presumably the most popular indirect measure of attitudes) and results on direct measures of attitudes do not differ in their relative success of predicting spontaneous evaluative responses versus deliberate evaluative responses (see section 1.3.3). This suggests that the postulation of two different kinds of attitudes (as measured by indirect and direct measures of attitudes) may not actually be necessary in order to predict and explain people's evaluative responses (see desideratum *D1* of a model of attitudes).

Against the standard view speaks also that the identification of attitudes with individual mental states is out of line with the folk psychological understanding of attitudes (see section 1.2.2). When we say that someone has a racist attitude (or is a racist), we do not seem to highlight a particular belief or association of the agent but seem to refer to a general trait of the agent. Of course, psychologists and philosophers are not required to (and may sometimes have good reason not to) employ the same concepts as folk psychologists. Yet, as I have mentioned in the introduction to this

---

2008; Holroyd, 2012). If these arguments are right, this puts pressure on the claim that control provides a criterion that would allow us to distinguish between explicit and implicit attitudes.

thesis, scholars working on important issues such as racism will find it immensely difficult to inform public discourse with their research if their use of the term “attitude” is very different from how the term is used in day-to-day discourse. It is thus worth examining whether there is a scientifically sound model of attitudes available that better corresponds to the folk conception of attitudes than the standard view and that may also appeal to psychologists as well as philosophers (see desideratum *D3* as mentioned in the introduction to this thesis).

In chapter 4, I will propose such a model. Building upon Machery (2016), I will argue that attitudes are complex traits of people. On this proposed model, the implicit-explicit distinction does not apply to attitudes, but each attitude is typically grounded in a variety of implicit and explicit mental states. In particular, I will argue that attitudes conceived as traits can be analysed as characteristic profiles of situation-specific evaluative response dispositions. The proposed model does justice to the fact (already touched upon in section 2.3.2) that evaluative responses towards a social group are highly context-dependent. Only a model of attitudes that acknowledges this context sensitivity can appropriately fulfil an explanatory/predictive function (see desideratum *D1* for a model of attitudes). Moreover, I will argue that attitude ascriptions as conceived on my model provide an accurate insight into a person’s moral character (see desideratum *D2*).

However, before I turn to my model of attitudes in chapter 4, I will elaborate in the next chapter on the relationship between cognitive stereotypes and affect (e.g., the relationship between Sarah’s conceptual association between BLACK PERSON and VIOLENCE and her affective reaction to be afraid of black people). This will allow me to draw some further conclusions about the mental states that underpin attitudes (see question *Q2* in the introduction to this thesis) and about attitude individuation (see question *Q1*). This, in turn, will help me to further motivate my model of attitudes. It will become clear that conceptual mental states (i.e., stereotypes) and affective mental states are causally so tightly linked to each other that it does not make sense to identify attitudes with either conceptual or affective mental states alone. Rather we should acknowledge that attitudes are grounded in clusters of mental states if we want the notion of an attitude to optimally fulfil an explanatory/predictive function.

Chapter 3: The relationship between mental stereotypes and affect

3.1 Introduction

In the last two chapters, I have elaborated on the assumed difference between implicit and explicit evaluative mental states (which some scholars have described as a difference between implicit and explicit attitudes). In this chapter, I will turn to another distinction that is often made in regard to those mental states that are candidate (components of) attitudes. It is common in the philosophy and psychology of inter-group relations to draw a distinction between stereotypes about a social group, on the one hand, and the affect that is elicited by and directed at the group, on the other hand (e.g., Blum, 2004, 2009; Greenwald & Banaji, 1995; Sie & van Voorst Vader-Bours, 2016). Sarah, for example, may associate black people with violence (or she may harbour the propositional mental state with the content "black people are violent"), which constitutes a stereotype, and at the same time *feel* scared of black people, which is an affective response. There are different views about what kind of cognitive structures underpin stereotypes (see Beeghly, 2015, for a review of these), but all these views have in common that they understand stereotypes as mental entities that link representations of a social group (e.g., black people) to representations of particular attributes (e.g., violence, laziness, athleticism).[70] The process of stereotyping, in a minimal sense, can be understood as the momentary activation of these stereotypes in a person's mind (Beeghly, 2015).[71] Affect, by contrast, includes basic feelings of like or dislike or fully-fledged emotions, such as anger, disgust, fear, or pity. Often the term "prejudice" is used in the literature to denote negative affective responses to an outgroup, and to contrast these with stereotypes (e.g., Blum, 2004;

[70] In line with Beeghly's (2015) "descriptive view" of stereotypes, I will use the term "stereotype" in this chapter in the broad sense of a trait assignment to a social group. Other authors have proposed additional criteria that are required for a trait assignment to qualify as a stereotype. For example, it has been argued that stereotypes are morally defective, false generalisations about social groups that are largely resistant to counterevidence (Blum, 2004) and that stereotypes are "socially shared cultural constructs" (Sie & van Voorst Vader-Bours, 2016: 94). The argument that I make in this chapter is not contingent on any particular stereotype definition. Beeghly (2015) notes that the term "stereotype" is ambiguous in so far as it may refer to a cluster of traits assigned to a social group ("the entire informational structure", p. 680) as well as to individual traits assigned to a social group ("parts of that structure", ibid). In this chapter, I will be concerned with individual trait assignments, if not mentioned differently.

[71] Beeghly (2015) distinguishes four different views about the nature of stereotyping that one may also interpret as stages in the process of stereotyping: stereotyping as momentary stereotype activation; stereotyping as stereotype use; primary influence of stereotypes on thoughts, emotions, and action as stereotyping; and stereotyping as stereotype communication.

Judd, Blair, & Chapleau, 2004; Sie & van Voorst Vader-Bours, 2016).[72, 73] The distinction between stereotypes and affect in regard to social groups (i.e., prejudice) can be seen as an instance of the general distinction between cognition and affect (Amodio, 2008).

Many scholars assume that the distinction between stereotypes about social groups and affect towards people qua members of social groups (henceforth referred to as "affect towards social groups" or simply "social affect") is not only a conceptual distinction but that these concepts in fact correspond to distinct mental kinds (e.g., Amodio, 2008; Judd, Blair, & Chapleau, 2004; Valian, 2005). I will follow Madva and Brownstein (2016) in calling this the "two-type model". On this view, Sarah's association between BLACK PERSON and VIOLENCE can in principle be separated (both conceptually and in terms of the kinds of mental states involved) from her fear response towards black people (and indeed any affective response towards black people). Accordingly, it would, at least in principle, be possible for her to have the stereotype activated without being in an affective state of fear (or without being in any affective state), and conversely, to be in an affective state of fear towards a black person without having any stereotype activated. Consequently, it is sometimes claimed that stereotypes are "cold" cognitive mental states, while prejudices are "hot" affective-motivational mental states (Valian, 2005).[74] Note that how we conceive of the relation between stereotypes and social affect has a bearing on the question about attitude individuation (see *Q1* in the introduction to this thesis). If the "two-type model" is right, we may ask whether stereotypes, or social affect (i.e., prejudice), or both of these components constitute people's attitudes.

However, some scholars have argued that the distinction between "cold" stereotypes and "hot" social affect does not hold up (Holroyd & Sweetman, 2016; Madva & Brownstein, 2016). They have argued that stereotypes inherently possess an affective valence and that social affect inherently possesses stereotypic conceptual content.[75] That is, for both what is commonly referred to as "stereotype" and for what is

[72] I use the term "outgroup" here as it is commonly used in social psychology to denote a social group that the person who we are referring to does not consider herself to be member of. For example, Muslims are a religious outgroup to a person who considers herself to be Christian. Conversely, Christians are that person's religious "ingroup".

[73] Blum (2004) notes that "stereotyping is not the same as prejudice, and neither requires the other" (footnote 4). Yet it shall also be mentioned that Blum (2009), by contrast, describes stereotypes as one component (alongside affect) of prejudice.

[74] Another author who adopts the "hot vs. cold-metaphor" is Anderson (2010). She claims that stereotypes "are *more* a matter of 'cold' cognitive processing than 'hot' emotion" (p. 45, my emphasis). This may suggest a gradual difference between "cold" stereotyping and "hot" emotion. Accordingly, I do not regard her, as Madva and Brownstein (2016) do, as a (clear) proponent of a two-type view.

[75] Wittenbrink and colleagues (1997) seem to defend at least the first of these two claims when they posit that implicit stereotypes "are colored by their valences, so that stereotyping and prejudice on the implicit level are conceptually intertwined" (p. 271).

commonly referred to as “prejudice” there are both affect and conceptual content involved.[76] Let us call this, again following Madva and Brownstein (2016), the “one-type model”. If this view is right, it is not sensible to ask whether stereotypes or social affect constitute our attitudes because these components cannot be separated.

In this chapter, I will argue that one-type theorists are right in so far as stereotypes about social groups and affects towards social groups form tight clusters. They are parts of a mental kind that I will call in accordance with Madva and Brownstein (2016) “evaluative stereotype”. This being said, I will also point out, contra to Madva and Brownstein (2016), that these clusters are not a unified mental state but are composed of different kinds of mental states (e.g., conceptual mental states and affective mental states) that are causally interconnected. Due to the tight causal connections between these mental states it is appropriate to say that stereotypes have an affective quality and that affect towards social groups has a conceptual or stereotypic quality. I will conclude that we need to acknowledge that attitudes are jointly constituted by conceptual (stereotypic) and affective mental states (plus maybe yet other kinds of mental states) if we want the notion of an attitude to optimally fulfil an explanatory and predictive role (see desideratum *D1* of a model of attitudes).

This chapter is structured as follows. In section 3.2, I will present the empirical evidence that Amodio and his colleagues have provided in support of the two-type model (Amodio & Devine, 2006; Amodio & Hamilton, 2012). In section 3.3, I will argue that there is an alternative explanation for Amodio and colleagues findings. Their findings can be explained by the operation of specific “evaluative stereotypes” (a term that I borrow from Madva and Brownstein, 2016) rather than the separate operation of evaluations (i.e., social affect) and stereotypes (section 3.3.1). I will show that it improves our predictions of discriminatory behaviour when we focus on the interaction between stereotypes and affective responses rather than emphasising their separateness (section 3.3.2). Moreover, I will argue that differential effects of induced emotions on IAT results are best explained on the assumption that there are evaluative stereotypes (see section 3.3.3). This leaves open the question of whether evaluative stereotypes are in fact unified mental states that blend conceptual and affective content, or whether they are constituted by distinct, but causally tightly linked, conceptual and affective mental states. I will address this question in section 3.4. In

[76] It is important to note that it is logically possible that “stereotype” and “prejudice” are different concepts although they refer to the same entity. This case would be analogous to Frege’s (1948) famous example of the words “morning star” and “evening star”, which he takes to have different senses while having the same referent, i.e. the planet Venus. The word “morning star” is roughly used in the sense “bright star that can be observed in the morning”, while “evening star” is used in the sense “bright star that can be observed in the evening”, and as it happens theses senses pick out the same object. Just as astronomical observations were needed to find out that the morning star and the evening star are the same planet, psychological experiments might inform us that stereotypes and prejudices are the same psychological kind after all.

section 3.4.1, I will discuss and reject Madva and Brownstein's (2016) one-type view according to which evaluative stereotypes "are best conceived in terms of mutually co-activating semantic-affective-behavioral 'clusters' or 'bundles'" that cannot be broken apart into more primitive mental states (p. 1). I will reply that Madva and Brownstein's "bundles" can better be understood as separate, but interacting, conceptual and affective mental states. In section 3.4.2, I will argue that due to the tight causal connections between these conceptual and affective mental states, one-type theorists are nonetheless right about the claim that stereotypes about social groups have an affective quality and that affect towards social groups has a conceptual (or stereotypic) quality. In section 3.5, I will then summarise my nuanced position on the relation between stereotypes and affect, and conclude that attitudes should be conceived of as being jointly constituted by conceptual and affective mental states (plus perhaps other kinds of mental states).

Before I start, one cautionary note is in order. In this chapter, I will mainly be concerned with mental states that scholars commonly describe as "implicit" (roughly, those mental states that drive spontaneous evaluative responses such as those that are expressed on indirect measures of attitudes) and not with mental states generally described as "explicit". I am confident that the conclusions that I reach in this chapter on the relation between implicit stereotypes and prejudices can be extended to mental states more commonly described as explicit.[77] Yet regardless of this possible extension, my argument as presented in this chapter lends support to the view that attitudes should not be identified with individual mental states but are better conceived of as grounded in clusters of mental states.

3.2 Empirical support for the two-type model

The empirical case for the distinctness of so-called implicit stereotypes and prejudices has most forcefully been made by Amodio and his colleagues (Amodio, 2008, 2014; Amodio & Devine, 2006; Amodio & Hamilton, 2012; Amodio & Ratner, 2011). It must be noted that they use the term "implicit evaluation" instead of "implicit prejudice" for affective responses towards social groups in order to "avoid invoking unintended connotations associated with the complicated construct of prejudice, such as consciously endorsed racist attitudes and beliefs." (Amodio & Devine, 2006: footnote 1). In what follows, I will adopt this terminology, which is also useful because "implicit

[77] Note also that it is not even clear whether there is something like *explicit* affect. Affective mental states do not exhibit those features that are usually seen as characteristic of explicit mental states: they are presumably not propositionally structured and neither subject to direct rational nor direct intentional control.

evaluation" may refer to both positive and negative affect, whereas the term "implicit prejudice" is typically connoted with negative affect.

A significant part of Amodio and colleagues' argument is based on neuroscientific evidence that supposedly shows that there are distinct neural systems for what they call semantic (i.e., conceptual) and affective processing (Amodio, 2008, 2014; Amodio & Ratner, 2011).[78] It is argued that this neural distinction supports a corresponding psychological distinction between conceptual stereotypes and affective evaluations. However, it is highly contentious whether such an inference from the distinction between neural systems to the existence of two distinct psychological constructs is valid.[79] In what follows, I will therefore not elaborate on the evidence for distinct neural underpinnings of stereotypes and prejudices but rather describe the direct psychological evidence for the claim that stereotypes and prejudices are distinct mental kinds. That is, I focus on what I take to be the strongest case for the two-type model. This is also in line with my general interest in the mental states (and *not* neural states) underpinning attitudes (see question *Q2* in the introduction of this thesis).

Amodio and colleagues discuss a range of psychological experiments that allegedly support the view that implicit stereotypes and implicit evaluations are distinct mental kinds (Amodio & Devine, 2006; Amodio & Hamilton, 2012). The alleged evidence for what for what I call in accordance with Madva & Brownstein (2016) "the two-type model" is threefold. Firstly, Amodio and Devine (2006) claim that people's results on IATs that are designed to assess implicit stereotypes do not correlate with people's results on IATs that are designed to assess implicit evaluations. Secondly, they assume that implicit evaluations and implicit stereotypes as measured with these different IATs are predictive of different kinds of behaviours. Thirdly, Amodio and Hamilton (2012) provide evidence that social anxiety induced by the prospect of an upcoming interaction with a black person affects scores on a racial evaluative IAT but not those on a racial stereotype IAT. In the following paragraphs, I will describe these pieces of evidence in turn. In the next section, I will then scrutinise the evidence and show that there is an alternative explanation for these findings that is compatible with a one-type view.

---

[78] In what follows I will use the term "conceptual" for what Amodio and colleagues call "semantic". The term "semantic" is ambiguous and what they actually seem to refer to is conceptual processing.

[79] Many contemporary philosophers of mind (and cognitive scientists) defend some form of non-reductive materialism according to which psychological states supervene on neural states but are not identical with them (Baker, 2009). For them, a difference between psychological kinds always implies a difference in the underlying neural states, whereas the reverse inference from a difference in neural states to a difference in psychological kinds is not valid. By contrast, identity theorists would challenge the very distinction between the psychological and the neural level because they believe that psychological states are identical to brain states (Lewis, 1994). See also section 6.2.8 in Madva and Brownstein (2016) for a discussion of the relation between psychological-level and neural-level distinctions.

Amodio and Devine (2006) created two different IATs to assess implicit evaluations and implicit stereotypes separately and to prove the independence of these psychological constructs (see study 1). The evaluative IAT (henceforth "Eval-IAT") was designed to measure how readily the white participants in the study associate white and black faces with pleasant and unpleasant words.[80] The pleasant words included "honor, lucky, diamond, loyal, freedom, rainbow, love, honest, peace, and heaven" and the unpleasant words included "evil, cancer, sickness, disaster, poverty, vomit, bomb, rotten, abuse, and murder" (p. 654). As predicted, the results indicated a significant evaluative bias against black people, presumably showing that participants more readily linked black people to unpleasantness than to pleasantness, and white people more readily to pleasantness than to unpleasantness.[81] The stereotype IAT (henceforth "Stereo-IAT") was designed to measure how strongly the same participants associate white and black faces with "mental" or "physical words". The "mental words" included "math, brainy, aptitude, educated, scientist, smart, college, genius, book, and read" and the "physical words" included "athletic, boxing, basketball, run, agile, dance, jump, rhythmic, track, and football" (p. 654). As predicted, the results were consistent with the assumption that white people are more strongly associated with mental attributes than with physical attributes and that black people are more strongly linked to physical attributes than to mental attributes.[82] Yet, although on both the Eval-IAT and the Stereo-IAT a significant bias was detected, participants' scores on these measures were uncorrelated. That means that participants who showed an evaluative bias against black people on the Eval-IAT did not reliably show stereotypic responses on the Stereo-IAT, and vice versa. Amodio and Devine argue that if evaluation and stereotyping were a unified mental kind, a correlation between the results on the Eval-IAT and the Stereo-IAT should be detectable. However, as no such correlation was found, evaluation and stereotypes must be seen as distinct mental kinds. Stereotypes can exert their influence on behaviour without affect-involving evaluative responses playing a role and, conversely, social affect can influence responses without stereotyping.

In two further studies, Amodio and Devine (2006) tested their hypothesis that implicit stereotypes and implicit evaluations are uniquely predictive of different kinds of behaviours (see studies 2 and 3). Based on previous research on explicit stereotypes

[80] Note that some authors call this kind of IAT "attitude IAT" (e.g., Oswald et al., 2013; Rudman & Ashmore, 2007).

[81] Amodio and Devine (2006) only report a composite Eval-IAT score. That is, in fact we do not know whether the effect found on the Eval-IAT is due to participants linking black people more readily to unpleasantness than to pleasantness, linking white people more readily to pleasantness than to unpleasantness, or due to a combination of both these factors.

[82] Again, Amodio and Devine (2006) only report a composite score. That is, in fact we do not know whether the effect found on the Stereo-IAT is due to a stronger association of white people with mental attributes than with physical attributes, a stronger association of black people with physical attributes than with mental attributes, or due to a combination of both these factors.

and explicit evaluation (e.g., Dovidio et al., 1996; Dovidio et al., 2003), they expected that implicit stereotypes would predict judgment formation about a social group (what they call "instrumental behaviour"), whereas implicit evaluations would predict basic approach or avoidance responses (what they refer to as "consummatory behaviours"). One of these studies (study 3) consisted of two sessions, in the first of which the Eval-IAT and Stereo-IAT were administered and in the second of which the measures of instrumental and consummatory behaviours were taken. The results showed that Eval-IAT scores were uniquely predictive of participant's seating distance to the belongings of an African American interaction partner (consummatory behaviour), whereas the Stereo-IAT scores were uniquely predictive of participants' assumptions about the interaction partner's performance on various academic and non-academic tasks (instrumental behaviour). The higher participants' evaluative bias against black people on the Eval-IAT, the further they sat away from the belongings of the African American, while the same pattern in the evaluative IAT results did not correlate with judgments about the black interaction partner's task performance. By contrast, the higher participants' stereotype bias (i.e., associating black people more strongly with physical than with mental words), the worse they assumed the interaction partner would perform on academic tasks, while this variation in stereotype IAT results did not correlate with seating distance.[83] These findings allegedly lend support to the view that stereotypes and evaluations (i.e., affect towards social groups) have different functional roles and thus are different mental kinds.

If stereotypes and evaluations have different functional profiles, we should also expect that they respond differently to situational input. This has been examined by Amodio and Hamilton (2012). They led half of their white participants to believe that they were about to interact with a black person (black partner condition), while the other half was led to believe that their interaction partner would be a white person (white partner condition). Those participants in the black partner condition showed subsequently more bias against black people on an Eval-IAT than those participants in the white partner condition. However, the two groups exhibited no difference in their biases on the Stereo-IAT. Crucially, self-reported feelings of anxiety were stronger for participants in the black partner condition than for participants in the white partner

[83] In the other study (Amodio & Devine, 2006: study 2), participants read a short writing sample of either a black or a white writer and were asked to form an impression of the writer. Afterwards, instrumental behaviour was assessed by asking participants to indicate which attributes they would use to describe the author and consummatory behaviour was measured by asking whether they would like to befriend the author as well as by asking them to rate on a feeling thermometer how warm they feel towards the author. Subsequently, the same participants completed a Stereo-IAT and an Eval-IAT. As predicted, a statistical procedure (hierarchical linear regression) revealed that the Stereo-IAT results were linked to participant's instrumental behaviour (whether they would describe the author in stereotype-conforming terms) but not to their consummatory behaviours, whereas the Eval-IAT results were predictive of the consummatory behaviours but not of the instrumental behaviours.

condition. Moreover, in the black partner condition, participants' level of anxiety was correlated with their Eval-IAT scores but not with their Stereo-IAT scores. For participants in the white partner condition, no such correlation could be found. Amodio and Hamilton conclude that social anxiety affects social evaluation (which involves affect) but not stereotyping. If evaluation and stereotyping were one psychological kind, we should expect that both are affected by the same factors. Amodio and Hamilton's experiment seemingly shows that there is at least one factor (i.e., anxiety) that affects evaluation selectively. Thus it seems that stereotyping and evaluation are functionally different.

To summarise, there are different pieces of evidence that have been taken to bolster the two-type model: results on measures of implicit stereotypes have been shown to be independent of results on measures of implicit evaluations. Moreover, stereotypes and evaluations seem to have different functional profiles: they contribute to different forms of behaviour and are not affected in the same ways by the same input.

3.3 Assessing the evidence for the two-type view

The argument that Amodio and colleagues (Amodio & Devine, 2006; Amodio & Hamilton, 2012) provide can be interpreted as an argument to the best explanation. They argue that the hypothesis that stereotypes and evaluations are distinct entities is the best explanation for (1) the low correlation between people's results on the Stereo-IAT and the Eval-IAT, (2) the fact that Stereo-IAT and Eval-IAT results predict different kinds of behaviours, and (3) the fact that results on Stereo-IAT and Eval-IAT are influenced differently by feelings of anxiety. In the following, I will argue that the reported findings can equally well be explained on a model that emphasises that there are tight links between stereotypes and evaluations, and that is compatible with a one-type view. That is, I will argue that the findings can be explained by the operation of particular "evaluative stereotypes" (a term that I borrow from Madva & Brownstein, 2016) rather than the separate operation of evaluations and stereotypes. I will make this case for each of Amodio and colleagues' pieces of evidence separately, starting with the evidence from the lack of correlation between Eval-IAT and Stereo-IAT results (section 3.3.1), followed by the evidence from behavioural prediction (section 3.3.2), and the evidence from the influence of social anxiety on social evaluation (section 3.3.3). At the same time, I will argue that a model that is based on the notion of evaluative stereotypes may lead research into more fruitful directions (section 3.3.2) and that it can explain at least one finding, from a different study (Dasgupta et al.

2009), that cannot be explained on a model that holds that evaluations and stereotypes are largely unrelated constructs (section 3.3.3).

3.3.1 Lack of correlation between Eval-IAT and Stereo-IAT results

There are two concerns that have been raised about Amodio and Devine's (2006) interpretation of the absence of correlation between Eval-IAT and Stereo-IAT results. Firstly, both IATs seem to utilise words that have at the same time conceptual and affective qualities and thus it is unclear how these IATs can track the distinction between conceptual stereotyping and affective evaluation (Holroyd & Sweetman, 2016). Secondly, the different IATs may not track the distinction between stereotyping and evaluation because the combination of face stimuli with word items on the two IATs may have triggered complex interactions between stereotypes and social affect (Madva & Brownstein, 2016). In section 3.3.1.1, I will address the first concern and argue that it can be dismissed once we properly understand the rationale behind Amodio and Devine's (2006) experiment. In section 3.3.1.2, I will elaborate on the second concern and insist that this indeed poses a challenge to Amodio and Devine's (2006) interpretation of their results.

3.3.1.1 Words with conceptual as well as affective content

Amodio and Devine (2006) take the non-correlation between Stereo-IAT results and Eval-IAT results to show that stereotypes and evaluations are functionally independent. People can stereotype social groups without having an affective response towards them and they can have an affective response towards other social groups without stereotyping them. However, it has been pointed out that both the words used for the Stereo-IAT and the words for the Eval-IAT possess conceptual content as well as an affective valence, and thus cannot track a difference between (affect-free) stereotyping and evaluation (Holroyd & Sweetman, 2016). Obviously, all the words on the Eval-IAT, such as the pleasant words "diamond" and "lucky" or the unpleasant words "evil" and "cancer", have a meaning that is not reducible to the valence of the word. They signify particular objects, such as diamonds, or states, such as being lucky. Similarly, all the words on the Stereo-IAT, such as the "mental words" "math" and "brainy" or the "physical words" "athletic" and "boxing", have both a particular meaning and a valence attached to them. Note, for example, that the positive valence of "athletic" can be seen if we compare it with other physical terms like "sluggish" or "weak". Importantly, the same is true for the category labels "physical" and "mental". Arguably, both these words

bring positive associations to mind (at least when considered in isolation).[84] Thus, Amodio and Devine's (2006) claim that these category labels are "relatively neutral" seems misplaced (p. 654).

Yet, properly understood, Amodio and Devine's claim is not that the lack of correlation between the scores on the Eval-IAT and the Stereo-IAT is due to the items on the Eval-IAT being purely affective and the items on the Stereo-IAT being purely conceptual.[85] Rather their claim seems to be that there is no correlation detectable because the effect on the Eval-IAT is driven by the affective aspects of the items used (and not by the conceptual aspects), whereas the effect on the Stereo-IAT is due to the conceptual aspects of the items used (and not due to their affective aspects). In short, on the Eval-IAT, affect is supposedly the "difference maker", whereas on the Stereo-IAT, conceptual content is supposedly the "difference maker". The rationale behind this claim is as follows. The effect found on the Stereo-IAT cannot be due to differences in affective responses because the physical and mental word lists used were of similar average valence (recall that physical words included for example "athletic" and "agile", while mental words included for example "brainy" and "educated"). By contrast, the Eval-IAT effect was due to different affective responses because the difference between the pleasant and unpleasant words was a difference in affective valence and not a difference in stereotypic conceptual content relating to black people (recall that pleasant words included for example "lucky" and "diamond", while unpleasant words included for example "cancer" and "disaster").[86] A pre-test, in which another group of participants rated the words in the word lists for their favourability, backed these claims about their average valences. Both the mental and physical words turned out to be positively valenced on average and, not very surprisingly, the pleasant words were rated much more positively than the words in the unpleasant word list.[87] Hence, according to Amodio & Devine (2006), the Eval-IAT and the Stereo-IAT may reveal a

[84] See next section (section 3.3.1.2) for an argument that "physical" can take on a negative valence when ascribed to black people.

[85] Madva and Brownstein (2016) seem to acknowledge this point (see endnote 10 in their paper). However, their discussion in the main text remains perplexingly out of line with this admission. In the main text, they continue to criticise that the category labels and words used on the Stereo-IAT are evaluatively-laden (p. 7).

[86] It should be noted, however, that some of the unpleasant words used (most notably "poverty", "abuse", and "murder") are presumably linked to common black stereotypes. This is an unfortunate oversight in Amodio and Devine's study design. I will grant for the sake of the argument that this oversight did not affect the results of their experiment and show below that even if this is granted, there is an alternative explanation for Amodio & Devine's results.

[87] Actually, things were a bit more complicated. Although both mental words and physical words were rated positively on average in the pre-test, the mental words were rated somewhat more favourably than physical words. Yet, the difference in favourability between the pleasant and unpleasant words of the Eval-IAT was much bigger than the difference in favourability between the mental words and the physical words of the Stereo-IAT. Moreover, Amodio and Devine (2006) used a statistical procedure (covariate analysis) to ensure that the effect on the Stereo-IAT was not driven by the difference in valence that was revealed on the pre-test. See footnote 3 in their paper on this.

difference between evaluation and stereotyping, even though the words used on these tests have both affective as well as conceptual qualities. Whereas the Stero-IAT effect is presumably due to a difference in stereotypic content (physical vs. mental) because the affective valence is held constant, the Eval-IAT effect is presumably due to a difference in affective valence (pleasantness vs. unpleasantness) because the words used presumably did not differ in stereotypic content relating to black people. As people's results on the two IATs are dissociated, Amodio and Devine (2006) conclude that stereotypes about black people and affect towards black people can operate independently of each other and are thus distinct mental kinds. Now that we have a better understanding of Amodio & Devine's argument, we can further assess its validity.

3.3.1.2 Results on both IATs are the result of "evaluative stereotypes"

In the last section, I pointed out that the fact that both the words used on the Stereo-IAT and the Eval-IAT have both affective and conceptual qualities does not undermine Amodio and Devine's (2006) claim that these tests tap into separate mental constructs: conceptual stereotypes and affective evaluations. However, this should not be taken to imply that dissociations between results on Eval-IAT and Stereo-IAT are in fact best accounted for by the separate operation of affective evaluations and conceptual stereotypes. In what follows, I will argue that on both the Eval-IAT and the Stereo-IAT, a combination of stereotypic conceptual content and affect may well have driven the observed effects. The combination of face and word stimuli on the two IATs presumably allows for complex interactions between the activation of stereotypes and affect, and different combinations of stereotypes and affect may have resulted in the dissociation between Stereo-IAT and Eval-IAT results. In other words, Amodio and Devine (2006) fail to establish that stereotypic content is the *sole* "difference maker" on the Stereo-IAT and that affect is the *sole* "difference maker" on the Eval-IAT. Consequently, they fail to establish that stereotypes about social groups and affect towards social groups can operate independently of each other.

Drawing on research by Degner and Wentura (2011), Madva and Brownstein (2016) point out that the valence that a given trait has for a person often depends on whom the trait is assigned to: "a trait like intelligence or being 'good at' some activity has a positive valence when it is attributed to oneself or one's ingroup, but a negative valence when attributed to an outgroup" (p. 7). Consequently, one can speculate that the mental attributes used on the Stereo-IAT might have taken on different valences when combined with black than when combined with white person stimuli. Amodio and Devine found that the valence of the mental attributes is positive on average when they tested them in isolation, but dependent on whom the trait is assigned to, the valence of

these traits may vary significantly. When the white participants were asked to respond with the same key on the keyboard to black person stimuli and to mental stimuli, this might have led to negative affect. By contrast, when participants were required to respond in the same way to white person stimuli and to mental stimuli, this might have led to positive affect. These evaluative responses might have contributed to the observed Stereo-IAT effect.

Similarly, the physical attributes might have taken on different valences for the two target groups on the Stereo-IAT. Blum (2004) rightly notes that the "[h]istorical and social context introduces an important level of complexity to the overall assessment of the content of a stereotype" (p. 278). As an example, he mentions that black people are often described as good dancers. On the face of it, this might seem to be a positive stereotype. However, seen in the historical context, this attribution suggests a negative evaluation. Drawing on work by Pickering (2001), Blum (2004) explains that this stereotype dates back to the slave era, where black people were seen as joyously entertaining subordinates, who are irrational, irresponsible, and lazy. Accordingly, the stereotype of black people as good dancers tends to invoke the deeply negative conception of black people as good at particular physical activities while being mentally weak. Even if people are ignorant about the origin of the stereotype, they may well be aware of its derogatory character (which is unbeknown to them historically rooted). By contrast, saying of a white person that she is a good dancer, or good at physical activities in general, does not carry this historical burden and can thus indeed be linked to a positive evaluation of the person (especially, if the evaluating person is white herself). Thus, it seems plausible that on Amodio and Devine's (2006) Stereo-IAT the pairing of white person stimuli with physical attributes evoked positive affect, whereas the pairing of black person stimuli with physical words was linked to negative evaluation of black people.

Taken together, these complexities that arise when mental and physical attributes are linked to black and white people shed doubts on the claim that differences in affective valence did not contribute to the Stereo-IAT effect in Amodio and Devine's (2006) study. Amodio and Devine take the Stereo-IAT to reveal "a pattern of stereotypic trait associations with Black and White faces" (p. 655). That is, participants presumably linked black faces more readily to physical attributes than to mental attributes and white faces more readily to mental attributes than to physical attributes.[88] Amodio and Devine seem to assume that these associations do not differ in affective valence because the

[88] Although, as already mentioned in footnote 82, it is not clear from the Stereo-IAT score reported by Amodio and Devine (2006) whether the effect is in fact due to people linking black people more readily with physical attributes than with mental attributes, linking white people more readily with mental attributes than with physical attributes, or due to a combination of both these factors.

attributes “physical” and “mental” are equally positively charged.[89] However, as I have argued above, this inference is misguided. Whereas an association between white people and mental attributes is likely linked to positive affect (at least in white people), an association between black people and physical attributes may well be linked to negative affect (at least in white people). All in all, the Stereo-IAT effect may reflect the affectively charged conception of black people as less intelligent than white people (and as more physical in a dehumanizing sense). Note that even the association of black people with physical attributes can be seen as expression of the demeaning conception that black people are mentally inferior. Results by Amodio and Hamilton (2012) support the suspicion that the Stereo-IAT is primarily driven by the conception of black people as unintelligent. They found a higher rate of attribute categorisation mistakes for the black-mental pairing than for any other pairing (black-physical, white-mental, and white-physical) on a Stereo-IAT, indicating that of all the pairings participants found it most difficult to link black people to mental attributes (see also Madva & Brownstein, 2016: 10). This is compatible with the interpretation that the Stereo-IAT effect is primarily driven by the negatively charged stereotype that black people are unintelligent.

So far I have argued that evaluations (i.e., affective responses) may well have contributed to the effect found on Amodio and Devine’s (2006) Stereo-IAT. Conversely, one may now wonder whether stereotyping may have influenced the Eval-IAT effect. In fact, the structure of the Eval-IAT does not rule out the possibility of stereotype content influencing the result. To be sure, most of the positive and negative word items used on the Eval-IAT (e.g., “diamond”, “peace”, “disaster”, and “vomit”) did not convey stereotype content.[90] However, it may well be that these words in conjunction with pictures of black and white persons activated stereotypes from memory, which matched the valences of the words. For example, when faces of white people were paired with positive words, this may have led to the activation of positive white person stereotypes. Likewise, when white faces were paired with negative attributes, this might have led to the activation of negative white person stereotypes. However, we can assume that people associate their own group more strongly with positive stereotypes than with negative stereotypes. Consequently, for participants in Amodio and Devine’s experiment, all of who were white, the pairing of positive words with white faces may have been more effective in triggering correspondingly valenced stereotypes than the

[89] It should be noted that although I speak here in accordance with Amodio & Devine (2006) loosely of associations, this is not meant to preclude the possibility that (some of) the relevant implicit mental states that link representations of people (e.g., black people) to representations of particular traits (e.g., physical attributes) may be propositionally rather than associatively structured (see last chapter). Whether a given mental state is associatively or propositionally structured does not make any difference to the argument presented here.

[90] Though see footnote 86 above.

pairing of white faces with negatively valenced words. Similarly, negative black stereotypes may have readily been activated in the white participants when they had to pair black faces with negatively valenced words. By contrast, the pairing of black faces with positive words may have been less effective in triggering the activation of positive black stereotypes. The Eval-IAT results may thus reflect that white people associate white people more strongly with positively valenced stereotypes than with negatively valenced stereotypes and black people more strongly with negatively valenced stereotypes than with positively valenced stereotypes. It is of course difficult to say which specific stereotypes may have been invoked on the test. However, a good guess would be that those stereotypes were triggered that were most accessible in participants' memory (Bargh & Pietromonaco, 1982). These considerations show that there is a possible alternative explanation of the Eval-IAT effect that hints at the mutual contribution of evaluation and stereotyping.

In sum, both the claim that the Stereo-IAT effect is exclusively driven by differences in stereotyping (and not in evaluation) as well as the claim that the Eval-IAT effect is exclusively driven by differences in evaluation (and not in stereotyping) can be called into question. Different evaluations of white and black people may well have been elicited on the Stereo-IAT depending on the association of these two groups with mental and physical attributes, respectively. Moreover, the pairings of positive and negative words with faces of white and black individuals on the Eval-IAT may well have triggered stereotypes that matched the valences of the words. The reason for the absence of correlation between Eval-IAT and Stereo-IAT is therefore not necessarily that the former test is exclusively influenced by evaluations (i.e., affect), whereas the latter test is purely influenced by stereotypes (i.e., conceptual content). A plausible alternative explanation is that the absence of correlation is a result of different, what I call following Madva & Brownstein (2016), *evaluative stereotypes* that the two tests evoke.[91] As already mentioned, the Stereo-IAT effect seems to be primarily driven by the negatively valenced stereotype that black people are less mentally capable than white people. Due to its different structure, the Eval-IAT is less likely to trigger a particular evaluative stereotype across participants. Rather the Eval-IAT can be assumed to evoke whatever stereotypes are strongest for each individual participant. Plausibly, participants whose performance on the Stero-IAT was driven by the negative

[91] On a more general note, my elaborations in this section point once again to the need to be cautious about the interpretation of psychological measurement results. I showed that a test that has been taken to reveal implicit stereotypes (the Stereo-IAT) may as well tap into implicit evaluations, while a test that has been taken to reveal implicit evaluations (the Eval-IAT) may as well tap into implicit stereotyping. Similar considerations may plausibly apply to other test like the affective priming task, which is commonly interpreted to reveal evaluations (Fazio et al., 1995) or the shooter task, which has been interpreted to reveal cultural stereotypes (Correll et al., 2002).

“black people are unintelligent” stereotype were not necessarily influenced by negative black stereotypes on the Eval-IAT and, conversely, those participants who were influenced by predominantly negative black stereotypes on the Eval-IAT were not necessarily influenced by the “black people are unintelligent” stereotype on the Stereo-IAT. The absence of correlation between Eval-IAT and Stereo-IAT is thus compatible with a view according to which stereotyping and evaluation are strongly intertwined (Madva & Brownstein, 2016: 10-11). It must be emphasised, however, that this alternative explanation leaves open the question of whether stereotypes and evaluations are distinct kinds of mental states that always or at least normally co-occur (which would arguably still be compatible with a two-type model), or whether evaluative stereotypes are, as proposed by Madva & Brownstein (2016), unified mental states that blend conceptual and affective content (which would speak for a one-type model). In section 3.4, I will tackle this question.

Before I do this, however, I will show that the presented account of evaluative stereotypes can also explain why Stero-IAT and Eval-IAT results predict different kinds of behaviours, and that the postulation of the construct of evaluative stereotypes has the potential to direct attitude research into more fruitful directions (section 3.3.2). Moreover, I will show that the evaluative stereotype account can explain why anxiety affects Eval-IAT results but not Stereo-IAT results, and that this account in fact provides the best explanation for selective effects of different emotions on different kinds of IATs (section 3.3.3).

3.3.2 Behavioural prediction from Eval-IAT and Stereo-IAT scores

As noted before, Amodio and Devine (2006) also suggest that the Stereo-IAT and the Eval-IAT are uniquely predictive of instrumental and consummatory behaviour, respectively. According to them, this finding points to differential functional roles of stereotypes and evaluations (i.e., social affect), supposedly confirming that stereotypes and evaluations are distinct psychological kinds. In the last section, I presented an alternative explanation for the absence of correlation between Stereo-IAT and Eval-IAT that does not depend on stereotypes and evaluations being independent. I have argued that different evaluative stereotypes (combinations of affective and conceptual mental content) may have been evoked when people participated in these tests. The behavioural prediction results can similarly be explained by the operation of different evaluative stereotypes. That is, they may be accounted for by differences in the content of particular evaluative stereotypes evoked on the Stereo-IAT and the Eval-IAT rather than by a difference between stereotyping and evaluation (Madva & Brownstein, 2016:

8-10).[92] I argued that the Stereo-IAT effect may be driven by the negatively valenced stereotype that black people are less mentally capable than white people, whereas the Eval-IAT effect may be influenced by a broader range of evaluative stereotypes (essentially, whatever stereotypes come readily to the mind of the participants). Against this backdrop, it is unsurprising that the Stereo-IAT was a good predictor of judgments about a black interaction partner's academic task performance (Amodio & Devine, 2006: study 3). By contrast, the Eval-IAT scores may have failed to predict these judgments because participants' Eval-IAT performance was presumably driven by a variety of evaluative stereotypes, only a few of which were about the purported mental inferiority of black people. It is also of no surprise that the negative conception of black people as mentally weak that may have driven the Stereo-IAT effect would not predict the seating distance to a black person. Increased social distance to stigmatised individuals can be due to anxiety (Dotsch & Wigboldus, 2008), yet there is no reason to assume that mentally weak individuals are perceived as a threat. Accordingly, it can be supposed that the Eval-IAT excelled in the prediction of seating distance because participant's performance on the Eval-IAT may have been a reflection of threat-related stereotyping rather than (or more than) intelligence-related stereotyping. There is thus a feasible alternative explanation for why Eval-IAT and Stereo-IAT results were uniquely predictive of different kinds of behaviours. These behaviours may not have been the function of different kinds of mental states (evaluations and stereotypes) as proposed by Amodio and Devine (2006) but the function of evaluative stereotypes with different contents.

This indicates that we may not actually need to separate the contributions of stereotypes and evaluations to make accurate predictions about people's responses towards others. In fact, Madva and Brownstein (2016) convincingly argue that discriminatory responses are best predicted by specific evaluative stereotypes (pp. 15-17). They review a range of studies that employed IATs that presumably tapped into specific evaluative stereotypes and show that results on these IATs excelled as predictors of discriminatory behaviours (Agerström & Rooth, 2011; Rooth, 2010; Rudman & Ashmore, 2007; Rudman & Kilianski, 2000; Rudman & Lee, 2002). For illustration, let us have a closer look at one of these studies. Rudman and Ashmore (2007) compared two different IATs for their ability to predict overtly discriminatory

[92] Madva and Brownstein (2016) point to the fact that we usually explain behavioural differences by differences in the content of mental states rather than by differences in the kinds of mental states involved (pp. 8-10). For example, two people may act differently because of differences in the content of their desires. Person A may belief that Bonnie is a drug dealer and desire to buy drugs, while person B may belief that Bonnie is a drug dealer and desire drug dealers to be punished (Madva & Brownstein, 2016: 9). The fact that person A buys drugs from Bonnie and that person B calls the police does not point to the fact that person A's and person B's behaviour is the function of different *kinds* of mental states but just to the fact that their desires differ in content.

intergroup behaviour. The first IAT included positively and negatively valenced words without stereotype content (positive: sunshine, smile, etc.; negative: filth, death, etc.) and thus closely resembled the Eval-IAT used by Amodio and Devine (2006).[93] The second IAT included positively and negatively valenced words with stereotype content (negative: lazy, shiftless, dangerous, etc.; positive: ambitious, industrious, ethical, etc.). This latter IAT can be described as *evaluative stereotype IAT* (henceforth, "Eval-Stereo-IAT").[94] In contrast to Amodio and Devine's (2006) Stereo-IAT, which was putatively designed to measure the influence of the conceptual content of stereotypes on behaviour, the Eval-Stereo-IAT was explicitly designed to assess the behavioural impact of *affectively valenced* stereotype content. In Rudman and Ashmore's (2007) first study, participants were asked to indicate how frequently they had performed certain discriminatory actions towards black people in their life after they had completed the two IATs. The behaviours were clustered into three groups. They included verbal behaviours, such as making ethnically offensive comments, defensive behaviours, such as avoiding certain groups, and offensive behaviours, such as physically hurting targets. In a second study, participants were asked to indicate to which student groups they would apply a necessary funding cut before taking part in the two IATs. Crucially, among the student groups listed were groups representing minority groups (Jews, Japanese, black people) that were subsequently also included in the IATs. For all these various behaviours in study 1 and 2, a statistical procedure (hierarchical regression analysis) revealed the Eval-Stereo-IAT to be the more effective predictor than the generic Eval-IAT. In both studies, Eval-Stereo-IAT scores predicted the reported and actual discriminatory behaviour even after controlling for the influence of explicit prejudice, whereas the Eval-IAT did not account for unique variance in the behavioural data. Strikingly, at least some of the behaviours examined in this study, namely the offensive and defensive behaviours, are behaviours that presumably fall under what Amodio and Devine (2006) describe as consummatory behaviours (involving approach and avoidance responses). On Amodio and Devine's account, consummatory behaviours are the function of affective responses and should thus be predictable by people's scores on the Eval-IAT. Yet in Rudman and Ashmore's study, the Eval-Stereo-IAT fared by far better in predicting these responses than the Eval-IAT.

Madva and Brownstein (2016) rightly stress that "[t]he Eval-IAT may be too coarse-grained to capture, let alone differentiate among, the many affect-laden responses most relevant to social behavior" (p. 14). It should come as no surprise that prejudice is not just a matter of generic negative affect (that the Eval-IAT is supposedly tracking)

[93] Rudman & Ashmore (2007) call this "the attitude IAT".
[94] One should not get distracted by the fact that Rudman and Ashmore (2007) refer to this latter IAT mostly as "stereotype IAT". They explicitly note that this IAT assesses both "cognitive and evaluative associations" (p. 361).

but often involves specific emotions, such as anger, disgust, fear, or pity (Cottrell & Neuberg, 2005; Inbar et al., 2009; Tapias et al., 2007). Note that I have mentioned above that differences in seating-distance to the African-American in Amodio and Devine's (2006) experiment may have resulted from differences in people's anxiety-related stereotyping. If this is the case, and we want to predict how far a white person will sit apart from a black person, we may be better advised to rely on a measure that is specifically designed to measure people's anxiety-related stereotyping in regard to black people rather than a measure like the Eval-IAT that may tap into whatever stereotypes about black people come readily to people's minds (or the Stereo-IAT that seeks to balance the involved valences of the stereotypes).[95] Note that different predictors will likely excel in different domains. While anxiety-related stereotyping is likely to be the best predictor of aversive or defensive behaviours (e.g., avoiding physical contact), anger-related stereotyping is likely to be a better predictor of offensive behaviours (e.g., physically harming others; Carver & Harmon-Jones, 2009; Mackie, Devos, & Smith, 2000; Madva & Brownstein, 2016: 21). This suggests that in order to yield optimal predictions of inter-group behaviour, researchers should adjust their measures in accordance with the task at hand. This has long been recognized as "the principle of compatibility" (Ajzen 1988: 96-98), but unfortunately this principle is often neglected in practice. [96]

Accordingly, measures like the IAT, if properly designed, might have some predictive value after all. Recall that recent meta-analyses of the IAT, and other indirect measures, revealed disappointingly low average correlations between people's scores on indirect measures and different forms of discriminatory behaviours (Oswald et al., 2013; Forscher et al., 2016; see section 1.3.3 in chapter 1). According to Madva and Brownstein (2016) these findings do not show that the IAT has low predictive validity per se but rather "that researchers too often use the *wrong* measures for a given task" (p. 16).[97] Most researchers rely on generic Eval-IATs or Stereo-IATs that seek to balance the affective valence of the stimuli. However, the research reviewed by Madva and Brownstein (2016) suggests that measures that tap into specific stereotypes with

[95] We may for example develop an Eval-Stereo-IAT that includes on the one hand stereotypic traits that are likely linked to anxiety (e.g., threat, assault, violence) and on the other hand words that are safety-related (e.g., safety, support, peace). Similarly, one could develop an affective priming task that requires participants to categorise stimuli as frightening or non-frightening, when they are primed by the previous presentation of black or white faces.

[96] Note that this principle is not only often neglected in indirect but also in direct attitude assessments. Note, for example, that feeling thermometers (a popular direct measure of attitudes) assess affect only very coarsely in terms of negative or positive valence. Accordingly, we should not expect people's responses on feeling thermometers to be predictive of their responses in specific situations, in which specific stereotypes and specific emotions are likely to become activated.

[97] See also Brownstein's and Madva's contributions (amongst others) to the roundtable discussion on the value of the IAT on The Brains Blog (accessed on 24/01/17): http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

specific affective implications show the greatest predictive success (Agerström & Rooth, 2011; Rooth, 2010; Rudman & Ashmore, 2007; Rudman & Kilianski, 2000; Rudman & Lee, 2000).

To conclude, there is abundant evidence that IATs that are designed to tap into both the conceptual content and specific affective implications of stereotypes (Eval-Stereo-IATs) have more predictive power than IATs that are designed to tap into generic evaluations (Eval-IATs) or the conceptual content of stereotypes (Stereo-IATs). To be sure, the results by Amodio and Devine (2006) suggest that Eval-IAT and Stereo-IAT results may, at least sometimes, lead to accurate predictions (studies 2 and 3). This may be due to the fact that these measures happen to tap into evaluative stereotypes after all (see section 3.3.1.2). Yet still, Eval-IATs and Stereo-IATs that merely incidentally tap into evaluative stereotypes will have less predictive success than Eval-Stereo-IATs that are specifically designed to tap into particular evaluative stereotypes relevant to the behavioural domain at hand. This shows that the emphasis that some psychologists have put on the purported independence of stereotypes about social groups and affect towards social groups may have led research in the wrong direction. Note that using the notion of evaluative stereotypes in our explanations for discriminatory behaviours may have pragmatic value in as far as it may nudge researchers to examine the interaction between affective and conceptual aspects of intergroup bias rather than examining the independent contributions of affective evaluations and cognitive stereotypes. This will ultimately lead us to a better understanding of those factors that actually drive people's responses towards social groups and to better predictions of intergroup behaviour.

Note that I have stressed in the introduction to this thesis that the notion of an attitude plays a crucial role in explanations and predictions of people's evaluative responses (see function *F1* mentioned in the introduction to this thesis). I argued that to optimally fulfil this explanatory and predictive function, our notion of an attitude towards a group X should pick out exactly those features of an individual's psychology that drive that person's evaluative responses towards that group X (see desideratum *D1*). As I have suggested in this section, stereotypes about social groups and affect towards social groups jointly drive people's responses towards social groups. Accordingly, we should want a model of attitudes that incorporates both affect and stereotypes regarding social groups (i.e., a model that incorporates evaluative stereotypes). Excluding any of these components would diminish the explanatory and predictive power of our model. Yet, it shall be emphasised again that it remains an open question whether stereotypes about social groups and social affect are distinct mental states that always (or at least normally) causally interact (which is arguably still compatible

with a two-type view) or whether they form unified mental states that blend conceptual and affective content (a question that I will tackle in section 3.4).

3.3.3 Effect of emotion on Eval-IAT and Stereo-IAT scores

So far I have presented alternative explanations for Amodio and Devine's (2006) finding that there is a low correlation between people's Stereo-IAT and Eval-IAT results (section 3.3.1) and their finding that Stereo-IAT and Eval-IAT results predict different kinds of behaviours (section 3.3.2). Now I turn to the final finding presented in section 3.2: the finding that social anxiety affects Eval-IAT scores but not Stereo-IAT scores (Amodio & Hamilton, 2012). According to Amodio and Hamilton (2012), this finding highlights that evaluations and stereotypes are distinct mental kinds that exhibit distinct functional profiles.[98] By now, it should come as no surprise that I insist that there is an alternative explanation for this finding that does not require evaluations and stereotypes to be functionally independent. Plausibly, the anxiety that was induced by the anticipation to interact with a black person may have affected the particular evaluative stereotypes that contributed to the Eval-IAT effect but not the particular evaluative stereotypes that contributed to the Stereo-IAT effect (Madva & Brownstein, 2016: 8). Amodio and Hamilton (2012) employed a very similar Eval-IAT and Stereo-IAT as Amodio and Devine (2006). We saw that the Stereo-IAT effect found in Amodio and Devine (2006) was plausibly driven by the negatively charged stereotype that black people are mentally inferior to white people. We may speculate that anxiety does not affect the extent to which this particular stereotype becomes activated but that it may affect the activation of other stereotypes, which played a role on the Eval-IAT (but not on the Stereo-IAT). For example, it seems plausible that anxiety would boost the activation of black-violence or black-dangerous stereotypes. Moreover, we can speculate that a different kind of emotion manipulation would in fact affect people's

[98] It should be noted that while Amodio and Devine (2006) have purportedly found a manipulation that affects evaluations (as assessed on the Eval-IAT) but not stereotyping (as assessed on the Stereo-IAT), they do not conversely report on a manipulation that affects stereotyping but not evaluations. In fact, Madva and Brownstein (2016) argue that a change in stereotyping without a corresponding change in evaluation would be "more diagnostic" for a two-type view than a change in evaluation without a corresponding change in stereotyping (p. 11). They support this statement with reference to a model by Gawronski and Bodenhausen (2006), which they take to be an example of a one-type view. They claim that on Gawronski and Bodenhausen's model occurrent affect towards a social group is determined by the activated stereotypes about that group, but not vice versa (i.e., occurrent affect does not influence stereotype activation). It is surprising that Madva and Brownstein base their argument here on a model according to which the relation between stereotyping and evaluation is one-directional because their own model seems to imply that the relation between stereotyping and evaluation is in fact bi-directional. They claim, for example, that "all putative implicit stereotypes are affect-laden and all putative implicit prejudices are 'semantic,'" (p. 1). On this view, we should neither expect that a manipulation that affects stereotyping fails to affect evaluations, nor should we expect that a manipulation that affects evaluations fails to affect stereotyping.

Stereo-IAT scores. For example, inducing feelings of superiority may boost the conception that black people are mentally inferior and thereby increase the observed bias on the Stereo-IAT.

Further experiments are certainly needed to assess whether these particular assumptions are right. However, there is already some evidence that specific emotions indeed selectively affect specific IATs (Dasgupta et al., 2009; Inbar et al., 2009; Inbar, Pizarro, & Bloom, 2012). Crucially, an evaluative stereotype model provides arguably the best explanation for these findings. Dasgupta and colleagues (2009), for example, showed in a series of experiments that induced disgust, but not induced anger, affects evaluations of homosexuals on an Eval-IAT, whereas induced anger, but not induced disgust, has an effect on the evaluation of Arabs on an Eval-IAT.[99] They conclude that "emotions may focus attention on semantically applicable features of outgroups" (p. 589) and that "negative emotions will only exacerbate implicit bias if they are applicable to the stereotypes and threats attached to the group" (ibid). Common Arab stereotypes such as "Arabs are terrorists" are arguably linked to the emotion of anger and so it comes as no surprise that anger increases bias against Arabs on Eval-IATs. Similarly, the increase in bias against homosexuals (as measured on an Eval-IAT) after disgust induction is exactly what we should expect given that common homosexual stereotypes such as "homosexuals engage in lewd conduct" are arguably linked to the emotion of disgust.

The finding that different emotions exhibit different effects on Arab and homosexual Eval-IATs is crucial because it provides us with an instance where the evaluative stereotype explanation is in fact the better and not just an equally plausible explanation. Note that the selective effects of disgust and anger on homosexual and Arab Eval-IATs are difficult to reconcile with the assumption that Eval-IATs measure generic likes or dislikes that are detachable from stereotypes. Disgust and anger are arguably both negative emotions and should increase biases on both Arab and homosexual Eval-IATs if these measured just basic positive and negative evaluations. By contrast, the selective effects of anger and disgust on the Eval-IATs is exactly what we should expect if these measures tapped into different stereotypes that are linked to anger and disgust, respectively. This further supports the claim, already touched on in the last section, that we need to go beyond the simple distinction between positive and negative affect if we want to understand the nature of people's attitudes. Different emotions, such as anger, disgust, fear, and pity, play a role in our responses to people

[99] See Tapias and colleagues (2007) for a related study (study 2). They found that a predisposition to feel anger (and not a predisposition to feel disgusted) predicted people's reports of their attitudes towards African-Americans, while a predisposition to feel disgust (and not a predisposition to feel anger) predicted people's reports of their attitudes towards homosexuals.

qua members of social groups, and these emotional responses are tightly linked to those stereotypes that we associate with these groups (Madva & Brownstein, 2016: 13-14).

3.3.4 Summary

To sum up, there are two competing explanations for the findings that (1) Eval-IAT and Stereo-IAT scores do not correlate well with each other, (2) that scores on these measures predict different kinds of behaviours, and (3) that scores on these measures are affected differently by anxiety. According to Amodio and colleagues (2006, 2012), these findings can be explained by the fact that the Eval-IAT and the Stereo-IAT tap into different mental kinds: the Eval-IAT measures affect towards black and white people and the Stereo-IAT measures stereotypes about black and white people. I presented an alternative explanation, inspired by Madva and Brownstein (2016), according to which both the Eval-IAT and the Stereo-IAT tap into interactions of conceptual stereotypes and affective evaluations. On this view, Eval-IAT and Stereo-IAT may tap into different evaluative stereotypes.

I also argued that there are reasons to prefer this alternative explanation. Firstly, I argued that using the notion of evaluative stereotypes in our explanations for biased social behaviour will likely have pragmatic benefits (see section 3.3.2). By using this notion, we nudge researchers to focus their investigation on the interactions between the conceptual and affective aspects of intergroup bias. Examining these interactions is arguably a more fruitful research programme than examining alleged differences between stereotyping and evaluation. One indicator of this is that measures that are explicitly designed to measure specific evaluative stereotypes (Eval-Stereo-IATs) possess more predictive validity than measures that are designed to measure exclusively evaluations (Eval-IATs) or stereotypes (Stereo-IATs). Secondly, an evaluative stereotype model provides *the better explanation* for the finding that different negative emotions (e.g., anger and disgust) affect Eval-IATs about different outgroups (e.g., Arabs and homosexuals) differently (see section 3.3.3). This effect is well explained by the fact that different Eval-IATs tap into different stereotypes that are linked to specific emotions but is difficult to reconcile with the idea that Eval-IATs tap into generic positive or negative affective responses that are detached from stereotypes. Overall, we are thus well advised to refer in our explanations of discriminatory conduct to the interactions between stereotypes and affect (i.e., evaluative stereotypes) rather than emphasising the individual contributions of either stereotypes or affect.

At this point, one may wonder how evaluative stereotypes relate to attitudes – the main theme of this thesis. If the notion of an attitude towards a group X is to denote those features of an individual's psychology that drive that person's evaluative responses towards people of group X (see desideratum *D1* in the introduction to this thesis), we should acknowledge that evaluative stereotypes are (at least partly) constitutive of attitudes.[100] After all, evaluative stereotypes seem to be an important factor that drives people's responses towards people qua members of social groups. However, this still leaves open the question whether stereotypes and evaluations are distinct mental states that always (or normally) interact (and thus are still, in a sense, distinct components of attitudes) or whether evaluative stereotypes are unified mental states that blend conceptual and affective content. I will address this question in the next section.

3.4 A model of evaluative stereotypes

In what follows, I will first consider the possibility that evaluative stereotypes are a unified mental state (section 3.4.1). Such a one-type view has been proposed by Madva and Brownstein (2016), who draw on Gendler's (2008a, 2008b) notion of "alief". They argue that evaluative stereotypes are "mutually co-activating semantic-affective-behavioral 'clusters' or 'bundles'" (p. 1). I will reply that their position is misguided because the components of these supposed clusters are better construed as distinct mental states rather than as parts of a unified mental state (Currie & Ichino, 2012; Dogget, 2012; Holroyd, 2016; Nagel, 2012). In the subsequent section (section 3.4.2), I will then argue that although evaluative stereotypes are composed of different mental states (including conceptual/stereotypic and affective mental states), the distinction between "cold" cognitive stereotypes and "hot" affective prejudice as suggested by proponents of the two-type view is misguided (e.g., Valian, 2005). Due to the tight causal connection between stereotypes about social groups and affect towards social groups, there is in fact a sense in which stereotypes are affective and in which social affect is conceptual or stereotypic.

3.4.1 Evaluative stereotypes as unified mental states?

Madva and Brownstein (2016) argue that evaluative stereotypes "are best conceived in terms of mutually co-activating semantic-affective-behavioral 'clusters' or 'bundles'" (p. 19). They claim that these bundles of semantic, affective, and behavioural content are

[100] I added "at least partly" in brackets because there may still be additional mental kinds that are constitutive of attitudes (e.g., endorsed beliefs). In the next chapter, I will elaborate more extensively on this.

sui generis mental states that cannot be broken apart into more primitive mental states. That is, they defend one-type model. Madva and Brownstein (2016) relate their model to Gendler's (2008a, 2008b, 2012) model of "alief" (see also Brownstein & Madva 2012a, 2012b). Gendler develops the notion of alief to account for behaviour that is neither fully intentional (i.e., based on beliefs and desires) nor fully reflexive (i.e., involving no or only minimal representational content that mediates between stimulus and behaviour; see in particular Gendler, 2012, on this contrast). Gendler defines a paradigmatic alief as an associative mental state that links representational (R), affective (A) and behavioural (B) content. She illustrates alief driven behaviour, amongst others, with the example of tourists walking on a glass walkway high above the floor of the Grand Canyon (Gendler, 2008a). Although the tourists typically *believe* that the platform is safe and evidently desire to step on the platform, they may experience feelings of anxiety or uneasiness and may only cautiously move forward. That is, they may *alieve* something different. According to Gendler (2008a), "[t]he alief has roughly the following content: 'Really high up, long long way down. Not a safe place to be! Get off!!'" (p. 635).[101] According to Gendler, the person in this example does not have separate representational, affective, and motoric mental states activated but is in a unified state of alief. We can say that the person aliefs R-A-B.[102] Gendler (2008b) herself describes implicit attitudes as a form of alief (pp. 574-576), and Madva and Brownstein (2016) build upon this idea to illustrate the nature of evaluative stereotypes (pp. 19-22). To give an example, a racist alief may consist of associations of BLACK PERSON with concepts such as DANGER and WEAPON (representational/conceptual stereotype content), which is associated with the feeling of fear (affect/evaluation) and the reading of a motor routine for flight (behaviour).[103] Similar to the person walking on the glass platform, who feels anxiety and moves only cautiously forward despite her belief that the platform is safe, the person with the racist alief (say Sarah) feels afraid of black people and is inclined to keep distance to them despite her anti-racist beliefs.

I would like to reply that it remains unclear why Gendler (2008a, 2008b) and Madva and Brownstein (2016) insist on the view that their supposed clusters of representational/conceptual, affective, and behavioural content constitute unified

[101] Other examples of alief driven responses that Gendler (2008a) mentions are, amongst others, being reluctant to drink from a glass of juice in which a sterilised dead cockroach has been stirred, being disgusted by eating fudge that has the form of dog faeces, or being less accurate in throwing darts at faces of beloved people than at faces of unknown people.

[102] Gendler (2008b) notes that "[t]hough this usage is approximate – and in that sense, misleading – it helps to emphasize the ways in which thinking in terms of alief differs from thinking in terms of the traditional cognitive and conative attitudes" (p. 559).

[103] Madva and Brownstein (2016) rightly note that Gendler's model of implicit attitudes can be seen as an adaption of the classical tripartite model of explicit attitudes (p. 19; Rosenland & Hovland, 1960).

mental states. Several commentators of Gendler's model have raised the worry that what she describes as instances of alief are in fact conglomerates of various interacting mental states (Currie & Ichino, 2012; Dogget, 2012; Holroyd, 2016, Nagel, 2012).[104] The challenge comes down to this: why should we prefer the alief account of unified representational, affective, and behavioural content over an account that holds that certain representational mental states (e.g., conceptual associations) closely interact with affective mental states (e.g., emotions) and motoric mental states?

Replying to her critics, Gendler (2012) claims that the representational, affective, and behavioural components of an alief are not "fully combinatoric", which according to her shows that they are not distinct mental states (p. 806). According to Gendler beliefs and desires are fully combinatoric. That means that, in principle, each belief can co-occur with any desire (and vice versa). For example, Aisha's belief that there is cake in the fridge may usually co-occur with her desire to eat cake but could, in principle, occur with any other desire (e.g., her desire to empty the fridge or her desire to eat ice cream). Gendler (2012) rightly notes that this is why it is misguided to speak of any belief-desire pair as a unified mental state.

Yet, it remains unclear why Gendler (2012) thinks that the components of so-called aliefs are not fully combinatoric. Let us consider that when Sarah encounters black people in a deprived neighbourhood, her weapon concept becomes activated, she is afraid, and inclined to run away. Contrary to what Gendler (2012) claims, these representational, affective, and behavioural components are arguably fully combinatoric. For example, the same fear response that co-occurs with an activation of the association between BLACK PERSON and WEAPON may under other circumstances co-occur with an activation of an association between BLACK PERSON and RAPE. Furthermore, fear conditioning may allow us to establish a link between literally any mental representation and Sarah's fear response (Davey, 1992; Rachman, 1991). Conversely, even when the association between BLACK PERSON and WEAPON is strongly linked to fear, conditioning procedures may allow us to establish a link between the association of BLACK PERSON and WEAPON to a different emotion, say anger or pity. Note that the fact that it may be very difficult to break apart the components of alleged aliefs does not imply that these components are not *in principle* fully combinatoric. The same is arguably true for many of our belief-desire pairs (Currie & Ichino, 2012: 790). Aisha's desire to eat cake may in fact always co-occur with her belief that cake is high in calories, but that does not imply that her desire to eat cake

[104] Madva and Brownstein (2016) mention this objection (p. 19) and admit that they "do not offer any knockdown argument" against it (endnote 35). What they offer is "a list of features of implicit mental states that ought to constrain theorizing about the nature of these states" (p. 19). Yet, each of the features on their list is compatible with both their "unified mental state model" and the "distinct but closely interacting mental state model" that I defend in the following.

could not *in principle* co-occur with any other belief (or no belief at all). As the representational, affective, and behavioural components of alleged aliefs can combine in multiple ways and are arguably not (relevantly) less combinatoric than paradigmatic examples of mental states such as beliefs and desires, we have reason to assume that these components of aliefs are in fact distinct mental states.[105] This undermines the very notion of alief. When an association of BLACK PERSON and WEAPON is reliably linked to fear and the readying of the motor routine for flight, this is better conceptualised as an instance of causally closely connected mental states rather than a single mental state of alief.

This is congruent with Holroyd's (2016) proposed "minimal model" of implicit bias, which "sees implicit biases as simply causally related, or co-activated representational contents, or affective and behavioural responses" (p. 175). Holroyd (2016) emphasises that a benefit of this model is that it alerts us to the heterogeneity of mental states (e.g., representational and affective mental states) that may play a role in implicit cognition (see also Holroyd & Sweetman, 2016). Being aware of this heterogeneity is important when one is exploring bias intervention strategies. Note that an intervention that successfully tackles one kind of mental state (e.g., a conceptual association) may not necessarily affect another kind of mental state (e.g., an affective disposition), although these stand in close causal relations.

I will leave it open just how many different kinds of mental states are involved in what Gendler calls alief. It may seem natural to assume that each of the supposed components of aliefs corresponds to one kind of mental state. This would leave us in fact with a three-type model according to which representational mental states (i.e., stereotypes), affective mental states, and motoric mental states closely interact in the production of implicitly biased responses. However, it should be noted that on many accounts of affect, it includes a motoric component (e.g., Frijda, 1986; Ekman, 2003). For example, fear may (partly) be constituted by a motor programme for flight. On such a view, evaluative stereotypes may be constituted by *two* different kinds of mental states: representational mental states (i.e., stereotypes) and affective mental states (which include motor programmes).[106] It would be beyond the scope of this chapter to

[105] Note also that it is unclear how much less combinatoric than paradigmatic mental states the components of so-called aliefs would need to be in order to justify the claim that these components constitute a unified mental state.

[106] If one would want to stick to the notion of an implicit attitude and to the idea that these can be identified with individual mental states, one would thus have to identify them with representational mental states (e.g., conceptual associations) and with affective mental states. Note that our answer to the question of whether so-called implicit attitudes are associative structures (see last chapter) may accordingly depend on how we flesh out the nature of affect. The crucial question would be whether affect can be understood, in any relevant sense, as an associative structure. In the implicit attitude literature, affect is often understood as a representation of a positive or negative valence that is associatively linked to other

argue for a particular account of affect, so I will remain non-committal on how exactly we need to divide those mental states that constitute evaluative stereotypes. The point I want to stress is just that evaluative stereotypes are not a unified mental state because at least some of its components (i.e., stereotypes and affect) can combine in multiple ways.

To conclude, what I refer to as "evaluative stereotype" can be understood as a cluster of causally tightly connected representational (i.e., stereotypic) and affective (and potentially motoric) mental states. Sarah, for example, may possess an evaluative stereotype with a content that we may broadly describe as "black people are dangerous". This evaluative stereotype may consist of various associations of the concept BLACK PERSON with concepts like VIOLENCE, WEAPON or RAPE (and/or corresponding propositional mental states); the emotion of fear, and various motor programmes (e.g., the motor programme for flight).

3.4.2 The affective quality of stereotypes and the conceptual quality of intergroup affect

The fact that evaluative stereotypes are composed of different kinds of mental states (including conceptual/stereotypic and affective mental states) should, however, not be taken to suggest that there is a divide, of the kind suggested by proponents of the two-type view, between "cold" cognitive stereotypes and "hot" social affect (Valian, 2005). In fact, my proposed account of evaluative stereotypes is compatible with the view that stereotypes are affective and that social affect is conceptual/stereotypic (Madva & Brownstein, 2016).

Let us first consider why one might think that stereotypes are non-affective. The stereotype "women are nurturing" may serve as an example. Valian (2005) emphasises that this stereotype (what she calls "a schema") can be recruited by different belief systems (p. 200).[107] She mentions that it could be recruited to rationalise a sexist belief system that dictates that parenting should be the primary role of women or contribute to an egalitarian belief system that advocates that more men should develop nurturing characteristics. Depending on the kind of belief system the stereotype is recruited by, it may be linked to negative affect (sexist belief system) or positive affect (egalitarian

representations (see section 1.2.1). Yet, it is unclear whether, and if so in what sense, complex emotions such as fear or disgust can be understood as associative structures.

[107] Valian (1999) explains that "schemas are similar to stereotypes but the term 'schema' is more inclusive and more neutral." (p. 1044). By this, she seems to refer to the fact that stereotypes are often characterised as morally problematic attributions of traits to social groups (see also footnote 70 above). She prefers the term schema because she believes that assertions such as "women are nurturing" are not inherently morally objectionable. See Blum (2004) for the opposing view that such schemas (or "stereotypes" as he calls them) are inherently morally objectionable.

belief system). Valian (2005) takes this to show that “cognitions [such as schemas] do not automatically carry a set of emotions and motivations with them” (p. 200).

I grant that in different people the stereotype “women are nurturing” is linked to different kinds of affective responses. However, I would like to object that this does not establish that this stereotype is affectively “cold” as suggested by Valian (2005). Note that in both the sexist and the feminist the stereotype *is* tightly linked to an affective response, even though the nature of the affective responses differs. It is by virtue of this link to affect that it is legitimate to say that the stereotype is affective. In the sexist, the stereotype “women are nurturing” is affective in the sense that it is disposed to elicit negative feelings such as contempt towards women (and in the sense that feelings of contempt may elicit the stereotype), and in the feminist, the stereotype “women are nurturing” is affective in the sense that it is disposed to elicit positive feelings such as admiration (and in the sense that feelings of admiration may elicit this stereotype). In fact, it can be assumed that the stereotype “women are nurturing” is linked in most, if not all, people who harbour this stereotype to one affective reaction or another. That is, this stereotype, like other stereotypes, has affective significance to the person who harbours the stereotype. Saying that the stereotype “women are nurturing” is affectively “cold” is thus misleading.[108]

Note also that, conversely, a given affective response may be linked to different (sets of) stereotypes in different people. In one person, feelings of contempt towards women may trigger the activation of stereotypes such as “women are nurturing”, while in another person the same feeling may trigger stereotypes such as “women are irrational”. In both persons, the affective response has conceptual implications by virtue of its causal connection to stereotypes. The same will be true of emotions such as anger, disgust, or pity in regard to particular social groups. These emotions elicit specific (sets of) stereotypes by virtue of which they can be said to have conceptual significance for the individual.

To conclude, there is a sense in which stereotypes are affective and in which social affect is conceptual. Stereotypes about social groups are affective in the sense that they are disposed to trigger affective responses towards social groups (and also in the sense that they can be activated by affective responses towards social groups). Conversely, affective responses towards social groups are conceptual (or we may say stereotypical) in the sense that they are disposed to activate particular stereotypes about social groups (and also in the sense that they can be evoked by particular stereotypes about social groups).

[108] This being said, Valian (2005) may well be right that the fact that the stereotype can be recruited by different belief systems establishes that the stereotype is not inherently sexist.

3.5 Conclusion

In the introduction to this chapter, I have presented two contrasting views on the relationship between stereotypes about social groups and affect towards people qua members of social groups (what I refer to as “affect towards social groups” or “social affect”). In short, while proponents of two-type models have emphasised that stereotypes about social groups and affect towards social groups are separable mental states (corresponding to “cold” cognition and “hot” affect), one-type theorists have stressed that stereotypes and social affect form inseparable clusters. Now it is time to evaluate how these views fare in the light of the conclusions that I have reached in this chapter, and to examine how this may inform our understanding of attitudes.

Recall that two-type theorists, like Amodio and colleagues (Amodio & Devine, 2006; Amodio & Hamilton, 2012), have emphasised that cognitive stereotypes about social groups and affect towards social groups are only weakly correlated, predict different kinds of behaviours, and are affected differently by emotion manipulations (see section 3.2). I have argued that the given evidence does not warrant these claims (section 3.3). Low correlations between results on Stereo-IATs that supposedly measure stereotypes and results on Eval-IATs that allegedly tap into affective responses can equally well be explained by the fact that these measures tap into different clusters of stereotypes and social affect (i.e., by postulating the existence of evaluative stereotypes; see section 3.3.1). Furthermore, the fact that the Eval-IAT and the Stereo-IAT tap into different evaluative stereotypes may also explain why results on these measures predict different kinds of behaviours (see section 3.3.2) and why these measures are differently affected by the induction of anxiety (see section 3.3.3). The evidence discussed by two-type theorists thus fails to establish that stereotypes and social affect can operate independently of each other. Quite to the contrary, much speaks in fact for the claim that stereotypes and social affect form clusters (i.e., evaluative stereotypes). The evidence suggests that we can yield better predictions of discriminatory behaviour when we focus on the interactions between stereotypes and affect rather than focusing exclusively on either stereotypes or affect (see section 3.3.2). Moreover, differential effects of different emotions on IATs involving different social groups can best be explained by the fact that particular stereotypes typically interact with particular affective mental states (see section 3.3.3).

However, this does not imply that evaluative stereotypes are unified mental states of the kind of Gendler’s (2008a, 2008b) aliefs or Madva and Brownstein’s (2016) semantic-affective-behavioural clusters. In particular, I argued against the idea that an evaluative stereotype can be construed as a sui generis mental state that joins stereotypic and affective content (see section 3.4.1). Against this view speaks that

stereotypes (e.g., the conceptual association between BLACK PEOPLE and WEAPON) and affective responses (e.g., fear towards black people) are, at least in principle, fully combinatoric. That is, a given stereotype can co-occur with different kinds of affective responses and a given affective response towards a social group can co-occur with different stereotypes. This speaks for the view that evaluative stereotypes are composed of different mental states (e.g., representational/stereotypic and affective mental states) that have tight causal connections to each other (see also Holroyd, 2016).

Given these tight causal links between stereotyping and social affect, Madva and Brownstein (2016) are arguably right about the claim that "all putative implicit stereotypes are affect-laden and all putative implicit prejudices are 'semantic,'" (p. 1). On my account, affect towards social groups (what Madva and Brownstein call "prejudices") are conceptual (what Madva and Brownstein call "semantic") in the sense that they are disposed to activate stereotypes about social groups (and also in the sense that they can be triggered by particular stereotypes about social groups). Conversely, stereotypes about social groups are affective in the sense that they are disposed to trigger affective responses towards social groups (and also in the sense that they can be activated by affective responses towards social groups; see section 3.4.2).

To conclude, I have argued for a nuanced view that does not readily fall into the one-type/two-type classification as it has been outlined at the start of this chapter. On my view, one-type theorists are right in so far as stereotypes about social groups and affects towards social groups form tight clusters (what Madva & Brownstein, 2016, call "evaluative stereotypes"). However, these clusters are not a unified mental state, as Madva and Brownstein (2016) assume, but are composed of different kinds of mental states (e.g., conceptual mental states and affective mental states) that are causally interconnected. Due to the tight causal links between these mental states, it is appropriate to say that stereotypes have an affective quality and that affect towards social groups has a conceptual or stereotypic quality. Although I showed that it is misguided to draw a line between "cold" cognitive stereotypes and "hot" affective prejudice, my view allows to identify mental states that are primarily conceptual (Sarah's association between BLACK PERSON and VIOLENCE) and only by virtue of their causal connection to other mental states affective, and mental states that are primarily affective (Sarah's fear response to black people) and only by virtue of their causal connection to other mental states conceptual. This may appeal to some authors

who seem to sympathise with a two-type view without (fully) endorsing the claim that stereotypes are non-affective (e.g., Anderson, 2010).[109]

What does the foregoing imply for the notion of an attitude? In the introduction to this thesis, I argued that a key function of the attitude notion is the explanation and prediction of people's evaluative responses in regard to other people (see function *F1* in the introduction to this thesis). Furthermore, I argued that in order to fulfil this explanatory and predictive function, our notion of a person's attitude towards a group X should pick out exactly those features of that person's psychology that drive that person's evaluative responses towards that group X (see desideratum *D1* in the introduction to this thesis). In this chapter, I argued that stereotypes about social groups and affective mental states interact tightly in the production of evaluative responses towards social groups. In fact, models that highlight the interactions between stereotypes and social affect produce better predictions than models that focus on one of these components alone (see section 3.3.2). If we would identify attitudes merely with affective mental states or merely with stereotypes, we would miss a crucial aspect of what is predictive of people's inter-group responses. Accordingly, we should acknowledge that both stereotypes (which may have associative or propositional structure) and social affect are part of people's attitudes. This raises the question as to what kind of ontological status attitudes have if they are jointly constituted by mental states of different kinds (see question *Q3* in the introduction to this thesis). This leads us straight to the issues to be addressed in the next chapter, in which I will evaluate and defend one account of the ontological status of attitudes: the view that attitudes are traits, which are based on various distinct mental states.

[109] Anderson (2010) claims that stereotypes "are *more* a matter of 'cold' cognitive processing than 'hot' emotion" (p. 45, my emphasis). She thus falls short of endorsing the claim that stereotypes are (always) non-affective.

Chapter 4: A trait view of attitudes

4.1 Introduction

In the last chapter, I argued that conceptual/stereotypic and affective mental states interact so tightly in the production of evaluative responses towards other people that it would not make sense to identify attitudes with either conceptual/stereotypic or affective mental states alone. In order for the notion of an attitude to optimally fulfil an explanatory/predictive function, we need to acknowledge that attitudes have both conceptual and affective components (see desideratum *D1* of a model of attitudes mentioned in the introduction to this thesis). However, it shall be noted that conceptual and affective mental states as described in the last chapter can hardly be the only components of attitudes. Note that Sarah, for example, does not only harbour evaluative stereotypes, such as her "black people are dangerous" evaluative stereotype (her associations of the concept BLACK PERSON with concepts such as VIOLENCE, WEAPON, or RAPE, the emotion of fear, etc.), but also certain moral beliefs (e.g., her belief that it is morally reprehensible to treat people differently because of their skin colour) and desires (e.g., the desire not to discriminate against black people). These beliefs and desires also influence (at least sometimes) the nature of her evaluative responses towards black people (see also Besser-Jones, 2008).[110]

Accordingly, we need a model of attitudes that takes all of the above mentioned mental states into account if we want to optimally explain and predict Sarah's evaluative responses towards black people in the various situations in which she encounters them (see desideratum *D1*). Note that dependent on the situation that Sarah finds herself in, different mental states may become activated and drive her responses towards black people. When Sarah walks through a deprived neighbourhood, she may be stressed and her reactions towards black people may primarily be driven by her "black people are dangerous" evaluative stereotype. However, when she walks through her own affluent neighbourhood, she may feel at ease and have more cognitive resources available to reflect on her egalitarian commitments and to keep the activation of negative evaluative stereotypes at bay when she encounters black people. Our model of attitudes should account for this situation-specificity of her evaluative responses. Moreover, we should want a model that is consistent with the character evaluative function of attitudes (see desideratum

[110] Besser-Jones (2008) argues that a person's moral character consists not only of that person's behavioural dispositions but also of that person's moral commitments (beliefs, desires, and intentions) and the extent to which that person's behavioural dispositions are influenced by these moral commitments.

D2) and that could potentially appeal to all parties that use the attitude concept (philosophers, psychologists, and ordinary people; see desideratum *D3*).

In this chapter, I develop a model that I take to be consistent with these desiderata. In short, I will argue that attitudes are complex traits. Each of these traits is grounded in a variety of mental states (conceptual associations, affects, beliefs, desires, etc.) and can be analysed as a profile of situation-specific evaluative response dispositions. On this view, Sarah can be said to exhibit the profile of an aversive racist (or to possess the trait of aversive racism): she is disposed to show favourable responses towards black people in situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments; and she is disposed to show negative responses towards black people in situations in which she does not have sufficient time (e.g., when she has to judge quickly whether she is in danger) or cognitive resources (e.g., when she is occupied with the detection of potential threats or when she is deeply engaged in a conversation with her patients) to reflect on and be guided by her endorsed egalitarian commitments.

I will develop my account of attitudes in response to Machery's (2016) trait view of attitudes, which I find appealing but which is not without its own flaws. The view that attitudes are traits is attractive because we ascribe attitudes to people for much the same reasons that we attribute traits such as courage to people: we want to explain/predict people's responses and convey information about people's characters (see functions *F1* and *F2* of the attitude concept mentioned in the introduction to this thesis). Machery (2016) argues that attitudes, just like traits such as courage, are broad-track dispositions, each of which is grounded in a variety of mental states (conceptual associations, emotions, moral beliefs, etc.). Machery further implies that attitudes can be characterised in terms of an aggregate strength and positive or negative valence. If, for example, the clear majority of a person's cognition, affect, and behaviour towards black people is reflective of a negative evaluation of black people, that person can be said to possess a strong negative attitude towards black people. If, however, only a small majority of a person's cognition, affect, and behaviour towards black people is reflective of a negative evaluation, while the rest of that person's responses reflect positivity towards black people, we can according to this view say that that the person has a *weak* negative attitude towards black people. I will argue that characterising attitudes in terms of aggregate strength and generic positive or negative valence obscures relevant evaluative complexities of attitudes: it conceals those evaluative conflicts that many people, such as Sarah, are experiencing and neglects that specific emotions, and not just generic positive or negative affect, are characteristic of attitudes. My own preferred model of attitudes – the view that attitudes

are profiles of situation-specific evaluative response dispositions – does justice to these complexities, while still allowing us to conceptualise attitudes as traits.

This chapter is structured as follows. In section 4.2, I will give a detailed account of Machery's (2016) trait view of attitudes. In section 4.3, I will discuss Machery's argument to the best explanation in favour of his trait view. I grant that Machery's model provides an explanation for a range of perplexing findings from the attitude literature, but I also point out that that there are other ways to conceptualise attitudes (including the view that attitudes are mental states) that have the same explanatory power. In section 4.4, I will highlight that the failure of Machery's argument to the best explanation notwithstanding, there are good reasons to adopt a trait model of attitudes. In particular, there are striking similarities in the explanatory, predictive, and character evaluative roles of trait and attitude ascriptions, and the trait view of attitudes aligns well with the folk psychological understanding of attitudes. In section 4.5, I will present an objection that any view that holds that attitudes are traits (including my own) must address. According to this objection, people possess no traits because their responses are largely determined by aspects of situations that they encounter and not by inner response dispositions of the kind that traits are usually identified with. I will give substance to this claim by presenting Doris' (2002) influential situationist argument against the existence of character traits (section 4.5.1) and will show that this argument, in slightly modified form, can also be applied to attitudes construed as traits (section 4.5.2). In section 4.5.3, I will then present a reply that is open to Machery (2016). According to this reply, the situationist argument does not establish that there are no attitudes construed as traits but rather that people's attitudes are oftentimes relatively weak. This reply rests on the idea that attitudes are characterisable in terms of an aggregate strength and valence. In section 4.6, I will argue that the characterisation of attitudes in these "aggregationist" terms misses the point because it obscures the complex structure of attitudes. In particular, it does not do justice to relevant differences in the affective content of attitudes and masks those evaluative conflicts that people are often experiencing in regard to social groups, such as when people feel alienated by their own racist dispositions or exhibit both benevolent and hostile sexist tendencies. In section 4.7, I will describe my own trait model of attitudes, which I argue neutralises the situationist challenge (described in section 4.5) without obscuring the evaluative complexities of attitudes (described in section 4.6). This is the view that attitudes are profiles of situation-specific evaluative response dispositions. The framework for my account is provided by Mischel and Shoda's (1995) influential cognitive-affective personality system model (section 4.7.1). On this model, some traits at least can be analysed as "distinctive and stable patterns of behavior variability across situations" (p. 246). Analogously, I propose that we can understand attitudes,

construed as traits, as stable patterns of evaluative response variation across situations (section 4.7.2). I will highlight that there are different legitimate ways to individuate attitudes on this account, which depend on the interests and purposes of attitude ascribers (section 4.7.3). Moreover, I will stress that attitudes as construed on my model have a psychological basis that is typically composed of a variety of implicit and explicit mental states (section 4.7.4) and point out what this implies for attitude measurement (section 4.7.5). Lastly, to further locate my account of attitudes in the literature, I will compare my account of attitudes to Schwitzgebel's (2013) dispositional model of attitudes (section 4.7.6).

4.2 Machery's trait view

Machery (2016) develops his trait view of attitudes as an alternative to what I described as the standard view of attitudes: the view that attitudes are mental states, which are either implicit or explicit.[111] He refers to this dominant view of attitudes in psychology and philosophy as "the Freudian picture of attitudes" (p. 105). Machery explains that on the Freudian view implicit attitudes are characterised as non-introspectable and automatic mental states, while explicit attitudes are described as mental states that are introspectable and whose impact on cognition and behaviour can intentionally be controlled.[112] He argues that understanding attitudes as traits, which are neither properly described as implicit nor explicit and which are based on a variety of different mental states, provides the better explanation for a range of findings from the attitude literature than the Freudian view (pp. 115-120). These findings include the low correlation between people's results on different indirect measures of attitudes (see section 1.3.2 in this thesis), the susceptibility of indirect attitude measures to contextual influences (see, for example, section 1.2.2, section 2.2.2.3, and section 3.3.3) and the low predictive power of people's results on indirect measures (see section 1.3.3).

[111] It shall be noted that Machery is certainly not the first author to link attitudes to traits. Ajzen (1988), for example, has pointed out that both the trait notion and the attitude notion are used in "dispositional explanations of behaviour" (p. 1). According to him, personality psychologists use the trait concept in dispositional explanations, while social psychologists use the attitude concept in dispositional explanations.

[112] Machery's characterisation of the "Freudian view" includes two of those criteria that are according to my review in chapter 1 often relied on to justify the distinction between implicit and explicit attitudes. These are the criteria of awareness (introspectability) and intentional control. In contrast to my characterisation of the standard view in chapter 1, Machery does not mention mental structure (i.e., associative vs. propositional mental structure) and rational control as further criteria. Note also that I showed in chapter 1 that even proponents of the standard view increasingly acknowledge that awareness does not provide a criterion by reference to which a distinction between implicit and explicit attitudes can be drawn. These differences between Machery's and my characterisation of the standard view (or the Freudian view) do not have any bearing on the arguments to come in this chapter.

Yet, before we can evaluate Machery's (2016) argument to the best explanation (see next section), we need to examine more closely what his claim that attitudes are traits comes down to. He characterises traits as follows:

> A trait is a disposition to perceive, attend, cognize, and behave in a particular way in a range of social and non-social situations. Within a species, there are individual differences with respect to a particular trait; some organisms have more of it, others less. This variation can be measured, and it is predictive of their behavior and cognition. (Machery, 2016: 111)

Machery uses, amongst others, the example of courage to give some context to this characterisation of a trait. Knowing that a person is courageous helps to predict the person's behaviour and cognition in a wide range of situations. Note that we have certain expectations of how a courageous person would behave in the face of a fire alarm, fierce criticism by a superior, a dangerous animal, etc. We can also fairly well predict some general cognitive and affective tendencies of a courageous person. For example, we would expect that the courageous person is not unduly swayed by fear and able to countenance due risk. Courage is thus at the same time a behavioural, affective, and cognitive disposition. Because traits such as courage manifest themselves in these different ways (and not just in a single way), Machery characterises them as broad-track dispositions rather than as narrow-track disposition (p. 111).[113]

He emphasises that courage, like any other trait, is not a mental state, and crucially, that it is not reducible to any of the occurrent mental states that it may manifest in. Rather courage is based on a range of mental states and processes, which compose the "psychological basis" of the trait and which jointly determine the degree to which the trait is possessed (i.e., determine the strength of the trait):

> A person's degree of courage depends on her moral beliefs (e.g. whether fear is shameful), on the nature of her fear reactions, on the strength of her pride, on her capacity for self-control, and so on. (Machery, 2016: 112)

By this he does not mean to imply that there is only one particular set of mental states that is sufficient for possessing a trait to certain extent. Rather different compositions of mental states can realise a specific trait to a particular degree.

Building upon this characterisation of a trait, he goes on to define attitudes as follows:

> [Attitudes] are broad-track dispositions to behave and cognize (have thoughts, attend, emote, and so on) toward an object (its formal object) in a way that reflects some preference. To have a positive attitude toward liberals is to be disposed to interact with

[113] Machery uses the terms "multitrack disposition" and "broad-track disposition" interchangeably.

> liberals in a way that reflects a positive evaluation and to have positive thoughts and emotions about them. (Machery, 2016: 112)

Accordingly, we can say that a person who has a negative attitude towards black people is disposed to behave, think, and feel in a way that reflects a negative evaluation of black people. This implies that knowing that a person has a negative attitude towards black people provides us with a good basis for predicting her behaviour as well as her thoughts and feelings towards black people. Just as any other trait of a person, attitudes according to Machery have a psychological basis, which is composed of a range of different mental states and processes. He provides the following example for this:

> A negative racial attitude toward blacks may depend on moral beliefs (e.g. for most of us the belief that racism is wrong or, for some racists, the belief that racism is right), on non-propositional associations between concepts (e.g. an association between the concept of a black man and the concept of danger), on emotions (e.g. fear when confronted with black men), and on weak self-control. This psychological basis is as heterogeneous as the psychological basis of courage. Some of the components may be conscious (perhaps some moral beliefs), while others (including associations between concepts) are likely to be inaccessible to introspection. (Machery, 2016: 112)

Machery argues that while the mental states and processes that form the psychological basis of the attitude (moral beliefs, conceptual associations, emotions, self-control processes) may correctly be described as implicit or explicit, this distinction does not apply to attitudes themselves. On this view, it makes no sense to ask whether a disposition to behave and cognise is introspectable or whether it operates in a controlled manner. Instead, this question can only reasonably be asked for the components of the psychological basis of the attitude (such as beliefs, associations, and emotions). Consequently, there are no such things as implicit or explicit attitudes.[114]

It is striking that Machery claims in the above quote that a negative attitude towards black people may partly be the function of the belief that racism is wrong. This may seem surprising because such a belief does not seem to imply a negative evaluation of black people. Quite to the contrary, such a belief can be expected to counteract negative cognitive and behavioural dispositions towards black people. What Machery seems to imply is that a person may possess a negative racial attitude *to a certain degree* if most components of the attitude imply a negative evaluation. Recall that Machery emphasises that traits such as courage can be possessed to different degrees and that the degree to which a trait is possessed ("the strength of a trait"; p.

[114] Accordingly, speech acts like "I like black people" do not express an explicit attitude according to Machery (2016: 114). He argues that they may express one's assessment of one's attitude, a command directed to oneself, a conscious emotional reaction, or a commitment to a moral norm.

112) is determined by the mental states and processes that form the psychological basis of the trait. Analogously, he assumes that the strength of an attitude (i.e., the degree to which an attitude is possessed) is the function of the mental states and processes that compose the psychological basis of the attitude. A person who believes that racism is wrong but who also harbours numerous mental states which imply a negative evaluation of black people (e.g., negative stereotypes and negative emotions) may mostly, albeit perhaps not always, exhibit negative cognitive, affective, and behavioural responses in regard to black people. Accordingly, the person can be said to possess a *relatively strong* negative attitude towards black people. As I will show in section 4.5.3, the idea that attitudes can vary in strength is a crucial component of Machery's attitude model because it helps warding off what I call the "situationist challenge" to the idea that attitudes are traits. Before I turn to this challenge, however, I will consider why we should adopt the view that attitudes are traits in the first place.

4.3 Assessing Machery's argument to the best explanation

Machery (2016) supports the trait view of attitudes with an argument to the best explanation (pp. 115-120). He argues that the trait picture provides a superior and unifying explanation for a range of perplexing results that have emerged in attitude psychology. These include: (1) the finding that scores on different indirect measures often do not correlate well with each other,[115] (2) the finding that scores on indirect measures are affected by various context effects, and (3) the finding that scores on indirect measures are poor predictors of evaluative responses. In what follows, I will discuss these findings in turn and show that any account of attitudes that is compatible with the idea that different indirect measures tap into different (sets of) mental states can explain them (see sections 4.3.1, 4.3.2, and 4.3.3). That is, Machery's trait model is certainly not the only model of attitudes that can account for (and would predict) these findings. This being said, I will stress that there are other reasons to prefer a trait view of attitudes over alternative attitude models in section 4.4: there is a striking similarity in the explanatory, predictive, and character evaluative roles of trait and attitude ascriptions, and a trait view of attitudes corresponds nicely with the folk psychological understanding of attitudes.

[115] I do not discuss separately a further finding that Machery (2016) mentions – "that the size of the correlation between two indirect measures can be manipulated" (p.117) – because it is closely related to this finding. See footnote 117 below.

4.3.1 Dissociation between results on different indirect measures of attitudes

Let us start with the finding that scores on different indirect measures of attitudes often show a relatively weak correlation (see section 1.3.2 in chapter 1 for a discussion of this finding). For example, people who show a bias against black people on an Eval-IAT do not necessarily show a bias against black people on an affective priming task (and vice versa; Fazio & Olson, 2003). Machery (2016) argues that this finding can be explained by the fact that different indirect measures tap into different components that constitute the psychological basis of an attitude.[116] To illustrate, recall that I argued in the last chapter that different kinds of IATs may plausibly tap into different evaluative stereotypes (i.e., clusters of conceptual content and affect) and that this may explain why results on these different measures are dissociated. If for example one IAT (primarily) taps into the "black people are dangerous" evaluative stereotype, while another IAT (primarily) taps into the "black people are musical" evaluative stereotype, we should expect a low correlation between people's scores on these measures. The claim that attitudes are traits that are based on various different mental states (including various evaluative stereotypes) that are accessed by different indirect measures accounts for the low correlations between these measures.[117]

However, note that all that is really needed to explain the comparatively low correlation between people's results on different indirect measures is the assumption that these measures tap into different (sets of) mental states. Machery's trait view implies this assumption, but other ways to individuate attitudes are certainly consistent with this assumption too. Note, for example, that we could identify attitudes with individual mental states (or specific clusters of mental states such as evaluative stereotypes) and say that different indirect measures tap into different attitudes. This would provide an equally good explanation for the finding that results on different indirect measures are often dissociated as the claim that these measures tap into different components of the psychological basis of a trait.

Machery's main point is that the trait model allegedly provides a better explanation for this finding than what he calls the Freudian view. He suggests that according to the Freudian view an agent harbours only one implicit attitude (construed as a mental state) in regard to a target, which is accessed by different indirect measures.[118]

---

[116] See also Huebner (2016: 67), who draws on Machery's argument.

[117] Note also that the more two indirect measurement procedures resemble each other, the higher the correlations between results on these measures can be expected to be. This helps to explain a related finding that Machery (2016) discusses: "that the size of the correlation between two indirect measures can be manipulated." (p.117)

[118] Machery is most explicit about how he construes the Freudian view in an endnote, in which he clarifies that the target of his paper is the view "that there is a single mental state that is people's implicit attitude." (endnote 5)

Accordingly, we should expect a substantial correlation between people's results on these measures (even if we allow for some measurement error). I would like to object that what Machery describes here is a caricature of the model of attitudes that is predominant in the attitude literature. In chapter 1 (section 1.2.2), I have argued that according to the most charitable interpretation of the standard view of attitudes, people can possess multiple implicit attitudes (identified with associative mental states) and multiple explicit attitudes (identified with propositional mental states) in regard to a social group. Arguably, such a view has sufficient resources to explain the fact that different indirect measures do not correlate well with each other. One could simply claim that different indirect measures tap into different implicit attitudes (i.e., different associative mental states).[119] Yet, note that even though the standard view is compatible with the low correlations between indirect measures, there are different reasons to think that the standard view does not provide us with an ideal model of attitudes (see previous chapters).

4.3.2 Contextual influences on indirect measures of attitudes

Machery (2016) further claims that the trait view of attitudes is better able to explain and predict contextual influences on indirect attitude measurement outcomes than the Freudian view. Let us first review some relevant findings. Some context effects are already familiar from last chapter. Recall that Amodio and Hamilton (2012) found that when their participants were made believe that they were soon to interact with a black person, they showed a stronger bias against black people on an Eval-IAT than when expecting to interact with a white person. That is, participants' Eval-IAT score was sensitive to the immediate situation in which they found themselves. Recall also that Dasgupta and colleagues (2009) showed that induced disgust increased participants' bias against homosexuals on an IAT, while induced anger increased their bias against Arabs on an IAT. This is again a case in which contextual factors (factors that induced disgust and anger) had an impact on indirect attitude measurement results. Machery (2016) argues that the trait view is well equipped to predict such contextual influences:

> [T]he trait picture hypothesizes that attitudes depend on psychological bases that encompass good old-fashioned mental states and processes, such as emotions, self-control, and so on, and that indirect measures tap into some of these components. Because we know a lot about these mental states – there is after all a lot of research on

[119] Machery anticpates this reply and objects that "[i]t is […] bad scientific practice to postulate a theoretical entity for every measure" (p. 117). Yet, it seems that Machery himself postulates a theoretical entity for every measure when he claims that these measures tap into different mental states. Note also that the assumption that people may possess multiple implicit attitudes can possibly be supported by independent considerations (e.g., the fact that people likely harbour multiple associative mental states that imply an evaluation of a social group; see section 1.2.2 in chapter 1).

> emotions and on what influences them – the trait picture leads us to predict which factors should modulate the measurement of attitudes by means of indirect measures. (p. 119)

Machery argues in this passage that the trait perspective helps us predict how particular contextual factors will affect particular measurement outcomes because it highlights that these measures tap into mental states of which we already know a great deal. If we know for example that a measure is particularly sensitive to negative emotions, we can predict that the prospect of interacting with a black person, which likely heightens feelings of anxiety in white people prone to negative evaluations of black people, will influence the measurement outcome.

Machery claims that the Freudian view does not allow for such specific predictions because on this view different indirect measures are assumed to tap into the same mental state (i.e., the same implicit attitude). According to him, this view has limited resources to explain why one indirect measure is influenced by a particular contextual factor, while another indirect measure is not affected by the same factor. Above, I have argued that, on a more charitable reading, the standard view of attitudes grants that people may harbour multiple implicit attitudes. Yet still, one may argue that Machery's trait view has more resources to explain and predict context effects because it holds that different indirect measures may tap into different kinds of mental states (e.g., emotions, associations, beliefs, etc.), whereas the standard view only holds that different indirect measures may tap into different mental states of the same kind (i.e., different associations).

However, this does still not establish that *only* Machery's trait view provides the resources to explain and predict these effects. In fact, all we need to explain and predict the differential context effects on different indirect measures of attitudes is the acknowledgement that these measures tap into different kinds of mental states (e.g., conceptual associations, propositional mental states, emotions) or different combinations thereof. This assumption is compatible with Machery's trait view but also with views that identify attitudes with individual mental states or specific clusters of mental states.

4.3.3 Low predictive validity of indirect measures of attitudes

Lastly, Machery (2016) argues that the trait view of attitudes provides the best explanation for the relatively low predictive validity of indirect measures of attitudes that has been revealed in recent meta-analyses (Greenwald et al., 2009; Oswald et al., 2013; see also section 1.3.3 of this thesis).[120] In chapter 3 (section 3.3.2), I have

[120] Note that a more recent meta-analysis, the one by Forscher and colleagues (2016), confirms the finding of low predictive validity.

argued that results on an indirect measure can predict clearly circumscribed responses if the measure has been customised for the specific task at hand. Machery would possibly agree with this assessment because he seems to base his argument merely on the claim that indirect measures are not reasonably good predictors of spontaneous evaluative responses across the board. He argues that on the trait view predictions that are based on people's scores on individual indirect measures are expected to be poor because any indirect measure only taps into a subset of the components that an attitude is composed of. As there are many components that influence people's responses towards a particular social group, a measure that only measures one of these components cannot predict responses towards the social group across the board (rather than just responses in circumscribed circumstances). On the Freudian view, by contrast, we should expect a greater predictive success of individual indirect measurement outcomes because, as Machery presents this view, it holds that different indirect measures tap into a single mental state (i.e., the implicit attitude), which influences a broad range of responses.

My reply should be familiar by now. What Machery presents here as the Freudian view is a non-charitable interpretation of the view that is predominant in the attitude literature. On a more charitable reading, the standard view of attitudes allows that people may harbour multiple implicit attitudes (i.e., multiple associative mental states in regard to a social group). So even the standard view allows for multiple determinants of people's responses towards a social group and could thus explain the low predictive validity of individual indirect measures.

4.3.4 Preliminary conclusion

To conclude, Machery (2016) fails to establish the superiority of the trait view with his argument to the best explanation. He argues that his trait view of attitudes provides the best explanation for (1) the fact that scores on different indirect measures often do not correlate well with each other, (2) the fact that scores on indirect measures are affected by various contextual factors, and (3) the fact that scores on indirect measures are poor predictors of evaluative responses. Yet, all we need to postulate to explain these findings is that different attitude measures tap into different (sets of) mental states. This is certainly compatible with the view that attitudes are traits that are based on a range of different mental states (e.g., associative mental states, propositional mental states, affective mental states, etc.) but also with models according to which attitudes can be identified with individual mental states (e.g., associative mental states, propositional mental states, affective mental states, etc.) or particular clusters of these.

4.4 Why conceptualise attitudes as traits?

The failure of Machery's (2016) argument to the best explanation notwithstanding, there are good reasons for adopting a trait view of attitudes. In the introduction to this thesis, I mentioned that the notion of an attitude plays a role in explanations/predictions of people's cognitive, affective, and behavioural responses towards other people (see function *F1* of a model of attitudes as mentioned in the introduction to this thesis) and in the assessment of a person's moral character (see function *F2*). It is widely acknowledged that trait ascriptions fulfil exactly these roles (Goldie, 2004: 3-6). If we are told that a person, say Frank, is arrogant, we may predict that he is likely to discount other people's opinions, is likely to show off, etc. We may also explain some of Frank's responses retrospectively by reference to the trait of arrogance. We may for example come to the conclusion that Frank boasted about his high salary to his colleagues because he is arrogant. Traits are usually understood to be the basic building blocks of a person's character. Knowing that Frank is arrogant does not only help us to explain and predict his responses but also to assess his character. As arrogance is commonly perceived to be a negative trait, we will likely come to the conclusion that Frank has a bad character (if all we know about him is that he is arrogant). Attitude ascriptions are strikingly similar in all these regards. They, too, help us to predict people's responses. If we are told that Frank has a negative attitude towards black people, we may predict that he will feel uncomfortable in the presence of black people, that he will likely discount the opinions of black co-workers, etc. Moreover, we can explain Frank's responses towards black people retrospectively by reference to his negative attitude towards black people. We may conclude that Frank keeps interrupting black co-workers in group discussions, while hearing out his white co-workers, because he has a negative attitude towards black people. Finally, the fact that Frank has a negative attitude towards black people may lead us to the conclusion that he has a bad character (given that this is all we know about Frank). Due to these striking similarities in the explanatory, predictive, and character evaluative roles of trait and attitude ascriptions, it is intuitive to assume that attitudes have in fact the ontological status of traits.

Note that the assumption that attitudes are traits evidently drives our day-to-day folk psychological judgments about people's attitudes. If someone tells us that Frank has a negative attitude towards black people, we intuitively infer that Frank is generally disposed to respond (cognitively, affectively, and behaviourally) in a negative manner towards black people. That is, we attribute a general trait to him rather than a particular mental state. As the trait conception of attitudes is clearly the prevalent conception of attitudes among folk psychologists, academic psychologists and philosophers may also

want to adopt a trait notion of attitudes (if they do not already do so) because this would facilitate exchange between academia and the wider public (see desideratum *D3* of a model of attitudes as mentioned in the introduction to this thesis). Scholars will find it difficult to inform public discourse on such important issues such as racism or sexism if their notion of an attitude is very different to the attitude notion that ordinary people employ.

To be sure, scholars may sometimes have good reasons not to use the same notions as folk psychologists. Folk psychology can be mistaken, in which case scholarship may actually aim to revise folk psychological notions rather than taking these notions for granted (P. M. Churchland, 1981; P. S. Churchland, 1986; Stich, 1983). Accordingly, if it turned out that folk psychologists are confused about the idea that attitudes are traits, philosophers and psychologists may try to replace the folk psychological notion of attitudes with a more accurate understanding of attitudes. As I will show in the following section, some scholars have in fact argued that folk psychologists are mistaken about the very idea that people possess traits (including attitudes conceived as traits). I will reply that this argument is misguided and that we can make sense of the notion that attitudes are traits after all. Philosophers and psychologists should consider adopting a trait notion of attitudes (if they do not already adopt such a notion) – not only because such a notion is useful for explanatory/predictive and a character evaluative purposes but also because this would make attitude research more accessible to the wider public.

4.5 The situationist challenge

There has been a long-standing debate in psychology as well as in philosophy about the question whether personality and character traits are real (e.g., Bowers, 1973; Hartshorne & May, 1928; Kamtekar, 2004; Merritt, 2000). Some scholars have argued that people's behaviour is largely determined by situational factors and that this speaks against the existence of traits as they are commonly understood, i.e., as psychological dispositions to behave (and cognise) in a consistent manner across different relevant situations (Doris, 2002; Harman, 1999; Hartshorne & May, 1928; Mischel, 1968; Peterson, 1968). According to this line of argument (henceforth, "the situationist challenge"), people are simply mistaken when they ascribe traits to themselves or others. A similar argument has been used to argue against the existence of attitudes understood as general evaluative dispositions (i.e., traits) of people (Schwarz, 2007; Schwarz & Bohner, 2001, Smith & Conrey, 2007; Wicker, 1969). If these authors are right, we cannot explain or predict people's responses towards people qua members of social groups by reference to attitudes understood as traits and neither can we invoke

such attitudes when assessing a person's moral character. All that we can refer to are situational influences on people's responses or perhaps the influence of situation-specific mental states.

In what follows, I will discuss the situationist challenge in some detail. In section 4.5.1, I will review Doris' (2002) argument against the existence of character traits and in section 4.5.2, I will describe how this argument can be applied to the case of attitudes. In section 4.5.3, I will then present a possible rejoinder to the situationist challenge that is open to Machery. It will become clear that on Machery's account the idea that attitudes can be characterised in terms of an aggregate strength and valence is crucial for fending off the situationist challenge. However, as I will show in section 4.6, describing attitudes in these aggregationist terms is problematic because it obscures relevant evaluative complexities of attitudes. This will lead me to develop an alternative trait model of attitudes in section 4.7. The proposed model can both ward off the situationist challenge (described in the present section) and do justice to the complexity of attitudes described in section 4.6.

4.5.1 A situationist argument against the existence of traits

Doris defends the view that character and personality traits as they are usually understood do not exist (Doris, 1998: 2002).[121] He argues that the notion of traits that is common in the philosophical as well as psychological literature and the common folk psychological notion of traits suppose that traits centrally involve dispositions to behaviour (Doris, 2002: chapter 2, chapter 5).[122] When we say that someone is compassionate we assume that the person is disposed to behave in a compassionate manner. Doris (2002) describes this common understanding of character traits by reference to what he calls the principle of consistency:

> Character and personality traits are reliably manifested in trait-relevant behaviour across a diversity of trait-relevant eliciting conditions that may vary widely in their conduciveness to the manifestation of the trait in question. (p. 22)

According to this, a person only possesses a trait if the behaviour that is expressive of the trait (i.e., the trait-relevant behaviour) is shown reliably across diverse situations in

[121] It shall be mentioned that Harman (1999) is another prominent defender of this view in philosophy.

[122] This being said, Doris acknowledges that character traits and virtues are not only expressed in overt behaviour but also in internal psychological processes. Yet, he regards these psychological processes as secondary for an account of traits in so far as they subserve behaviour (Doris, 2002: 17). Doris' focus on behavioural dispositions has been widely criticised (Besser-Jones, 2008; Kamtekar, 2004; Webber, 2006; Webber, 2013). Machery (2016), by contrast, does not seem to prioritise behavioural over psychological dispositions in his characterisation of traits. This implies that the situationist objection must be slightly adjusted to be applied to Machery's account of attitudes (see next section).

which the behaviour is appropriate.[123, 124] An example will help to point out what Doris means by "trait-relevant behaviour", "trait-relevant eliciting conditions", and by "conduciveness to the manifestation of the trait" (p. 22). Let us consider the trait of compassion, which Doris also uses as his central test case. Compassionate behaviours, such as helping or comforting other people, are the trait-relevant behaviours in this case. In general we would expect the compassionate person to behave compassionately in situations in which she is confronted with the suffering of other beings. The distress of others is thus a trait-relevant eliciting condition. Yet, situations in which one is confronted with other's distress may differ in how conducive they are to compassionate behaviour. For example, it is more difficult to act compassionately when one is generally in a bad mood or under time pressure than when one feels elevated and has plenty of spare time. According to Doris, we should expect of a compassionate person that she acts compassionately even when it is relatively difficult to do so. Thus, compassion-relevant situations that are not especially conducive to compassionate behaviour are especially diagnostic when it comes to ascribing the trait of compassion to a person (Doris, 2002: 19). Doris is quite aware that it is a delicate matter to decide how consistently a person must behave compassionately across situations in which compassionate behaviour is appropriate in order to justify the ascription of the corresponding trait (Doris, 2002: 18-20). Yet, he is convinced that people's failure to act compassionately across different compassion-relevant situations is in fact so severe that there remains no doubt that people do not possess a trait of compassion as it is usually understood.

To prove his point, he draws on a vast number of psychological experiments that indicate that seemingly irrelevant situational variables determine whether people act compassionately (Doris, 2002: chapter 3). For example, there is evidence that people are much more likely to help another person who is apparently in distress when they recently had good luck (Isen & Levin, 1972) and less likely to do so when they are in a hurry (Darley & Batson, 1973) or when there is another person present who could help but who remains passive (Latané & Darley, 1970). Moreover, the famous Milgram experiments have revealed that most people can be persuaded to apply severe electric shocks to another person if the instructions are given by an authority figure who insists that the electric shocks are necessary for the success of the experiment (Milgram, 1974). Doris takes these findings to show that aspects of the situations in which people

[123] Doris mentions two further principles: The principle of stability, which is implied in the principle of consistency (if trait-relevant behaviour is consistent across different trait-relevant situations, it can also be expected to be stable across repeated occurrences of the same trait-relevant situation), and the principle of evaluative integration, which concerns the relation between different traits that constitute a person's character.

[124] Very similar characterisations of trait possession can be found in Goldie (2004: 50) and Merritt (2000: 365).

find themselves determine whether they behave compassionately or not. Accordingly, it would be wrong to assume that some people possess a robust trait that disposes them to behave compassionately across different trait-relevant situations.

Yet, Doris acknowledges that people may possess "highly contextualised dispositions or 'local' traits" (Doris, 2002: 64). For example, someone may consistently help people in distress when being in a good mood and when not being under time pressure. Accordingly, we may say that that person possesses "good mood and spare time compassion", while lacking any form of "bad mood" or "in a hurry compassion". Doris stresses that this "localised" way of speaking about traits is at odds with the usual model of traits as general behavioural dispositions.

4.5.2 A situationist argument against the existence of attitudes conceived as traits

In section 4.3.2, we saw that Machery (2016) claims that his trait view may help to predict how particular contextual factors affect the measurement of attitudes. Yet, one may object that contextual influences of this sort undermine the very notion of attitudes conceived as traits in the first place. In analogy to Doris argument against the existence of personality traits, one can construct the following argument against the existence of attitudes conceived as traits:

P1) If attitudes are traits of a person, a person's evaluative responses towards members of a particular social group (i.e., the attitude-relevant responses) should be consistent across various situations in which members of the group are present (i.e., across attitude-relevant eliciting conditions).

P2) Yet, a person's evaluative responses towards members of a particular social group (i.e., the attitude-relevant responses) are not consistent across various situations in which members of the group are present (i.e., across attitude-relevant eliciting conditions).

C) Hence, people do not possess attitudes understood as traits.

This argument is implicit in the work of a range of authors who call into question the notion of attitudes as broad response dispositions (Schwarz, 2007; Schwarz & Bohner, 2001; Smith & Conrey, 2007; Wicker, 1969). I deliberately refer to "attitude-relevant-responses" instead of "attitude-relevant behaviours" in the above argument because cognitive and affective responses are commonly understood to be as indicative of attitudes as behavioural responses. Recall, for example, that on Machery's (2016) model, attitudes are "broad-track dispositions to behave and cognize (have thoughts,

attend, emote, and so on) toward an object […] in a way that reflects some preference" (p. 112). Thus, according to premise 1 of the above argument, if a person possesses a negative (or positive) attitude towards black people, we should expect that that person *consistently* shows negative (or positive) behavioural, cognitive, and emotional responses towards black people across different situations. According to premise 2, this is not what we actually find. In the remainder of this section, I will review some evidence that supports premise 2 and elaborate on the conclusion that follows from this if we also accept premise 1. In the next section (section 4.5.3), I will then show that Machery's account provides us with resources to reject premise 1. Yet, this reply comes at a price as I will show in section 4.6.

There is plenty of evidence of situational influences on people's evaluative responses towards social groups (for reviews see Blair, 2002; Dasgupta, 2013; Smith & Semin, 2004, 2007). In fact, I have already mentioned various examples of the situation-specificity of evaluative responses in previous parts of this thesis (see, for example, section 1.2.2; section 2.2.2.3; section 3.3.3; and section 4.3.2). In section 4.3.2, I mentioned for example the finding that induced disgust promotes people's bias against homosexuals on an IAT, while induced anger fosters bias against Arabs on an IAT (Dasgupta et al., 2009). Note that the emotions, induced by thinking about disgusting or anger-eliciting autobiographical events, can be seen as aspects of the situation that the participants found themselves in, and which influenced their evaluative responses.[125] Recall also that Amodio and Hamilton (2012) found that when their participants were made to believe that they were soon to interact with a black person, they showed a stronger bias against black people on an Eval-IAT than when they expected to interact with a white person. This shows again that aspects of the immediate situation that people find themselves in (i.e., whether they expect to interact with a black or white person) affect their evaluative responses. Richeson and Ambady (2001) conducted a similar experiment with gender as target category. They found that male participants who expected to interact with a woman in a superior role relative to them showed a bias against women on a gender Eval-IAT, whereas male participants who expected to interact with an equal-status or subordinate-status female partner showed a bias in favour of women on the Eval-IAT. One may want to object that participants in this experiment did not exhibit different evaluative responses towards women in different situations but different responses towards different kinds of women (female superior vs. female non-superior). Yet, it must be stressed that if we are interested in a person's attitude towards women in general, characteristics of a

[125] Note also that the evidence that Mandelbaum (2016) provides in support of his view that implicit attitudes are not associative can similarly be interpreted as evidence for the situation-specificity of evaluative responses (see section 2.2.2.3 in chapter 2).

particular woman that a person is confronted with (or expects to be confronted with) can count as situational factor on a broad reading of "situation".[126] It is also worth noting that, on a broad reading of "situation", the common finding that people report positive attitudes towards black people when directly asked for their attitude but exhibit biases against black people on indirect measures of attitudes can be construed as the result of situational influences (e.g., Dovidio et al., 1997; Fazio et al., 1995; see also section 1.3.2). That is, we can construe the different ways that attitudes are accessed (by asking the participant directly or by engaging the participant in a categorisation task such as the IAT) as different situations that trigger different responses. One situation (when participants are directly asked for their attitude) allows the participant to deliberate on her response (thus allowing for a controlled response), whereas the other situation (when participants need to react as quickly as possible on a categorisation task) prevents such deliberation. Analogously, we can interpret the low correlations between people's results on different indirect measures of attitudes (see section 4.3.1 of this chapter and section 1.3.2 of chapter 1) as the result of situational influences. Different measures, such as the affective priming task and the Eval-IAT, involve different procedures, which can be construed as situational factors that influence the measurement outcomes. As a consequence, we should expect people's results on these measures to be dissociated. Also, if responses on indirect measures are highly sensitive to situational factors (such as the details of the measurement procedure), it should come as no surprise that outcomes on these measures are relatively weak predictors of discriminatory responses in real-world contexts (see section 4.3.3 of this chapter and section 1.3.3 of chapter 1).

To sum up, there is plenty of reason to believe that premise 2 of the above argument is true: people's evaluative responses towards members of a particular social group are not consistent across various situations in which members of the group are present (i.e., across attitude-relevant eliciting conditions). Accordingly, one may claim in the spirit of the situationist argument that people do not possess attitudes understood as general dispositions to show evaluative responses of a certain valence (i.e., attitudes understood as traits). Instead we may say that people possess "local attitudes" (Doris, 2002: 87). People may for example exhibit the "being confronted with a subordinate women attitude" and the "being confronted with a superior women attitude" or the "having been asked to report an attitude towards black people attitude". Such a "local" situation-specific conception of attitudes has indeed many proponents in psychology (Conrey & Smith, 2007; Mitchell, Nosek, & Banaji, 2003; Schwarz, 2007;

[126] As I will point out in section 4.7.3, depending on our interests and purposes, we may want to individuate situations in different ways.

Schwarz & Bohner, 2001; Smith & Conrey, 2007).[127] These scholars identify attitudes with highly *situation-specific occurrent evaluative responses*, sometimes also called "constructed" attitudes (Schwarz, 2007). However, it does not seem accurate to say that our attitudes towards a social group cease to exist when we do not currently undergo a response to that group (see Schwitzgebel, 2010: 543, for a similar argument). I therefore prefer to conceptualise local attitudes as dispositions. They can be construed as highly situation-specific dispositions to exhibit evaluative responses of a certain kind. On this view, people can possess multiple attitudes towards a social group at the same time, all of which are tied to particular situations. Yet, we may not even need to bother about the accurate conceptualisation of attitudes from a localist perspective if we can make sense of the notion of attitudes as "global" traits after all. In the next section, I will show how this may work. I will argue that the notion that attitudes have an aggregate strength and valence provides us with a reply to the situationist challenge. However, as I will show in section 4.6, this reply comes at a price: describing attitudes in these aggregationist terms is problematic because it obscures relevant evaluative complexities of attitudes.[128]

4.5.3 A rejoinder to the situationist argument against the existence of attitudes conceived as traits

Although Machery (2016) does not directly discuss the situationist argument as I have presented it above, his characterisation of attitudes suggests that he would likely reject premise 1.[129] That is, his model allows rejecting the claim that we can only ascribe an attitude conceived as a trait to a person if that person shows consistent evaluative responses across different attitude-relevant situations. Recall that attitudes, conceived as traits, can vary in strength according to Machery. Accordingly, Machery can insist that we would only expect total consistency in a person's evaluative responses towards a social group if that person has an *extremely strong* attitude towards that group. Consider a case in which all of a person's mental states and processes in regard to black people reflect negativity. Here, we would in fact expect that irrespective of the situation in which the person encounters a black person, the person will show

[127] The situationist perspective on attitudes is in fact as old as attitude psychology. Already Allport (1935), reviews (and rejects) in his seminal article on attitudes what he calls "The Case for Specificity" (p. 820).
[128] The issue of local attitudes will be taken up again in section 4.7.3.3, in which I will describe what status local attitudes have on my proposed trait model of attitudes.
[129] Machery (2016) presents the situationist argument in very general terms and his response is somewhat sketchy: he claims that the outcome of the person-situation debate has been that "[p]roperties of the person and situational features both influence behavior" (p. 121). Moreover, he emphasises the role of aggregation over various responses of a person in determining a person's attitude (see main text below for explanation).

cognitive, affective, and behavioural responses that reflect negativity. By contrast, if someone possesses a *weak* negative attitude towards black people, we would actually expect that that person's cognitive, affective, and behavioural responses towards black people are somewhat inconsistent across situations. In particular, the person who has a weak negative attitude towards black people may show negative cognitive, affective, and behavioural responses towards black people in most but not in all situations in which she encounters black people. The reason may be that she harbours some mental states in regard to black people that do not imply a negative evaluation, such as the belief that racism is wrong. In some situations this belief may become sufficiently activated to counteract the influence of negative evaluative responses.[130]

On Machery's model, we can determine whether a person possesses a positive or negative attitude towards a social group by aggregating over a person's evaluative responses on various occasions. If a person shows negative responses towards black people across all (or nearly all) observed situations, we can be reasonably confident that that person has a strong negative attitude towards black people.[131] If a person shows negative evaluative responses towards black people in most but clearly not all instances, we may say that the person has a weak negative attitude towards black people. Finally, we can infer that a person lacks an attitude towards black people if there are as many instances in which the person shows positive evaluative responses towards black people as there are instances in which the person shows negative evaluative responses towards black people. That is, a person does not possess an attitude if aggregated over various occasions the person exhibits no preference for the group. This latter thought is expressed in the following quote, in which Machery (2016) argues against the possibility of ambivalent attitudes:

> [T]he trait picture denies (except perhaps in pathological cases) that people have ambivalent attitudes. If the hypothesized coreferential, differently valenced mental states do not lead to a broad-track disposition to behave and cognize in a way that expresses either a positive or a negative preference, then people simply do not have an attitude toward the relevant object. They will act and cognize in a way that expresses a positive preference in some contexts and a negative preference in other contexts, and their aggregate behavior cannot be predicted (even imperfectly) by postulating a trait. (p. 124)

[130] The point that strong attitudes lead to consistent responses, while weak attitudes are associated with rather inconsistent responses, is also emphasised by Webber (2013, 2016b).

[131] Note that this can only be an estimate because the number of observations will always be limited. We are restricted to observations of evaluative responses because we cannot directly observe the entirety of mental states that a person harbours in regard to a social group and that may issue in evaluative responses towards that group. Note also that it remains unclear how we would determine the valence and strength of a person's attitude if we had access to this set of mental states. In particular, would each mental state figure in the same manner in the calculation or would we give more weight to those mental states that get activated more frequently?

A situationist about attitudes may see this as grist to her mills and argue that people never exhibit an aggregate preference towards any social group because their responses are solely the function of situational influences. However, such a claim is at odds with the empirical evidence as Machery rightly points out (p. 121). Decades of research in attitude psychology have shown that attitude measures that aggregate across various evaluative responses of a subject on various occasions (e.g., various responses of the subject to questionnaire items) do oftentimes reveal aggregate preferences of the subject that have (some) predictive validity (Ajzen, 1988: chapter 3; Epstein, 1983). That is, for many social groups individuals exhibit at least a minor aggregate cross-situational preference, which allows ascribing at least a weak attitude to them (i.e., a weak positive attitude or a weak negative attitude).

To conclude, there is a possible reply to the situationist argument against the existence of attitudes construed as traits. According to this reply, situationists set the bar too high when they claim that in order to possess an attitude towards a social group, one's evaluative responses towards members of that social group must be consistent across various situations in which members of the group are present.[132] According to Machery, attitudes can vary in strength (i.e., degree of possession), and thus we should only expect very strong attitudes to lead to perfectly consistent evaluative responses across different situations. Weak attitudes, by contrast, are characterised by imperfect cross-situational consistency in evaluative responses. Empirically, we can estimate the strength and valence of a person's attitude towards a social group by aggregating over various observed cognitive, affective, and behavioural responses of that person towards members of that social group.

The foregoing suggests that aggregation carries much weight in Machery's model of attitudes because without this aspect, Machery's trait view would fall victim to the situationist challenge. In the next section, I will point out that presenting attitudes in these aggregationist terms problematically obscures many of those complexities that are characteristic of attitudes. In section 4.7, I will then present an alternative view of attitudes as traits that wards off the situationist challenge without obscuring the evaluative complexities of attitudes. This is the view that attitudes can be identified with stable profiles of evaluative response variation across situations.

[132] It must be noted that even though this response to the situationist challenge may be successful in the case of attitudes (though see next section), it may not work for all kinds of traits. Doris (2002) acknowledges that different traits have different attributive standards (p. 18-20). He further notes that the attribution of virtues such as loyalty or compassion requires more than just broad behavioural trends over multiple situations (pp. 73-75). According to him, individual situations carry significant weight when it comes to the ascription of these virtues. One single instance of unfaithfulness in the face of sexual temptation may suffice to conclude that a person is not loyal (to any degree). Similarly, we may not want to ascribe the trait of compassion (to any degree) to a person who exhibits a broad cross-situational tendency to help people in distress but who administers severe electric shocks to a person when instructed to do so by an authority figure (Milgram, 1974).

4.6 The complexity of attitudes

We have seen that Machery (2016) characterises attitudes in aggregationist terms. By aggregating across a person's positively and negatively valenced cognitive, affective, and behavioural attitude-relevant responses in various situations, we can estimate how strongly positive or negative the attitude is overall. The fact that attitudes can vary in strength provides us with a plausible response to the siutationist argument against the existence of attitudes conceived as traits. However, as I will argue in this section, Machery's characterisation of attitudes is problematic because it does not do justice to the complex structure of people's attitudes.

Before I turn to this argument, it is worth pointing out that estimating the aggregate strength and valence of an attitude is beset with various difficulties. Note, for example, that there is often no simple fact of the matter as to whether a given response expresses a negative or positive evaluation. Suppose that someone pities black people. We may regard this affective response to be expressive of a positive evaluation if it results from that person's acknowledgement that black people are structurally disadvantaged in society. Yet, we may see it to be expressive of a negative evaluation if that person pities them for supposed inferior mental capacities. Relatedly, we cannot even be sure whether the bias against black people that is usually found on race IATs is (entirely) of negative evaluative nature (Oswald and colleagues, 2013: 186-187). For example, the bias may just reflect (amongst others) that participants are less familiar with black than with white faces. Note that if it is difficult to determine for any single response whether it expresses a positive or negative evaluation (or no evaluation at all), it will be incredibly difficult to determine the overall strength and valence of a person's attitude. After all, we have to aggregate over the individual response tendencies to estimate the overall strength and valence of the attitude. Note also that it remains unclear how we would make valences of cognitive, affective, and behavioural responses commensurable. How would we for example set off against each other the supposed positive valence of a person's belief that Chinese people are clever and the supposed negative valence of the same person's envy of Chinese people? In what follows, I will assume for the sake of the argument that these difficulties can be dealt with. Let us assume that there is a convincing and reliable way to determine (or at least estimate) the aggregate strength and valence of a person's attitude towards a social group on the basis of that person's evaluative responses towards that group. This leaves us still with the question as to whether this is a desirable characterisation of attitudes. In what follows, I will argue that it is not.

Let us consider again the case of Sarah and assume for the moment that she responds slightly more often in a negative way towards black people (e.g., based on

various negative evaluative stereotypes that she harbours in regard to black people) than in a favourable manner (e.g., based on her egalitarian beliefs, her desire not to behave in a racist manner, etc.). On Machery's model we would have to say that she has a weak negative attitude towards black people. Note that this characterisation obscures the evaluative conflict that Sarah is experiencing. If someone told us that she has a weak negative attitude towards black people, we may get the false impression that all her cognition, affect, and behaviour in regard to black people has a somewhat negative valence. This is clearly different from Sarah's actual evaluative stance towards black people, which is not that homogeneous.

Let us now suppose a slightly different scenario. Suppose that Sarah responds as frequently in a favourable manner towards black people (e.g., based on her egalitarian beliefs, her desire not to behave in a racist manner, etc.) as she responds in a negative way towards black people (e.g., based on various negative evaluative stereotypes that she harbours in regard to black people). On Machery's attitude model, we would have to say that Sarah lacks an attitude towards black people in this case. This is because her "aggregate behavior cannot be predicted (even imperfectly) by postulating a trait" (Machery, 2016: 124). That this is unconvincing becomes clear when we compare her to a person, say Liang, whose entire cognitive, affective, and behavioural responses towards black people are pretty much neutral in valence. We may consider that Liang's beliefs about black people do not have any particular valence, that he does not experience any noteworthy affect in regard to black people, and that his behavioural responses in regard to black people are not any different to his behavioural responses towards white people. On Machery's trait view, both Sarah and Liang would be characterised as lacking an attitude towards black people. After all, both Liang's and Sarah's responses towards black people are neutral on balance. Yet, describing them in the same way is clearly problematic because this obscures crucial differences between Liang and Sarah. Note that Sarah and Liang will rarely respond to black people in the same manner. Whereas Liang may in fact be said to lack an attitude towards black people, Sarah's attitude seems to be conflicted rather than non-existent.[133] We cannot just average across Sarah's cognitive, affective, and behavioural dispositions in regard to black people if we want to give an accurate account of her attitude. Only a model of attitudes that does justice to the complexities of Sarah's aversive racism will allow for accurate predictions of her responses and an appropriate evaluation of her character (see functions *F1* and *F2* of the attitude notion as mentioned in the introduction to this thesis). In the next section, I will outline such a

[133] Yet, note that even in the case of Liang, we may not want to say that he lacks an attitude towards black people (or that he possesses a neutral attitude towards black people) but that his attitude towards black people is one of indifference.

model. Before I turn to this account, however, it is worth emphasising that aversive racism is not the only phenomenon that Machery's model struggles to give an appropriate account of.

This becomes clear when we turn our attention to a phenomenon that Glick and Fiske (1996, 1997) have called "ambivalent sexism". They provided evidence that people's (both men's and women's) attitudes towards women often combine hostile and benevolent sexist elements. Women are, for example, often stereotyped as warm, friendly, nurturing, gentle, or understanding (Eagly & Mladinic, 1994). These stereotypes are benevolent in the sense that the stereotype holder perceives them as positive characteristics and that they lead to behaviour that is generally considered to be favourable (e.g., the protection of women from potential threats). Yet, people who hold these benevolent stereotypes are also more likely to harbour distinctively negative stereotypes that characterise women for example as incompetent. Benevolent and hostile sexism often go together because "benevolent sexism may be used to compensate for, or legitimate, hostile sexism" (Glick & Fiske, 1996: 492). Moreover, benevolent and hostile sexist beliefs jointly justify existing social power relations (Jost & Kay, 2005). For example, a man, say Jack, who believes that women in high-status occupations (e.g., university professor, judge, CEO) are not fit for their job and must have been lucky to obtain these positions (hostile sexism) may also believe that women have great social skills, which makes them supposedly better suited for assisting or caring roles (benevolent sexism). Note that ambivalent sexism is also reflected in Richeson and Ambady's (2001) aforementioned finding that men who expect to interact with a woman in a superior role relative to them, exhibit a bias against women on a gender Eval-IAT, whereas men who anticipate to interact with an equal-status or subordinate-status female partner exhibit a bias in favour of women on the same measure (see section 4.5.2). Following Machery's trait view, ambivalent sexists like Jack may appear to lack an attitude towards women because the disposition to show negative responses towards women in superior roles and the disposition to be kind to women in same status or subordinate roles may balance each other. Recall that according to Machery "the trait picture denies (except perhaps in pathological cases) that people have ambivalent attitudes" (p. 124). Yet, contra Machery, it seems intuitive to say that Jack has a (non-pathological) ambivalent attitude towards women.

Note that ascribing an ambivalent sexist attitude to Jack fulfils both an explanatory/predictive and a character evaluative function (see functions *F1* and *F2* of the attitude notion as mentioned in the introduction to this thesis). While Machery (2016) may be right that we cannot predict Jack's "aggregate behaviour" on the basis of such an ascription (p. 124), we can make specific predictions about Jack's likely

responses in regard to different kinds of women. Knowing about his ambivalent sexism (and not just about the aggregate strength and valence of his attitude), we may for example predict that he will likely feel uncomfortable about having a female supervisor at work and that he will feel approval towards females who spend a lot of time with their kids. Moreover, this ascription gives us information about Jack that we can take into account when assessing Jack's character. I will further elaborate on the nature of ambivalent sexist attitudes in the next section.

For now though, it is important to point to another problem with Machery's (2016) trait model: it obscures the variety of affects involved in attitudes. In the last chapter, I stressed following Madva and Brownstein (2016) that prejudice is not just a matter of generic feelings of like or dislike but often of specific emotions such as anger, disgust, fear, or pity. For example, I mentioned experiments that indicated that in many people prejudice towards homosexuals is linked to disgust, while prejudice towards Arabs is linked to anger (Dasgupta et al., 2009; see section 3.3.3). On Machery's account, both the person who is disgusted by homosexuals and the person who is angered by Arabs will likely be described as having a negative attitude towards the respective group (if most of their response dispositions towards these groups are reflective of negativity). However, characterising their attitudes merely in terms of negative valence obscures the affective differences between these attitudes. Note that disgust and anger are not only experienced differently by the subject but are also linked to different kinds of behaviour. While one may be inclined to avoid the person whom one is disgusted by, one may approach and confront the person who is the target of one's anger. Hence, acknowledging the particular nature of the affective reaction of a person towards a social group helps us making better predictions about that person's responses to members of that group (see section 3.4.2 in the previous chapter). A model of attitudes should do justice to these affective complexities.

To sum up, I have argued in this section that Machery's account of attitudes problematically obscures crucial details that are characteristic of people's attitudes. It obscures evaluative conflicts that are often characteristic of people's attitudes, such as when people are alienated from their own racial biases or when people exhibit both benevolent and hostile sexist dispositions. Moreover, describing attitudes just in terms of strength and generic positive or negative valence does not do justice to relevant differences in the emotional reactions that people exhibit in response to different social groups. Taken together, Machery's trait model obscures factors that would help us to predict people's evaluative responses and to assess their character. In the following, I will argue for an alternative characterisation of attitudes that is supposed to do justice to these aforementioned complexities but still allows conceptualising attitudes as traits.

4.7 Profiles of situation-specific evaluative response dispositions

I propose that we can address the complexity of attitudes by characterising them as complex profiles of situation-specific evaluative response dispositions. In the following sub-sections, I will give substance to and flesh out this characterisation. I will start with an elaboration on Mischel and Shoda's (1995) influential cognitive-affective personality system model (CAPS) because it provides the framework for my proposed account of attitudes (section 4.7.1).[134] In response to the situationist challenge against the notion of personality traits (see section 4.5.1), Mischel and Shoda argue that people exhibit stable profiles of behaviour variation across situations that can be identified with personality traits. In other words, traits are profiles of situation-specific behavioural dispositions. In section 4.7.2, I will show that this idea can also be applied to attitudes understood as traits. According to this, attitudes are stable profiles of situation-specific (cognitive, affective, and behavioural) evaluative response dispositions. For example, a person's ambivalent sexist attitude can be understood as the profile to be disposed to respond in a benevolent manner towards women in same-status or subordinate roles and to be disposed to respond in a hostile manner towards women in superior roles. I will argue that this conception of attitudes both fends off the situationist challenge (mentioned in section 4.5) and gives justice to the evaluative complexities of attitudes (mentioned in section 4.6). In section 4.7.3, I will point out that on the proposed account of attitudes there are different legitimate ways to individuate attitudes, which depend on our interests and purposes as attitude ascribes. In section 4.7.4, I will elaborate on the psychological basis of attitudes as understood on this account. In section 4.7.5, I will point out that the proposed profile view of attitudes can equally well explain those findings that Machery (2016) claims are best explained on his model of attitudes (see section 4.3). Finally, in section 4.7.6, I will point out some significant similarities and differences of my model of attitudes to Schwitzgebel's (2013) dispositional model of attitudes to further locate my account in the literature.

[134] Another author who links attitudes to the CAPS model is Webber (2013, 2016b). However, his project is different from mine. Whereas I provide an account of attitudes towards social groups in terms of traits, he provides an account of traits such as circumspection or cruelty in terms of attitudes. He argues for example that circumspection can be analysed as a strong positive attitude towards caution or that cruelty can be analysed as a strong positive attitude towards other people's suffering (Webber, 2013: 1086-1087). Due to this difference in our projects, I believe that my account is not necessarily incompatible with his. It could be that attitudes towards social groups are characterisable as specific traits that are not again characterisable in terms of attitudes (see section 4.7.2 below), while Webber (2013) is right that traits such as circumspection and cruelty are analysable in terms of attitudes. In short, our accounts may be concerned with different kinds of traits. Webber does not directly discuss attitudes towards social groups. Yet, he argues that one can reduce one's susceptibility to implicit biases by "instil[ing] in oneself a few firmly held moral attitudes, such as attitudes in favour of fairness or against discrimination" (Webber, 2016b: 149).

4.7.1 Cognitive-affective personality system

Mischel and Shoda (1995) developed their model of a cognitive-affective personality system (CAPS) in response to the situationist challenge against the notion of personality traits. They argue in their highly influential paper that the fact that situational factors strongly influence people's behaviours does not undermine the claim that people possess traits and stable personalities. They stress that, quite to the contrary, "behavioral variation in relation to changing situations constitutes a potentially predictable and meaningful reflection of the personality system itself" (p. 255). This is because the personality system generates "stable *if...then...* profiles of behavior variability across situations" (p. 252).

Evidence for these stable situation-behaviour profiles has been gathered, amongst others, in a large-scale field study in which children's behaviour was observed across a variety of situations in a summer camp setting (Shoda, Mischel, & Wright, 1994). Over the course of 6 weeks, children's behaviour was recorded on various behavioural dimensions (e.g., verbal aggression, whining, complying) as it occurred in five different types of interpersonal situations. These types of situations included situations in which a peer teased the child, situations in which a peer initiated positive social contact with the child, situations in which an adult warned the child, situations in which an adult punished the child, and situations in which an adult praised the child. The researchers found that the observed children exhibited characteristic profiles of behaviour variation across the five types of situations. Consider for example a child who shows a high level of verbal aggression when being teased by another child, a medium level of verbal aggression when being warned or being punished by an adult, and a low level of verbal aggression when being positively approached by another child or when being praised by an adult. To say that this situation-behaviour profile is characteristic of the child's personality (i.e., reflective of a trait of the child) is to say that the child tends to show the same (or a very similar) situation-behaviour profile whenever it encounters this set of situations and that the profile can be distinguished from other children's profiles. Shoda and colleagues (1994) found stable intra-individual situation-behaviour profiles of this sort for all behavioural dimensions that they analysed. Mischel and Shoda (1995) take this to show that behavioural variation across situations is not due to random fluctuations or "due to situation rather than the person" but in fact the product of a stable personality system (p. 257). They note that it is common in psychology to average behavioural indices of traits across different situations and stress that this is highly problematic because "it actually removes data that may alert us to the person's most distinctive qualities and to his or her unique intraindividual patterning of social behaviour" (p. 251).

According to Mischel and Shoda, the personality system, which gives rise to the above mentioned situation-behaviour profiles, is characterised by the cognitive and affective mental states that are available to the subject, the distinct set-up of interrelations between these cognitions and affects, and the relation of these cognitions and affects to aspects of situations (Mischel & Shoda, 1995: 254). They refer to these cognitions and affects as "cognitive-affective units" and posit that these include encodings (i.e., categories), beliefs (e.g., about the social world), affects (e.g., feelings and emotions), goals, and behavioural scripts. Aspects of situations activate and deactivate subsets of these cognitive-affective units. The activated units in turn activate or deactivate those units to which they are connected and so forth, ultimately producing a specific behavioural output. Mischel and Shoda call such a sequence of activation and deactivation of cognitive-affective units "processing dynamics" (p. 257) and emphasise that these dynamics may "operate at many levels of awareness, automaticity, and control" (p. 259). As different situational features activate different cognitive-affective units and thus trigger different processing dynamics, people behave differently in different situations. However, as the personality system is relatively stable, people exhibit stable situation-behaviour profiles across time. Moreover, as people differ in cognitive-affective units available to them and in how these cognitive-affective units are related to each other and to situational features (i.e., the processing structure), different people exhibit different situation-behaviour profiles.

Yet, it is possible to identify personality types or personality traits that are shared by different individuals according to Mischel and Shoda (1995: 257-258). As the structure of the personality system is reflected in situation-behaviour profiles, commonalities in people's situation-behaviour profiles can be taken as evidence for commonalities in people's personalities (i.e., for commonalities in the organisation of cognitions and affects that constitute the personality system). A personality trait that can be identified by analysing situation-behaviour profiles is for example rejection sensitivity (Mischel & Shoda, 1995: 258; Ayduk & Gyurak, 2008). We can identify rejection sensitivity by closely observing a person's behaviour towards his romantic partner (or another loved person) in various situations. If a person tends to behave in exceedingly kind and tender ways towards his partner across a range of situations in which the partner's attention is on him but tends to show aggressive or abusive behaviours in situations in which his partner's attention is on other people or behaves in a way that may be interpreted as uncaring, we may infer that the person is rejection sensitive. Rejection sensitive people share certain processing dynamics (Ayduk & Gyurak, 2008). They characteristically exhibit anxious expectations of rejections, which leads them on the one hand to behave in exceptionally kind ways towards their partners to prevent rejection. Yet, their anxious expectations of rejection make them on

the other hand more likely to perceive and interpret their partner's behaviour as reflective of rejection. Such perceptions in turn lead to feelings and thoughts of hostility towards their partner that can result in aggressive behaviour. Thus, although the person's behaviours may seem inconsistent (tender and aggressive behaviours towards a loved one), these behaviours are actually expressions of the same personality trait (i.e., rejection sensitivity), which can be analysed as a characteristic profile of situation-specific response dispositions towards a loved one.

4.7.2 Attitudes on the profile view

Mischel and Shoda's (1995) CAPS model provides us with a useful framework for conceptualising attitudes in a way that neutralises the situationist challenge presented in section 4.5 without obscuring the evaluative complexities of attitudes that have been mentioned in section 4.6. Just as we can conceptualise traits such as rejection sensitivity as a profile of situation-specific response dispositions, I propose that we can conceptualise attitudes as profiles of situation-specific response dispositions. On this view, people can possess robust attitudes, conceptualised as traits, despite the pervasive influence of situational factors on evaluative responses. In order to contrast the proposed view to Machery's (2016) trait view of attitudes, I will call this the profile view of attitudes. This should not be taken as indication that attitudes are not traits. They are traits, but traits of a different sort than proposed by Machery as I will outline in what follows.

In section 4.6, I argued that evaluative complexities, such as when people harbour at the same time hostile and benevolent sexist dispositions towards women, should be accounted for by our model of attitudes. On the profile view, such an attitude towards women (i.e., an ambivalent sexist attitude) can be understood as the profile to be disposed to respond in a benevolent manner towards women in same-status or subordinate roles and to be disposed to respond in a hostile manner towards women in superior roles. To use an expression by Mischel and Shoda (1995), the ambivalent sexist, say Jack from my example in section 4.6, exhibits a "stable *if...then...* profile[] of behavior variability across situations" (p. 252).[135] If confronted with a woman in an equal-status or a subordinate role, he shows benevolent behaviour (as well as

[135] One may object (as already hinted at in section 4.5.2) that this attitude is not a profile of *situation*-specific response dispositions but a profile of *person*-specific or *role*-specific response dispositions. I grant that these would be valid ways to describe the attitude. Yet, it must be stressed that a broad characterisation of a social situation encompasses characteristics of the person that one is confronted with. I prefer to characterise attitudes as profiles of *situation*-specific response dispositions because this is general enough to qualify as a characterisation of various different attitudes, including the aversive racist attitude as described below. As I will argue in the next section, we may want to individuate situations (and thus attitudes) in different ways dependent on our interests and purposes.

benevolent affective and cognitive responses) and if confronted with a woman in a superior role, he shows hostile behaviour (as well as hostile affective and cognitive responses). This stable profile, or we can say this trait, is Jack's attitude towards women. It is thus wrong to assume, as situationists do, that the fact that people show different evaluative responses towards members of a social group in different situations speaks against the existence of attitudes understood as traits (see section 4.5.2). Quite to the contrary, it is a crucial feature of attitudes on the profile view that they incorporate response dispositions that are tied to particular situations.

Recall that on Machery's (2016) trait view, a person who exhibits hostile and benevolent evaluative responses towards women in equal measure would be said to lack an attitude towards women. After all, the person's response dispositions are neutral on balance. This is a problematic characterisation because the ambivalent sexist clearly differs from a person whose entire responses towards women are neutral in valence (and who could indeed be said to lack an attitude towards women). If we adopted Machery's trait view of attitudes, we would not be able to use the attitude notion to convey information about what responses the ambivalent sexist is likely to show towards different kinds of women or about the ambivalent sexist's moral character (see functions *F1* and *F2* of the attitude notion mentioned in the introduction to this thesis). The profile view, by contrast, highlights the complexity of the ambivalent sexist's attitude towards women and thus facilitates predictions about the agent's likely responses towards women and an assessment of the agent's moral character.

Note also that the profile view is well suited to account for the complexity of Sarah's aversive racist attitude. Sarah's attitude can broadly be analysed as the profile to show favourable responses concerning black people in situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments and to show negative responses with respect to black people in situations in which she does not have sufficient time (e.g., when she has to judge quickly whether she is in danger) or cognitive resources (e.g., when she is occupied with the detection of potential threats or when she is deeply engaged in a conversation with her patients) to reflect on and be guided by her endorsed egalitarian commitments. Knowing that Sarah has an aversive racist attitude helps us to predict whether she will respond in a favourable or negative way towards a given black person, dependent on the situation that Sarah is in. If Sarah is under time pressure or engaged in an attention-demanding task, she is likely to exhibit negative responses towards the person, but if she has time and the appropriate cognitive resources to reflect on her responses, she is likely to exhibit a favourable response towards the person. Note also that we may predict Sarah's results on indirect and direct measures of attitudes based on the fact that she exhibits the profile of an aversive racist (see section 1.3 in chapter

1 for an elaboration on these measures). On an IAT, for example, Sarah is required to respond as fast as possible in a categorisation task that requires her full attention. We should thus expect that her responses will not be guided by her anti-racist commitments (e.g., she may respond quicker when black people and negative words are paired than when white people and negative words are paired). Direct measures, by contrast, allow Sarah to deliberate on her responses. Accordingly, we would expect that her responses on these measures fall in line with her egalitarian commitments (e.g., she may attribute as many positive and negative traits to black people as to white people).

So far I have presented the profile view of attitudes by reference to the examples of ambivalent sexist and aversive racist attitudes. However, I would like to stress that attitudes can take a variety of forms on the profile view. Empirical research may reveal what profile attitudes are prevalent in a given population of people. Recall that Shoda and colleagues (1994) identified different profiles of verbal aggressiveness in children by observing their behaviour in different situations. I propose that we can identify attitudes in a similar manner by observing people's cognitive, affective, and behavioural responses towards members of the target group in different situations.[136] This research will likely reveal that there are many details in which the attitudes of different people differ. Yet, it should be possible to identify relevant similarities among these profiles and to group them in terms of certain attitude types. For example, relevant types of attitudes towards women may include the following:

- Ambivalent sexist attitude: benevolent responses towards women in same status or subordinate roles; hostile responses towards women in superior roles.
- Aversive sexist attitude: favourable responses towards women in situations in which the agent has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments; negative responses towards women in situations in which the agent does not have sufficient time or cognitive resources to reflect on and be guided by her endorsed egalitarian commitments.

[136] As mentioned earlier, attitudes are not only constituted by behavioural dispositions but also include cognitive and affective dispositions. Note that a person whose thoughts and emotions in regard to women in superior roles reflect hostility and whose thoughts and emotions in regard to women in subordinate roles are benevolent may still be said to have an ambivalent sexist attitude, even if this attitude is not (or only rarely) translated into overt behaviour. This implies that when determining the types of profile attitudes that people harbour, we should not restrict our investigation to the observation of overt behavioural responses (although these are of course as well important indicators of cognitive and affective responses). We can also use physiological measures to determine the affective responses that people exhibit in different situations or use questionnaires to determine situation-specific judgment dispositions. In short, to cover the whole range of attitude-relevant responses, our examination should take into account an array of cognitive, affective, and behavioural measures.

- Moralising sexist attitude: favourable responses towards women whom the agent perceives to be chaste or demure; hostile responses towards women whom the agent perceives to be unchaste or indecent.
- Attractiveness-dependent sexist attitude: friendly responses towards women whom the agent perceives to be attractive; unfriendly responses towards women whom the agent perceives to be unattractive.
- Firm sexist attitude: hostile responses towards women in all situations.
- Firm egalitarian attitude: neutral or favourable responses towards women in all situations.

This typology of attitudes towards women is not meant to be exhaustive. Nor do I claim that the way in which I flesh out the content of the individual profile attitudes is the only appropriate one (see section 4.7.3 below on attitude individuation). The above list is merely meant to give a first impression of the diversity of attitudes that different people may possibly exhibit in regard to a single social group. Note that if we would characterise attitudes in aggregationist terms, as Machery (2016) does, we would run the risk of missing this diversity. Following Machery, we would have to say that the ambivalent sexist, the aversive sexist, the moralising sexist, and the attraction-dependent sexist all lack an attitude towards women if their respective positive and negative response dispositions in regard to women balance each other. This problematically obscures the differences between the evaluative stances of these persons towards women. Moreover, it deprives us of the ability to use the attitude notion to convey information about the persons' likely responses towards women in different kinds of situations (or towards different kinds of women) and about their moral character (see functions *F1* and *F2* of the attitude notion).

Yet, although I have developed the profile view of attitudes to account for evaluative complexities of attitudes, the profile view does not exclude the possibility of what we may call "uncomplex" attitudes, such as the firm sexist and firm egalitarian attitudes in the above typology of attitudes towards women. Some people may exhibit hostile (or favourable) responses towards women in all or the vast majority of situations in which they encounter women. In this case it may seem tempting to follow Machery (2016) to say that these people have a negative (or positive) attitude towards women, understood as a general tendency to show negative (or positive) responses towards women across a wide range of situations. Yet, I would like to stress that cases like this are the exception rather than the rule. More often than not, people exhibit a range of differently valenced evaluative responses towards members of a particular social group in different situations, and this should be reflected in our model of attitudes rather than be obscured. An "uncomplex" attitude can be described on the profile view as an

extreme case in which a person's profile of situation-specific response dispositions is unusually invariant. Moreover, I suspect that attitudes are more likely to appear "uncomplex" when we analyse them at what I describe below as low levels of situation-specificity and response-specificity (see section 4.7.3.2).

4.7.3 Interest-dependence of attitude individuation

On my view, there are often various legitimate ways to individuate a person's attitude(s) towards a social group if that person's evaluative responses vary across different situations. Of course, attitude ascriptions should track actual dispositions of the agent to be appropriate. However, these dispositions may often be so numerous and diverse that attitude ascribers (including psychologists, philosophers, and ordinary people) need to pick out especially salient or relevant patterns of evaluative responding to give an intelligible account of a person's attitude(s). As I will show in what follows, this process of highlighting particular profiles of situation-specific response dispositions is influenced by our interests and purposes as attitude ascribers. In section 4.7.3.1, I will show that dependent on what situations and responses we take to be relevant, we may end up with different descriptions of a person's attitude(s). In section 4.7.3.2, I will build on this and point out that we may vary how finely we differentiate between different situations and different responses dependent on our purposes as attitude ascribers. In section 4.7.3.3, I will then also stress that dependent on our purposes, we can ascribe relatively global (e.g., attitudes towards women) or relatively local attitudes (e.g., attitudes towards women in superior positions).

4.7.3.1 Picking out relevant situations and responses

Above I have mentioned that it is unlikely that a person's evaluative responses towards members of a social group will be totally invariant across different situations. This becomes evident when we consider the sheer amount of different situations in which we may encounter members of a target group. Jack, for example, may encounter women in countless different contexts (at his workplace, at the supermarket, at a sports club, at a parent's evening at school, etc.), and he may encounter women with innumerable different traits (different age, different looks, different professions, different status, etc.). It is extremely unlikely that Jack's evaluative responses towards women will be the same under all these circumstances. Rather we would expect that he exhibits an immensely complex mesh of situation-specific response dispositions that we cannot even comprehend in its entirety, let alone convey to others in conversation. To give an intelligible account of his attitude(s) towards women, we need to extract

especially salient and noteworthy patterns of evaluative responses (i.e., profiles of situation-specific response dispositions) from the more complex mesh of response dispositions that Jack exhibits.

Crucially, what appears salient or noteworthy to us depends on our interests and purposes, and these may vary. We may be interested in how Jack responds to women in different roles (e.g., superior, same-status, subordinate roles), which may lead us to realise that Jack exhibits the profile of an ambivalent sexist as described above (see section 4.7.2). Yet, if we are interested in how Jack responds towards women whom he perceives to be attractive versus women whom he perceives to be unattractive, we may (or may not) find that he exhibits an attractiveness-dependent sexist attitude as described above. Note that Jack's responses towards women may possibly be influenced by both of these factors (perceived attractiveness and role), yet, dependent on our explanatory or predictive purposes, we may decide to emphasise only one of these. Consider that we notice that Jack is very friendly to some of his female co-workers but very rude to others, although all of them occupy roughly the same level in the company's hierarchy as Jack. Although Jack may harbour an ambivalent sexist attitude, referring to this fact does nothing to explain/predict his differential responses towards his equal-status female co-workers. Yet, Jack's responses may possibly be explained by reference to the fact that he has an attractiveness-dependent sexist attitude that leads him to respond in a friendly manner towards female co-workers whom he perceives to be attractive and in an unfriendly manner towards female co-workers whom he perceives to be unattractive. Thus, when asked for an explanation for Jack's unequal treatment of his female co-workers, we provide valuable information when we hypothesise that he may have an attractiveness-dependent sexist attitude. Recall that I have characterised attitudes as profiles of situation-specific evaluative response dispositions. The foregoing shows that how we characterise a person's attitude(s) depends strongly on what situational contrasts we take to be noteworthy (e.g., situations in which the agent is confronted with women of different perceived attractiveness or situations in which the agent is confronted with women of different status).

At the same time, attitude individuation is partly a function of what we (as philosophers, academic psychologists, or folk psychologists) take to be the relevant *kinds of responses*. Note, for example, that instead of characterising Jack's responses towards women of varying perceived attractiveness in terms of friendliness or unfriendliness, we could as well describe his responses in terms of the interest or disinterest/disregard that he exhibits in regard to women of varying perceived attractiveness. Crucially, these dimensions may be independent. Someone who responds in an unfriendly manner towards another person is not necessarily

disinterested in that person and someone who responds in a friendly manner towards another person is not necessarily particularly interested in that person. As a consequence, Jack may possess an attractiveness-dependent attitude towards women characterised on the friendliness/unfriendliness dimension, while lacking an attractiveness-dependent attitude characterised on the interest/disinterest dimension (and vice versa). There is no fact of the matter as to which dimension is more adequate. They are both legitimate, and it depends on our interests as attitude ascribes which one we prefer (or whether we want to take both into account).

The fact that we can highlight those evaluative response dimensions that are of interest to us when ascribing attitudes allows us to mark differences in affective responses that would go unnoticed on Machery's (2016) account. In section 4.6, I have argued that attitudes are not only a matter of generic positive or negative valence, as Machery's (2016) describes them, but also of specific emotions. The profile view can account for this. We can, for example, characterise a person's attitude towards homosexuals in terms of the degree of disgust that the person exhibits in regard to homosexual persons in different contexts. Moreover, for different groups different emotions may be of central interest (Dasgupta et al., 2009; Inbar et al., 2012). When it comes to people's attitudes towards Arabs, for example, we may want to characterise the attitude in terms of the degree of anger that people are experiencing rather than focussing on disgust.

4.7.3.2 Situation-specificity and response-specificity

Above, I showed that dependent on what we (as philosophers, academic psychologists, or folk psychologists) take to be the relevant kinds of situations and responses, our characterisations of people's attitudes will differ. Another point that I want to stress is that we may also vary how finely we differentiate between different situations and different responses that we take to be relevant. In the above typology of attitudes towards women (see section 4.7.2), attitudes are characterised rather coarsely in terms of two (or three in the case of the ambivalent sexist) contrasting types of situations. For example, the attractiveness-dependent attitude is characterised in terms of situations in which the agent is confronted with a woman whom he perceives to be attractive and situations in which the agent is confronted with a woman whom he perceives to be unattractive. Note that we may want to make finer distinctions between situations to characterise the attitude in more detail. That is, we may want to describe the attitude on what I call "a higher level of situation-specificity". For example, we could divide the space of possible situations into five classes: situations in which the agent is confronted with a woman whom he perceives to be (1) extremely attractive, (2)

somewhat attractive, (3) neither attractive nor unattractive, (4) somewhat unattractive, (5) extremely unattractive. In principle, we could even assume an infinite number of situations differing in the perceived attractiveness of the target person. This would be the highest possible level of situation-specificity.[137] In the above typology of attitudes, responses, too, are characterised rather broadly (e.g., favourable vs. hostile responses). Again, we may choose to make finer-grained distinctions (i.e., apply a higher level of response-specificity) to characterise attitudes in more detail. We may for example divide responses into five different classes: extremely favourable responses, somewhat favourable responses, neutral responses, somewhat hostile responses, and extremely hostile responses. Alternatively, we could imagine an infinite number of possible responses along the favourable-hostile continuum, which would amount to the highest possible level of response-specificity.[138]

The two dimensions that I have mentioned here, level of situation-specificity and level of response-specificity, are independent in the sense that we can vary the level of response-specificity of our attitude characterisation without varying the level of situation-specificity of our attitude characterisation, and vice versa. Normally, however, we would choose comparable levels of situation-specificity and response-specificity when describing profile attitudes because the reasons that speak for choosing a particular level of situation-specificity equally apply to the choice of a level of response-specificity, and vice versa. Sometimes one may want to identify broad types of attitudes that are prevalent in a given society to get an overview of common evaluative response patterns towards a given social group (as in the typology of attitudes towards women in section 4.7.2). These broad types of attitudes that are shared among a large number of people (e.g., the aversive sexist attitude and the ambivalent sexist attitude) are more likely to be found when one chooses relatively low levels of situation-specificity and response-specificity. Naturally, the more we dissect different types of situations and different types of responses, the more differences we will find in people's profiles of situation-specific response dispositions. When we want to convey information about a person's attitude to others, it is convenient to draw on attitudes at low levels of situation-specificity and response-specificity. These attitudes can be conveyed by using succinct labels such as "aversive sexist" or "ambivalent sexist", which simplifies communication. Yet, descriptions of attitudes on higher levels of situation-specificity and response-specificity also bring a clear advantage with them: they allow for better predictions of people's cognitive, affective, and behavioural responses towards social

[137] To put it in psychometric terms, one could model attractiveness as a continuous variable.
[138] Note that I have tacitly chosen a high level of response-specificity when I suggested above that one may describe a person's attitude towards homosexuals or Arabs amongst others in terms of the *degree* of disgust or anger that the person exhibits in response to homosexual or Arabian persons in different contexts.

groups. Obviously, the more detailed our descriptions of profile attitudes are in terms of responses and situations, the more precise can the predictions of future responses be. For example, knowing that Jack tends to respond with extreme unfriendliness to women whom he perceives to be very unattractive and with milder unfriendliness to women whom he perceives to be somewhat unattractive may help us to make more nuanced predictions than just knowing that Jack tends to respond with unfriendliness towards women whom he perceives to be unattractive.

In short, dependent on our purposes, we may choose to describe profile attitudes at different levels of situation-specificity and response-specificity. If we are interested in making accurate predictions about a person's likely responses towards members of a particular social group, we do better to describe the profile attitude at higher rather than lower levels of response-specificity and situation-specificity. To speak figuratively, we need to zoom in on the attitude. If we are, however, more interested in identifying common patterns of evaluative responding that are also easily conveyable to others, we do well to describe profile attitudes at relatively low levels of response-specificity and situation-specificity. In short, we need to zoom out on the attitude. Yet, we can also zoom out to much. This is the case with Machery's (2016) description of attitudes. By aggregating across all situations in which the agent may encounter a member of the target group, Machery cancels out any situation-specific evaluative response tendency in his characterisation of attitudes. This leaves us with a characterisation that is blind to any of the evaluative complexities that are characteristic of attitudes (see section 4.6).

4.7.3.3 Local and global attitudes

Another way in which attitude individuation is interest-sensitive is marked by the contrast between global and local attitudes. One may wonder why we should say that a person, say Jack, has an ambivalent sexist attitude towards women rather than saying that Jack has both a hostile sexist attitude towards women in superior roles relative to him and a benevolent sexist attitude towards women in equal-status or subordinate roles. Note that this latter way of speaking about Jack's attitude(s) is reminiscent of the "local attitude view" as it has been presented in section 4.5.2. I grant that these ways of speaking about Jack's attitude(s) are complementary. If someone asks us about Jack's general attitude towards women it is appropriate to say that he has an ambivalent sexist attitude (i.e., it is appropriate to refer to the "global" attitude).[139] Yet, if someone

[139] However, it would be misleading to respond to this question that Jack harbours two *conflicting* attitudes towards women. This is because hostile and benevolent sexism are mutually supportive rather than opposing each other. As already mentioned in section 4.6, benevolent sexist stereotypes can be used to legitimate hostile sexist stereotypes, and both forms of sexism justify and support existing social power relations.

asks us more specifically about Jack's attitude towards females in superior roles and females in subordinate roles it is appropriate to say that he has a hostile attitude towards the former and a benevolent attitude towards the latter (i.e., it is appropriate to refer to the "local" attitudes). It must be stressed, however, that these "local" attitudes, too, can be analysed as profiles of situation-specific response dispositions. For example, Jack is likely to respond in slightly different ways to females in different roles superior to him, such as female police officers, female professors, female judges, and so on. That is, Jack shows a distinctive profile of situation-specific response dispositions towards women in higher ranking positions. Again, instead of saying that Jack has a complex attitude towards women in superior roles that is composed of his response tendencies towards females in different superior positions, we may equally well say that Jack harbours different attitudes towards female police officers, female professors, female judges, and so on. And again, these attitudes can themselves be analysed as profiles of situation-specific response dispositions. Jack may for example show slightly different responses towards female professors depending on their specialisation (e.g., female professors in the sciences vs. female professors in the humanities). Accordingly, his attitude towards female professors can be described as a profile of situation-specific response dispositions. As should be clear by now, the predicate "local" as used in "local attitude" is relative (and so is the predicate "global"). Jack's attitude towards female professors is local in relation to his attitude towards females in superior roles and his attitude towards females in superior roles is local in relation to his attitude towards women. Dependent on our interests and communicative purposes, we can ascribe attitudes that are relatively local (e.g., attitudes towards female police officers) or relatively global (e.g., attitudes towards women).[140]

4.7.4 The psychological basis of attitudes

Recall that according to Mischel and Shoda (1995) the personality system is characterised by a distinctive organisation of relationships between cognitive-affective units. Individual traits are based on subsets of these cognitive-affective units. Rejection sensitivity, for example, may be based on a desire not to be rejected, anxious expectations of rejection, the disposition to feel anger in response to rejection, and so

[140] It is important to note that the global-local dimension is independent of the situation-specificity and response-specificity dimensions mentioned in the last section. Note for example that one can describe a global attitude (a person's attitude towards women) at different levels of situation-specificity and response-specificity. Similarly, one can hold the level of situation-specificity and response-specificity fixed while varying the attitude description along the global-local dimension. For example, one can describe two different response dispositions that are tied to two different situations (a relatively low level of situation- and response-specificity) as two distinct attitudes of the person (local attitudes) or as two components of one attitude (global attitude). See the example of the ambivalent sexist in the main text.

forth (Ayduk & Gyurak, 2008). By attributing the trait of rejection sensitivity to a person, we essentially pick out these cognitive-affective units. Similarly, by attributing to a person an attitude towards a group X, we pick out certain cognitive-affective units in that person's personality system (and the links between these) that are involved in evaluative responses towards members of group X. To borrow a term from Machery (2016), we can say that these units constitute an attitude's "psychological basis". In different situations that involve members of group X, different units that form the psychological basis of the attitude become activated. These units may include, amongst others, evaluative stereotypes as characterised in the last chapter (including conceptual and affective mental states), moral beliefs, desires, and behavioural scripts. For example, an ambivalent sexist attitude may be based on conceptual associations that link women to features such as incompetence, kindness, and sociability, various affective dispositions that are linked to these stereotypes, the belief that women are not apt for leadership roles, the belief that women are good carers, the desire to dominate women, and so forth. These various mental states may dispose the ambivalent sexist person to show benevolent cognitive, affective, and behavioural responses towards women in subordinate positions and to show hostile cognitive, affective, and behavioural responses towards women in superior positions.

Recall also that according to Mischel and Shoda (1995), the personality system operates "at many levels of awareness, automaticity, and control" (p. 259). This is arguably also true for those cognitive-affective units that form the psychological basis of profile attitudes: some of them may be easier to introspect than others (see section 1.2.5 in chapter 1); some of them may only be subject to indirect rational control, while others are subject to direct rational control (see section 2.3.1 in chapter 2); and some of them may only be subject to indirect intentional control, while others are subject to direct intentional control (see section 2.3.2 in chapter 2). That is, the psychological basis of profile attitudes may include both mental states that would commonly be described as implicit and mental states that would commonly be described as explicit (and possibly mental states that lie somewhere in-between).

4.7.5 Attitude measurement

In section 4.3, I presented Machery's (2016) argument to the best explanation. According to him, his trait view of attitudes provides the best explanation for various findings from the attitude measurement literature, which include (1) the finding that scores on different indirect measures often do not correlate well with each other, (2) the finding that scores on indirect measures are susceptible to various context effects, and (3) the finding that scores on indirect measures are poor predictors of behaviour. I

argued that all we need to explain these findings is the assumption that different indirect measures often tap into different mental states that determine evaluative responses. Machery's trait view is compatible with this assumption but so is my profile view. This is because both models assume that attitudes have a psychological basis which is composed of a variety of mental states. Different attitude measures may tap into different subsets of these mental states. Accordingly, people's results on these measures will often only be weakly correlated, be influenced by different kinds of contextual factors, and only be predictive of a narrow range of responses (see section 4.3). For instance, the performance of an ambivalent sexist person on an IAT may primarily be influenced by benevolent stereotypes that the person harbours in regard to women, such as associations between women and kindness or sociability. At the same time, the person's performance on an affective priming task may primarily be influenced by stereotypes about women with negative affective implications (e.g., the stereotype that women are incompetent). Recall that Machery's (2016) denies the existence of ambivalent attitudes. Yet, his model allows for weak attitudes that are based on mental states and processes with somewhat heterogeneous evaluative implications (see section 4.5.3). However, to reiterate, characterising attitudes in terms of an aggregated strength and valence is problematic because it does not do justice to the complexities of people's evaluative stances towards social groups (see section 4.6). Whereas Machery's characterisation of attitudes obscures the fact that people often harbour competing evaluative dispositions in regard to a single social group (that are tied to particular situations), these complexities are a crucial feature of attitudes on my profile view.

4.7.6 Relation to Schwitzgebel's dispositional account of attitudes

My account of attitudes bears resemblance to Schwitzgebel's (2013) dispositional account of attitudes, so it is worth pointing out where I agree and disagree with his characterisation. Schwitzgebel initially proposed a dispositional account of believing (Schwitzgebel, 2001, 2002, 2010) but has more recently extended his dispositional account to attitudes of different kinds, including attitudes towards groups of people (Schwitzgebel, 2013). This is how he characterises having an attitude, including propositional attitudes, reactive attitudes, but also attitudes towards groups of people:

> To have an attitude is, primarily, to have a dispositional profile that matches, to an appropriate degree and in appropriate respects, a stereotype for that attitude, typically grounded in folk psychology. (Schwitzgebel, 2013: 78)

I agree with Schwitzgebel insofar as I also view attitudes (whereby I refer only to attitudes towards people qua members of social groups) as dispositional profiles.

Schwitzgebel (2013) characterises a dispositional profile as "a suite of dispositional properties, or more briefly dispositions" (p. 78). He acknowledges that the dispositions that form part of the dispositional profiles may include amongst others cognitive, affective, and behavioural dispositions (pp. 87-88), and that these dispositions are very often tied to particular situations (e.g., pp. 85-87). All this is in line with my description of attitudes as profiles of situation-specific response dispositions (although I put greater emphasis on the situation-specific character of these dispositions than Schwitzgebel does).[141]

Yet, I disagree with Schwitzgebel's suggestion that we can only say that someone possesses an attitude if she possesses a dispositional profile that matches a (folk psychological) stereotype for that attitude. Consider again the case of the aversive racist (referred to as the "implicit racist" by Schwitzgebel). Schwitzgebel (2013) comes to the conclusion that this is a case of in-between attitude possession (pp. 85-87; see also Schwitzgebel, 2010). According to him, the aversive racist only partly exhibits the dispositional profile that ordinary people would commonly regard as characteristic of an egalitarian attitude and only partly the dispositional profile that is commonly regarded as stereotypic of a racist attitude. In short, the aversive racist neither fully possesses a racist attitude, nor does she fully possess an egalitarian attitude. What Schwitzgebel apparently does not want to say is that the person in question has an aversive racist attitude. That is, he does not want to say that the person shows the dispositional profile that is commonly regarded as characteristic of an aversive racist attitude. This is because there is presumably no folk psychological stereotype of an aversive racist attitude. Note that for him "[to] have an attitude […] is mainly a matter of being apt to interact with the world in patterns that ordinary people would regard as characteristic of having that attitude" (p. 75). Ordinary people have a conception of what it is to be a racist or to be an egalitarian but possibly not of what it is to be an aversive racist.

I find it problematic to tie the possession of an attitude to stereotypes for that attitude that are grounded in folk psychology. This may seem surprising because I argued previously that it is an advantage of a trait view of attitudes that it corresponds well with the folk psychological conception of attitudes (see section 4.4). However, it is important to note that saying that it is beneficial to adopt a view of the *ontology of attitudes* that corresponds to the folk conception of attitudes is not to say that *the content of these attitudes* must be restricted by folk psychology. A model of attitudes that psychologists, philosophers, and ordinary people can agree on (see desideratum *D3* of a model of attitudes in the introduction to this thesis) is not necessarily a model

[141] Schwitzgebel (2013) also emphasises the structural similarities between traits and attitudes (pp. 81-82). Yet, he stops short of saying that attitudes *are* traits. This does not necessarily contradict my claim that attitudes are traits because I have a narrower notion of attitudes in mind (i.e., attitudes towards people qua members of social groups) than Schwitzgebel.

that only acknowledges the existence of attitudes for which there are already folk psychological stereotypes. Note that even if there is no (folk psychological) stereotype of an aversive racist attitude, someone may notice that Sarah tends to show favourable responses towards black people in situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments and negative responses towards black people in all other situations. Irrespective of whether we have a label such as "aversive racism" for this profile that Sarah exhibits, it is a fact that Sarah exhibits this profile. It seems natural to me that someone who discovers this pattern in Sarah's responses towards black people has in fact discovered an attitude of her even if there is (yet) no stereotype for such an attitude in the community of folk psychologists (or in a more narrowly specified community).

It must be emphasised that the notion of an attitude is at least as much a scientific notion as it is a folk psychological one and that science can inform folk psychology. Even if folk psychologists are not (yet) familiar with the notion of an aversive racist attitude, it makes sense for psychologists (and philosophers) to ascribe this and other profile attitudes to people. As mentioned earlier, describing people in terms of profile attitudes can help to identify common patterns of evaluative responding and help making predictions about likely responses of an agent. Notions that scientists use may in turn be taken up in ordinary parlance. Schwitzgebel (2013) himself notes this in passing:

> [S]cience can legitimately lead us to adjust our superficial stereotypes, either by producing entirely new stereotypes or by modifying existing stereotypical structures to incorporate rising knowledge. Psychological research on sexism, for example, can coin a new type – "the implicit sexist" – and also modify our existing stereotypes of sexism and egalitarianism *simpliciter*. Folk psychological stereotypes won't sit still, anyway, and are always to some extent influenced by scholarship and science, hence "phlegmatic", "extravert", "agnostic", and our post-Freudian sense of how desires might manifest. (pp. 94-95)

While Schwitzgebel's point is a descriptive one (science may eventually change our folk psychological stereotypes about attitudes), my stance is a more decisively normative, or one could say ameliorative, one.[142] Scholarship of the type that I am pursuing in this thesis can tell us how folk psychological stereotypes about attitudes *should* be changed given that attitude ascriptions are linked to certain functions such as response prediction and character assessment. In other words, science and scholarship can inform us how we should speak about attitudes in order for attitude ascriptions to optimally fulfil these purposes. It may well be that our current folk

[142] Talking about social kind concepts Haslanger (2005) explains that "[a]meliorative analyses elucidate 'our' legitimate purposes and what concept of F-ness (if any) would serve them best (the target concept)" (p. 20). See also Haslanger (2000) on what she there calls the "analytical approach" (p. 33).

psychology only contains stereotypes of generic positive or negative attitudes but not stereotypes of more complex attitudes, such as aversive racist or ambivalent sexist attitudes. Yet, my claim is that we should integrate complex attitudes, such as aversive racist or ambivalent sexist attitudes, into our folk psychological attribution repertoire if science shows that these are prevalent profiles that people exhibit and if ascribing these profiles serves our purposes well. In particular, including more complex attitudes in our folk psychological attribution repertoire will help us to convey more accurate information about people's evaluative stances towards other people, to make better predictions of people's responses towards other people, and to reach more appropriate conclusions about the attitude holder's moral character. As a consequence, the model of attitudes that I have proposed in this chapter is a model that ordinary people may find appealing, even though it allows for the existence of attitudes for which there are not yet folk psychological stereotypes. At the same time, it is a model that scholars in psychology and in philosophy may find appealing because it is a model that is not unduly restricted by folk psychology. In short, it is a model that psychologists, philosophers, and ordinary people can possibly agree on (see desideratum *D3* of a model of attitudes).

4.8 Conclusion

The similarity between attitude ascriptions and trait ascriptions are undeniable. Both can be seen as parts of what Goldie (2004) calls "personality discourse":

> Personality discourse is everywhere largely because it serves a purpose, or rather, because it serves several purposes. We use personality discourse to describe people, to judge them, to enable us to predict what they will think, feel and do, and to enable us to explain their thoughts, feelings and actions. (pp. 3-4)

Ascribing a trait such as compassion to a person provides a description of the person on the basis of which we may judge the person (e.g., assessing the person's moral character) and predict or explain the person's cognitive, affective, and behavioural responses in relevant situations. We can say exactly the same about the ascription of an attitude to a person (see section 4.4). We describe people in terms of attitudes in order to predict and explain their cognitive, affective, and behavioural responses towards other people (see function *F1* in the introduction to this thesis) and to assess their character (see function *F2*). It is evident that folk psychologists, at least, conceive of attitudes as traits of persons. As I showed in this chapter, there is a model of attitudes available that captures the notion that attitudes are traits and that may also appeal to scholars in psychology and philosophy (see desideratum *D3* of a model of attitudes).

I started out by discussing the trait model of attitudes that has recently been proposed by Machery (2016). According to this model, "[attitudes] are broad-track dispositions to behave and cognize […] toward an object […] in a way that reflects some preference" (Machery, 2016: 112). These broad-track dispositions are based on a variety of mental states and processes, which is why Machery's trait model can account for some perplexing results from the attitude literature (though it is certainly not the only model that can account for these findings; see section 4.3). Also, Machery can parry the situationist objection against the notion that attitudes are traits by characterising these attitudes in terms of an aggregate strength and valence (see section 4.5). However, this characterisation comes at a price: it obscures evaluative conflicts and ambivalences, and masks relevant differences in people's affective responses (see section 4.6).

I argued that these complexities are better addressed by characterising attitudes as profiles of situation-specific response dispositions (see section 4.7). For example, Sarah's attitude towards black people can broadly be analysed as the profile to show favourable responses towards black people in situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments and to show negative responses towards black people in situations in which she does not have sufficient time or cognitive resources to reflect on and be guided by her endorsed egalitarian commitments. It is misguided to assume, as situationists do, that it speaks against the existence of attitudes understood as traits if people exhibit different evaluative responses towards members of a particular social group in different situations (see section 4.5.2). Quite to the contrary, I have argued that it is a defining feature of attitudes that they are composed of dispositions that are tied to particular situations. Accordingly, the proposed profile view of attitudes both does justice to the evaluative complexities of attitudes and neutralises the situationist challenge against the notion of attitudes construed as traits.

The assumption that attitudes are traits provides an answer to the question about the ontological status of attitudes (see question *Q3* in the introduction to this thesis). These traits are based on a variety of distinct kinds of mental states, which may include conceptual associations, affects, beliefs, and desires. This provides an answer to the question about the mental states that underpin attitudes (see question *Q2*). Lastly, I have dealt with the question about attitude individuation in this chapter (see question *Q1*). In particular, I argued that we may individuate attitudes in different ways dependent on our interests and purposes as attitude ascribers (see section 4.7.3). Saying that attitude individuation is interest dependent is of course not meant to imply that that there are no constraints on attitude individuation. We cannot just ascribe any attitude to a given person. Of course, attitude ascriptions are only accurate as long as

they correspond to actual dispositions of the agent. However, it is an epistemic requirement that we need to extract especially salient or relevant patterns of evaluative response dispositions if we want to give an intelligible account of people's attitude(s). It is this process of highlighting relevant response patterns (i.e., highlighting profiles of situation-specific response dispositions) that is influenced by the attitude ascriber's interests and purposes. Dependent on their interests and purposes, attitude ascribers may highlight different situational contrasts and different kinds of responses (section 4.7.3.1), they may vary how finely they differentiate between different situations and different responses (section 4.7.3.2), and they may vary the scope of their attribution along the local-global dimension (section 4.7.3.3).

Ascribing attitudes understood as profiles of situation-specific response dispositions to people fulfils an explanatory and predictive function (see function *F1* of the attitude notion in the introduction to this thesis). For example, knowing that Jack has an ambivalent sexist attitude helps us to make specific predictions about Jack's responses towards women in different roles and in different contexts (and to explain these responses retrospectively). If Jack is confronted with a female supervisor at work, we may predict that he will feel uncomfortable and will behave in a negative way towards her. By the same token, we can be reasonably certain that Jack will behave in a friendly manner towards female assistants. If we would insist, as Machery (2016) does, that attitudes have either a positive or a negative valence and that ambivalent attitudes are thus impossible (p. 124), we would not be able to use the notion of an attitude to make such predictions. In contrast to Machery's model, the here proposed profile model of attitudes allows us to pick out exactly those features of an agent's psychology that drive that person's evaluative responses towards the target group (see section 4.7.4). The profile view thus fulfils desideratum *D1* of a model of attitudes as it was mentioned in the introduction to this thesis.

I also insist that the profile notion of attitudes is well-suited to contribute to the moral assessment of a person's character (see function *F2* of the attitude concept). Note that I have described profile attitudes as being rooted in a person's cognitive-affective personality system. By learning, for example, about Sarah's aversive racist attitude, we learn something about Sarah's personality structure, which we can then consider in our moral evaluation of her. By contrast, when describing attitudes in terms of aggregate strength and valence, as proposed by Machery (2016), we are likely to miss important aspects of what would be relevant for character assessment. Following Machery's model we may, for example, end up with the perplexing conclusion that Sarah lacks an attitude towards black people (see section 4.6).

However, there remains one possible caveat. As I will show in the next chapter, one may want to argue that my model of attitudes leads to a wrong assessment of

Sarah's character because it implies that Sarah's attitude towards black people is partly based on those mental states that she does not identify with (e.g., her association between BLACK PERSON and DANGER).[143] According to this objection, my model does not fulfil desideratum *D2* of a model of attitudes: it is not appropriately sensitive to the difference between aspects of a person's psychology that can rightly be said to be constitutive of that person's moral character and those aspects that are not part of that person's moral character.

[143] Note that if this argument was successful, this would be equally damaging to Machery's (2016) account of attitudes as to my account because both models imply that mental states that a person does not identify with can form part of that person's attitude.

Chapter 5: Attitudes and character evaluation

5.1 Introduction

In the last chapter, I have presented my positive account of the nature of attitudes. According to this, attitudes are collections of response dispositions that form characteristic profiles (hence, "the profile view of attitudes"). I mentioned that an aversive racist attitude can be analysed as a profile that consists of broadly two situation-specific evaluative response dispositions: the disposition to show favourable responses concerning black people in situations in which the agent has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments and the disposition to show negative responses concerning black people in situations in which the agent does not have sufficient time or cognitive resources to reflect on and be guided by her endorsed egalitarian commitments. It is worth emphasising again that I take attitude ascriptions to fulfil a character evaluative role. When we learn about someone's attitude, we learn something about that person's character that we can take into account in our moral evaluation of the person (see function *F2* of the attitude notion as mentioned in the introduction to this thesis). Accordingly, saying that Sarah has an aversive racist attitude suggests that we can evaluate her for both her positive and her negative response dispositions in regard to black people.

Proponents of so-called "real self theories" (also sometimes called "deep self theories") may find this implication of my account problematic (e.g., Frankfurt, 1971; Stump, 1988; Velleman, 1992; Watson, 1975). On these models, only those dispositions that the agent identifies with or that conform to the agent's considered values and rational judgments constitute the persons "real self" for which she is morally evaluable. All other dispositions are external to what constitutes the morally evaluable self of the person. On the real self model, Sarah's disposition to respond in negative ways towards black people (e.g., when she is under time pressure) does not count as part of her real self because it conflicts with her endorsed egalitarian values and reasoned judgments. If we would include this disposition in our moral evaluation of Sarah's character, we would make a mistake according to this line of reasoning because this disposition is not part of what she really stands for. As attitude ascriptions are commonly seen as guides to character evaluation, my account of attitudes that describes non-endorsed dispositions as parts of attitudes would according to this argument invite misguided judgments about the moral characters of attitude holders. In

other words, my account of attitudes may fail to satisfy the second desideratum of a model of attitudes as it has been stated in the introduction to this thesis:

(D2) To optimally fulfil its role in character assessment, our notion of a person's attitude towards group X should be sensitive to any difference that there may be between aspects of that person's psychology that can rightly be said to be constitutive of that person's moral character and those aspects that are not part of that person's moral character.

Note that Sarah's disposition to show negative responses towards black people when she does not have sufficient time or cognitive resources to reflect on her endorsed egalitarian commitments is presumably (mostly) grounded in mental states that we would commonly regard as implicit (e.g., conceptual associations and affects linked to these). In chapter 2, I have already suggested that those mental states commonly described as implicit can be said to form part of a person's moral character because they are not in fact completely outside of the agent's control. They are subject to forms of ecological control (indirect rational control and indirect intentional control; see section 2.3). Yet still, one may object that the fact that Sarah does not identify with those mental states that imply a negative evaluation of black people (and thus with her disposition to show negative responses towards black people) suffices to establish that these mental states do not form part of her moral character, irrespective of whether Sarah could in principle take ecological control of these mental states. Accordingly, my account of attitudes would be at odds with desideratum *D2* after all. In this chapter, I will refute this position and argue in support of the view that even those evaluative mental states (and the resulting dispositions) that an agent does not identify with can form part of the agent's moral character. This supports my account of attitudes, according to which attitudes may be composed of both endorsed and non-endorsed response dispositions. I will point out that this account is not only in line with relevant pre-theoretical intuitions but also commendable for pragmatic reasons.

I will proceed as follows. In section 5.2, I will present the real self view and elaborate on its possible implication that evaluative response dispositions that an agent does not identify with do not form part of that agent's attitudes. In section 5.3, I will stress that if the real self perspective should be right, we need to give up on the common conception that attitude ascriptions can fulfil at the same time an explanatory/predictive and a character evaluative function. In section 5.4, I will show by reference to the case of Huckleberry Finn (which plays a prominent role in the philosophical moral psychology literature) that we routinely take non-endorsed response dispositions into account when evaluating other persons' characters, which is

at odds with real self theory. In section 5.5, I will stress that many of us also take non-endorsed response dispositions into account when evaluating our own character. I will further argue that when people judge that their problematic non-endorsed response dispositions are not part of their real self, this is likely the result of a self-serving bias rather than an honest assessment. In section 5.6, I will build on this and show that there is also a pragmatic reason for including non-endorsed response dispositions in our conception of attitudes: it nudges people to tackle their problematic biases that are harmful to others.

5.2 Real self and attitudes

Real self theories hold that there is a distinction to be made between mental states and processes that profoundly belong to an agent (thus constituting the "real self" of the agent) and those mental states and processes that, albeit operating within the agent, cannot be attributed to the agent.[144] Many contemporary real self models are influenced by or take as reference point Harry Frankfurt's (1971; 1988) model of the structure of volition. Frankfurt proposes a hierarchical model of volition according to which a person has first-order desires to perform one or another action and second-order volitions concerning what first-order desires she wants to be effective in action.[145] He illustrates this model by reference to the example of an unwilling drug addict (Frankfurt, 1971: 12-13). Frankfurt describes this person as having two first-order desires that stand in conflict with each other: based on his substance dependence, he has a desire to consume the drug, but at the same time he also has a desire to refrain from taking the drug (a desire that may, for example, be driven by health concerns). The unwilling drug addict wants the latter desire to gain the upper hand over his behaviour. That is, he has the second-order volition that his first-order desire to refrain from taking the drug may become effective in action. According to Frankfurt, through this kind of second-order endorsement the unwilling drug addict makes the desire to refrain from taking the drug "more truly his own" (Frankfurt, 1971: 13). To be sure, his behaviour may still be driven by the first-order desire to take the drug, but as the person does not identify with this particular desire, it constitutes "a force other than his own" (Frankfurt, ibid). One can say, although these are not Frankfurt's own words, that the desire to take the drug does not reflect his "real self" (Arpaly & Schroeder, 1999: 165).

[144] See Arpaly and Schroeder (1999), Lippert-Rasmussen (2003), or Sripada (2016) for reviews of real self accounts.

[145] Frankfurt (1971) attaches a special meaning to the term "person". According to him, an agent is only a person if she is capable of having second-order volitions. I will not follow this distinction and instead use the terms "agent" and "person" interchangeably in what follows.

The details of Frankfurt's hierarchical model of volition have been subject to wide criticism and a range of modified and alternative models have been suggested to account for the assumed difference between mental states that form part of a person's real self and those that are external to it.[146] For example, it has been suggested that the real self is grounded in second-order volitions that are the result of reasoning (Stump, 1988), a person's "valuational system" (Watson, 1975: 215), or an agent's "desire to act in accordance with reason" (Velleman, 1992: 479).[147] Despite these differences in real self accounts, these views have in common that they divide the self in two parts that can broadly be described as "Reason" and "Appetite" as is aptly described by Arpaly & Schroeder (1999):

> [T]hese theories all share the assumption that there is a sharp separation in the structure of the self which roughly follows the Platonic distinction between reason and the appetites, and they all identify the agent's Real Self with the part of her self that is the counterpart of reason in the Platonic model. If there is a clash between what a person desires or prefers to do and what she decisively thinks she should desire or prefer, these theorists identify the person's Real Self with her conviction regarding what she should desire rather than with the conflicting desire. Whether they will interpret the situation as a clash between a desire and a second-order volition, a desire and a value, a desire and a reasoned second-order volition, or a desire and the desire to be rational, they will all identify the agent with the conviction or judgment rather than with the conflicting desire, with the equivalent of Platonic reason (hereafter "Reason") rather than the equivalent of Platonic appetite (hereafter "Appetite"). (p. 170)

Although Arpaly & Schroeder (1999) focus on desires here, it should be noted that not only desires but also other kinds of mental states, such as emotions, associations, or propositional mental states, can clash with what they describe as "Reason".[148] Note for example, that one may be afraid of black people, associate black people with violence, and think that black people are dangerous, although one may have rationally formed higher-order desires not to be afraid of black people, not to associate black people with violence, and not to think that black people are dangerous. Accordingly, one may want to say that these mental states are not part of the person's real self.[149] When I speak of

[146] Frankfurt himself has modified his original account as presented in his 1971 paper to some extent. See for example Frankfurt (1988).
[147] See Arpaly & Schroeder (1999: 165-166) for a short review of these real self models.
[148] Presumably, real self views have focused on desires because cases in which an action follows from a desire that the agent is estranged by are seen as paradigm cases in which an agent's autonomy or freedom is compromised.
[149] The question whether there can be beliefs that are not part of a person's real self is trickier because the answer depends on one's stance on belief fixation (Mandelbaum, 2014; see also section 2.2.1.2 on this). On what can be seen as the standard view of belief fixation, the Cartesian view, people have the ability to entertain propositions in their mind without assenting to the proposition. In a separate step they may endorse or deny the proposition. By virtue of endorsing the proposition they form a belief with the respective proposition as content. On this common view, it is incoherent to say that one believes a certain proposition but that one does not endorse the belief. By contrast, on what has been described as the Spinozan view of belief fixation, people automatically believe those propositions that they entertain (Huebner, 2009, 2016, Mandelbaum, 2014). On Mandelbaum's (2014) interpretation of this account, the automatic acceptance of a proposition is a subpersonal process that can be followed by a

"real self theory" in what follows, I refer to those models that identify the real self broadly with what Arpaly and Schroeder (1999) describe as "Reason". I take this to be the predominant type of model of the real self in the philosophical literature. This being said, there are alternative models of the real self (e.g., Sripada, 2016), but these will not be the target of my argument in this chapter.

The distinction between "Reason" and "Appetite" is reminiscent of the distinction between reason-responsive explicit and reason-insensitive implicit attitudes as reviewed in chapter 1 (see especially section 1.2.3). Recall that I have argued in chapter 2 that implicit mental states are not completely reason-insensitive. In fact, they are at least indirectly reason-responsive (i.e., they are subject to indirect rational control) and can thus potentially be brought into line with the agent's commitments (see section 2.3.1). Yet still, according to real-self theory it does not suffice that a mental state can potentially be brought into line with "Reason" (i.e., the agent's second-order volitions, values, the desire to be rational, etc.) for that mental state to be part of that person's real self. On this view, only those mental states that *de facto* align with "Reason" (i.e., the agent's second-order volitions, values, the desire to be rational, etc.) can be regarded as part of the agent's real self.

In what follows, I will say that a person identifies with a mental state or endorses a mental state if that mental state is in accordance with that person's (reasoned) second-order volitions, values, and the desire to be rational (i.e., when it is in line with "Reason"). Conversely, I will say that a person does not identify with or does not endorse a mental state ("non-identification") if that mental state conflicts with that person's (reasoned) second-order volitions, values, and the desire to be rational (i.e., when it conflicts with "Reason").[150, 151] Of course, different real self accounts will

---

reflective endorsement or rejection of the belief on person-level. This view thus allows for the existence of beliefs that the person does not endorse or would indeed reject on reflection and which accordingly may be said to be external to the real self of the person. This is not the place to adjudicate between these different accounts of belief fixation. For my present purposes, it suffices to point out that both views agree that a person can reject the truth of mentally entertained propositions (be they beliefs or not).

[150] Sometimes people may neither identify nor disidentify with a mental state (Lippert-Rasussen, 2003: 371-373). For example, a person may not possess any values that would conflict with or conform to a particular desire. We may treat this as a border-line case: it is neither a prototypical case of a mental state that forms part of the real self of the person, nor is it a prototypical case of a mental state that is external to the real self of the person. In what follows, when I speak about "non-identification" I mean to refer to clear cases of a mental states being at odds with (reasoned) second-order volitions, values, and the desire to be rational and when I speak of identification, I have prototypical cases of endorsement in mind, in which the agent is not just indifferent in regard to the mental state.

[151] Instead of saying that a person does not identify with a mental state, one may choose to say that the person is alienated from the mental state (Glasgow, 2016). However, this terminology does not quite capture what I mean by non-identification. Alienation is usually understood as a negative feeling state and not simply as a relation between a mental state and the agent. Note that Sarah, for example, need not be aware of a mental state (e.g., her association between BLACK PERSON and VIOLENCE) to "non-identify" with the mental state in my proposed sense of the term "non-identification". All that non-identification implies is that a mental state is in

sometimes imply different verdicts about whether a person identifies with a given mental state and, consequently, about whether a given mental state belongs to the real self of a person (dependent on whether they focus on second-order volitions, values, the desire to be rational, etc.). Yet, for my present purpose, which is to outline what the real self view may imply for our understanding of a person's attitude, it suffices to attend to clear cases in which different real self accounts of the kinds mentioned above would lead to the same conclusion.

Let us again consider the case of Sarah the aversive racist. Sarah tends to exhibit favourable (cognitive, affective, and behavioural) responses towards black people when she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments and tends to show negative responses with respect to black people when she does not have sufficient time or cognitive resources to reflect on and be guided by her endorsed egalitarian commitments. Sarah's former disposition (her disposition to show favourable responses towards black people) is presumably grounded in her endorsed egalitarian values, her belief that it is morally reprehensible to treat people differently because of their skin colour, and her higher-order desire not to discriminate against black people. Let us also assume that Sarah's disposition to show negative responses towards black people is grounded, amongst others, in stereotypes that link black people to negative attributes such as danger or violence, an emotional disposition to feel afraid of black people, and perhaps a barely conscious desire to keep distance to black people (in short, the evaluative stereotype "black people are dangerous"; see section 3.4.1). We can say that Sarah does not identify with these mental states (and thus with her disposition to show negative responses towards black people in certain situations) because they are in conflict with her endorsed egalitarian values, her conviction that it is wrong to treat people differently because of their skin colour, and her higher-order desire not to discriminate against black people. Hence, following the real self view, those mental states on which Sarah's situation-specific disposition to show negative responses in regard to black people is based are not part of who Sarah really is. Sarah's real self is rather to be identified with her endorsed values, her higher-order desires, or considered judgments, which ground her situation-specific disposition to show positive responses towards black people. On this account, it would be unfair to base our moral evaluation of Sarah (even partly) on her situation-specific disposition to show negative responses towards black people. As attitude ascriptions are commonly taken as a guide to moral character evaluation, it would accordingly be problematic to include Sarah's problematic response disposition

---

conflict with the person's values, second-order volitions, the desire to be rational, etc. (i.e., with "Reason"). Yet, there is a relation between non-identification and feelings of alienation: when Sarah becomes aware of the mental state that she does not identify with, she would likely experience feelings of alienation.

towards black people in our notion of her attitude towards black people. Rather than saying that she has an aversive racist attitude towards black people as proposed on my model of attitudes, we may want to say that she has a favourable attitude towards black people.[152] This would provide us with a more accurate description of Sarah's moral character according to the real self view.[153]

5.3 Prediction and character evaluation revisited

It should be noted that saying that Sarah has a favourable attitude towards black people is not particularly helpful when it comes to the prediction of Sarah's responses towards black people. Some of Sarah's responses towards black people will be driven by those mental states that she does not identify with (e.g., her association between BLACK PERSON and VIOLENCE). Thus, if we are told that Sarah has a favourable attitude towards black people, we may form wrong expectations about her likely responses towards black people. Saying that Sarah has a favourable attitude towards black people directs our attention to her endorsed egalitarian values, her conviction that it is wrong to treat people differently because of their skin colour, and her higher-order desire not to discriminate against black people. On the basis of this attribution we would expect her to respond exclusively in a neutral or favourable manner towards black people. Note that we are more likely to form correct predictions about her responses when we are told that she has an aversive racist attitude. That is, we will form more accurate predictions when our attention is drawn to the fact that she also harbours an emotional disposition to feel afraid of black people, stereotypes that link black people to negative attributes such as violence, etc. Yet, if we include these non-endorsed mental states in our notion of Sarah's attitude, we include mental states that are, according to real self theory, not part of Sarah's moral character. As a result, if the

[152] Instead of saying that Sarah has an aversive racist attitude, my profile view of attitudes also licences to say that Sarah has a favourable attitude towards black people that is tied to situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian values, and a negative attitude towards black people that is tied to situations in which she does not have sufficient time or cognitive resources to reflect on her egalitarian commitments (see section 4.7.3.3). From a real self perspective, this way of speaking about Sarah's attitudes is just as problematic as saying that she has an aversive racist attitude. Sarah only identifies with her disposition to show favourable responses towards black people and thus only this disposition should be referred to as her attitude, given that attitude ascriptions are a guide to moral character evaluation.

[153] It shall be stressed that even if the real self perspective is correct, it may still be legitimate to ascribe ambivalent sexist attitudes to people. Note that the ambivalent sexist may identify both with those mental states that ground his disposition to show hostile responses towards women in superior roles (e.g., an association between WOMAN and INCOMPETENCE) and with those mental states that ground his disposition to show benevolent responses towards women in same status or subordinate roles (e.g., an association between WOMAN and SUPPORT). That is, even on the real self view, both of these dispositions may be regarded as components of the person's moral character.

real self view is right, attitude ascriptions cannot adequately fulfil a predictive and a character evaluative function at the same time. On this view, we could either say that Sarah has a favourable attitude towards black people and thus fulfil the character evaluative function but give up on the predictive function, or we could say that Sarah has an aversive racist attitude towards black people and thus fulfil the predictive function but give up on the character evaluative function.

As I have mentioned in the introduction to this thesis, we (as folk psychologists) normally take an attitude ascription to be informative about a person's likely cognitive, affective, and behavioural responses towards other people (function *F1*) *as well as* about the person's moral character (function *F2*). Is this common conception thus misguided? In what follows, I will argue that it is not. I will show that the real-self view is not unequivocally supported by our intuitions and that there are good reasons not to take those intuitions that speak in favour of the real self view at face value. When we evaluate the characters of other people, we tend to base our judgment on both endorsed and non-endorsed evaluative response dispositions (section 5.4). We often do the same when we evaluate our own character, and when we do not, this is likely due to a self-serving bias rather than an honest assessment (section 5.5). Building on this, I will stress that including both endorsed and non-endorsed evaluative response disposition into our conception of attitudes has pragmatic benefits: when we acknowledge that non-endorsed evaluative response dispositions are part of our attitudes, we are more likely to tackle problematic biases than when we regard them as aspects of our psychology for which we are not morally evaluable (section 5.6).

5.4 Third-person moral character assessment

That we intuitively take into account non-endorsed response dispositions of an agent when evaluating the agent's character can be shown by reference to the case of Mark Twain's (1884) fictional character Huckleberry Finn, which plays a prominent role in the moral psychology literature (e.g., Arpaly, 2002; Arpaly & Schroeder, 1999; Bennet, 1974; McIntyre, 1993; Smith, 2004). On the one hand, Huckleberry Finn (henceforth "Huck") identifies with the racist principles of the society that he grew up in. On the other hand, he feels deep compassion towards his slave friend Jim, which leads him to help Jim escape from his owner Miss Watson. On reflection, Huck comes to the conclusion that it was wrong to help Jim escape and feels deep regret about his deed. There is no doubt in his mind that Miss Watson is the rightful owner of Jim. On their journey on the Mississippi, Huck is on the verge of turning Jim in to the authorities. Yet, ultimately his non-endorsed sympathy for Jim wins over his conviction. When he is asked by bounty hunters whether the other man on his raft is black or white, he lies to

them. He fails to do what he believes to be the right thing to do and considers himself weak for this. In a way, Huck represents the reversal of an aversive racist. The aversive racist endorses non-racist ideals, yet frequently behaves in a racist manner. Huck, by contrast, endorses racist ideals, yet tends to behave in a non-racist manner towards Jim and would probably behave in a similar manner towards other black people. Accordingly, we may say that Huck is an aversive egalitarian.

On a real self account as specified in section 5.2, we would have to say that Huck's sympathy towards Jim is not part of his real self and thus not part of his moral character (Arpaly & Schroeder, 1999). Huck does not want to be moved by his sympathy towards Jim (his sympathy is at odds with his second-order desires), he believes that helping Jim is immoral (helping Jim is in conflict with his values), and his reasoned judgment is that he should turn Jim in (Huck's sympathy towards Jim is at odds with his desire to be rational). On this view, Huck is a racist. Yet, this verdict does not match our intuitions as readers of Twain's (1884) novel. It seems to us that Huck is at the core of his heart a good person, who is somewhat led astray by the ideals that the racist society that he lives in has imposed on him (Arpaly & Schroeder, 1999; Smith, 2004). Our intuition is presumably guided by the fact that Huck's disposition to feel sympathy towards Jim is very deep. This sympathy is expressed in much of his affective, cognitive, and behavioural responses towards Jim. His racist ideals manifest themselves in the guilt that he feels about helping Jim, but they do not actually motivate him to action. The fact that Huck's sympathy towards Jim persists despite his questionable moral convictions underlines the deepness of these feelings (Smith, 2004: 343). This being said, we would appreciate Huck's moral character even more if he did not buy into racist ideals at all (other things being equal). Just as Huck does not seem to us to be a pure racist, he does not seem to us to be a pure egalitarian either. In short, we intuitively take into account both Huck's non-endorsed response dispositions and his endorsed values or reasoned judgments when evaluating his moral character. That is, our intuitions are at odds with real self theory when it comes to the assessment of Huck's character.[154] Note that if we did not know about Huck and were told by someone that Huck exhibits a racist attitude, we would likely form an impression of Huck that would diverge from the impression that we would form of him as readers of Twain's Novel. We would form the impression of a person whose evaluative

[154] It shall be emphasised again that what I label as "real self theory" encompasses models that identify the real self roughly with what Arpaly and Schroeder (1999) have described as "Reason" (see section 5.2). Hence, to be precise, my argument is that our intuitions about Huck's case are at odds with the predominant model of the real self as I have specified it in section 5.2. This being said, one may of course choose to defend a different type of model of the real self (or deep self) on which Huck's sympathy towards Jim turns out to be reflective of his real self. In fact, Arpaly and Schroeder (1999) defend a model (which they call "Whole Self theory") that captures the intuition that Huck's non-endorsed dispositions reflect on his character due to their "deepness" (see also Sripada, 2016).

dispositions in regard to black people are unequivocally negative. If we were instead told that Huck exhibits an aversive egalitarian attitude, we would form a much more appropriate impression of Huck: we would form the impression of a person who shows egalitarian dispositions but who does not identify with these.[155]

Note that if it is appropriate to say that Huck has an aversive egalitarian attitude, it should also be appropriate to say that Sarah has an aversive racist attitude. After all, the two cases are structurally similar except for the fact that Sarah shows racist dispositions that she does not endorse, while Huck shows egalitarian dispositions that he does not endorse. If we are happy to include Huck's non-endorsed disposition to show favourable responses towards Jim in our evaluation of his character, we should also be happy to include Sarah's non-endorsed disposition to show negatively valenced responses towards black people in our assessment of her character.

What I have presented in this section may not provide a knockdown argument against the real-self view as presented in section 5.2. For my present purposes, it suffices to have shed some doubt on the intuitiveness of the claim that only response dispositions that the agent identifies with reflect on the moral character of the agent. I will continue with this programme in the next section, where I will emphasise that not only in third-person but also in first-person character evaluations, we often take non-endorsed response dispositions into account. Together the evidence from third-person and first-person character evaluation indicates that including non-endorsed evaluative response dispositions in our attitude model is in line with common intuitions about the scope of people's morally evaluable self.

5.5 First-person moral character assessment

If real self theory as outlined in section 5.2 was right, we should base the evaluation of our own moral character only on those cognitive, affective, and behavioural dispositions that we identify with. All other dispositions (i.e., non-endorsed dispositions) do not express our real self. This view is at odds with how many of us react when we become aware of response dispositions that we do not identify with. Smith (2004) correctly notes that we often treat it as deeply revelatory about us and as a call for moral self-improvement when we realise that we respond in ways that conflict with our considered judgments and values (p. 344). Cases in which we learn about ourselves by paying attention to our spontaneous reactions are in fact ubiquitous: by noticing our

[155] Obviously, we would form the most accurate impression of Huck if we were told in detail about his various dispositions. This is arguably what happens when we read Twain's novel. However, in ordinary conversational contexts what we can say is limited. Therefore, we sometimes need to rely on simple labels such as "aversive egalitarian attitude" or "aversive racist attitude" (see also section 4.7.3.2).

surprise about the fact that the CEO of a company is female, we may realise that we associate men more strongly with leadership than women (Valian, 1999); by noting that we mistook an object in a black person's hand for a gun, we may realise that we are in fact disproportionately afraid of black gun violence (Payne, 2001); and by realising that we prefer to keep distance to obese people, we may become aware of the fact that we are disgusted by them (Bessenoff & Sherman, 2000). In all these cases we may genuinely be surprised by our reactions because they are in conflict with our endorsed beliefs and values: we may not endorse the belief that men are better leaders then women, we may find it unjustified to be afraid of black gun violence, we may consider it wrong to react with disgust to obese individuals. Nevertheless, as Smith (2004) notes, we often treat these incidents as revelatory of what kind of person we are rather than as mere lapses that happened to us. This is indicated by the fact that we are often genuinely embarrassed about these incidents (Smith, 2004: 344; Smith, 2005: 264). Arguably, struggling against non-endorsed motives is experienced as being so painful because we in fact wrestle with *ourselves* in these cases and not just with some external influence on our behaviour (Smith, 2004: 339). We are not merely motivated to tackle our non-endorsed dispositions because they are potentially harmful to others. Instead, much of our motivation arguably stems from the fact that these dispositions reflect negatively on us as persons. Certainly, endorsed response dispositions belong to us in a profound way as they reflect sincere commitments of us. Yet, it is not uncommon for us to also ascribe to ourselves – and crucially to take to reflect on ourselves – evaluative response dispositions that we do not endorse.

To be sure, what I have said here is certainly not true of all people all of the time. While some people (at least sometimes) tend to regard it as revelatory about their moral character when they learn about response dispositions that they do not endorse, other people are certainly happy to accept, in line with real self theory, that problematic dispositions that they do not endorse do not reflect on their moral character. For example, some people would certainly deny that it is reflective of their character that they tend to keep excessive distance to black people in conversation because this is at odds with their non-racist commitments. In short, people's intuitions about what belongs to their morally evaluable self differ. However, note that this alone suffices to establish that real self theory, as characterised in section 5.2, is not unanimously supported by our intuitions.

Moreover, there is reason to question the credibility of people's judgments that non-endorsed evaluative response dispositions do not reflect on their moral character. These judgments may well be the result of a self-serving bias rather than an honest assessment (Shepperd, Malone, & Sweeny, 2008). We generally want to see ourselves in a positive light and real-self theory allows for such a positive self-image if only we

endorse the right values. That is, we can boost our self-esteem by declaring problematic response dispositions to be external to our real selves. For example, denying that it is reflective of one's moral character when one tends to keep excessive distance to black people in conversation will help to keep up or establish a positive self-image. However, the fact that the real self perspective helps us to boost our self-esteem is certainly not a good reason to believe that it is true. Being a good person is certainly not only to endorse the right values but also to live up to them (Schwitzgebel, 2010: 546-548; Schwitzgebel, 2013: 87-88). This leads me to my pragmatic argument for the inclusion of non-endorsed response dispositions in our attitude model.

5.6 A pragmatic argument

So far I have argued that when making third-person and first-person moral character assessments, we in fact often take non-endorsed response dispositions into account. When we do not take them into account in first-person moral character evaluations, this may well be due to a self-serving bias rather than an honest assessment. As a consequence, even on the assumption that attitude ascriptions fulfil a character evaluative function (and not only on the assumption that attitude ascriptions fulfil an explanatory/predictive function), it is justifiable to include non-endorsed evaluative response dispositions in our model of attitudes. Now I will turn to a pragmatic argument to further support the claim that we *should* include non-endorsed response dispositions in our model of attitudes (see Schwitzgebel, 2010: 546-548, and Schwitzgebel, 2013: 87-88, for a related argument). In particular, I will argue that it may help to reduce problematic biases if we encourage people to regard non-endorsed evaluative response dispositions as part of their attitudes.

My point is this: if Sarah regards her disposition to show negative responses towards black people as alien to her real self, and thus not to be reflective of her moral character (maybe because of a self-serving bias; see last section), she is less likely to do something against her problematic responses than in the case in which she takes ownership of this response tendency. My argument is based on the well supported assumption that people normally want to see themselves (and want to be seen by others) in a positive light and accordingly tend to be motivated to modify response tendencies that reflect negatively on them in order to avoid feelings of self-dissatisfaction (Devine et al., 1991; Monteith, 1993; Rokeach & Cochrane, 1972; Rokeach & McLellan, 1972). People presumably lack this motivation if they regard problematic response dispositions as being external to their morally evaluable self.[156]

[156] My point is not that people lack *any* motivation to tackle problematic biases if they do not regard them to be reflective of their moral character. They may, for example, still be motivated

As a consequence, if Sarah does not believe that the biases that she exhibits (e.g., keeping excessive distance to black interlocutors) reflect negatively on her moral character, she is less likely to tackle these biases, which in turn makes it likely that these biases will persist. Note that there is the danger of a vicious circle: the tenacious persistence of the bias may reinforce Sarah's impression that she is the passive victim of the bias, which will in turn keep her from tackling the bias and so forth (Smith, 2004: 347).

To effectively reduce problematic biases in people's behaviour, it would thus be beneficial if we could nudge people to regard these biases as reflections of their moral character (if they do not already do so; see last section). As has been mentioned numerous times now, attitude ascriptions are commonly regarded as fulfilling a character evaluative function. By ascribing attitudes to people that include non-endorsed response dispositions, we thus encourage the perception that these dispositions are expressive of people's moral characters. It certainly makes a difference for our self-image whether we (or others) ascribe to us a favourable attitude towards black people, even though we show negatively valenced responses towards black people in a subset of situations, or whether we (or others) ascribe to us an aversive racist attitude. In the former case, the negatively valenced response dispositions are readily ignored. In the latter case, by contrast, it is highlighted that we exhibit these racist dispositions and we are nudged to try to get rid of these because they reflect negatively on us.

However, there remains a caveat. Saul (2013) has remarked that it may create "defensiveness and hostility" if we encourage people who exhibit problematic biases to view themselves as "one of those bad racist or sexist people" (p. 55). It must be stressed that ascribing a negative trait to a person is not yet to say that the person is blameworthy for the trait. As Holroyd and colleagues (2017a) note we "might invoke an evaluative judgement about the agent and her character – she is cruel, or she is racist – without taking a stance on whether this is her fault" (p. 5; see also Watson, 2004). However, Saul's worry may remain. People may *feel* blamed when racist dispositions are ascribed to them and accordingly react with defiance rather than with determination to tackle their biases. In response, it shall be emphasised that when we say that someone is an aversive racist we highlight both desirable characteristics of the agent (the agent's anti-racist commitments) and undesirable characteristics (the agent's non-endorsed racist tendencies). Yet still, one may argue that someone who is called an aversive racist may feel blamed for the biases that he exhibits despite the fact that his

---

to tackle biases because they sympathise with the victims of the bias. My point is merely that the motivation to tackle the bias is likely to be stronger if the bias is perceived as being reflective of one's moral character rather than as external to the morally evaluable self.

egalitarian commitments are also acknowledged. I do not believe that this undermines the here presented pragmatic argument for the inclusion of non-endorsed response disposition in our attitude model. Quite to the contrary, there is evidence that people who feel blamed for their own biases are more likely to form the intention to do something against these biases than people who do not feel blamed (Czopp, Monteith, & Mark, 2006; Scaife et al., 2016). In a study by Scaife and colleagues (2016), participants took part in a shooter task that required them to distinguish quickly between armed and unarmed individuals that appeared on the computer screen in front of them and to "shoot" only the armed individuals (experiment 3; see also Correll et al., 2002, 2007). Subsequently, participants in the experimental group were blamed by the experimenter for a bias against black individuals that they allegedly had exhibited on the task. Participants in the control group, by contrast, were not blamed for their performance on the task. After having completed some other tasks, all participants were then asked the generic question "Do you intend to try to change your future behaviour as a result of your experience in this experiment?" (p. 17). Participants who had been blamed reported on average a stronger intention to change their future behaviour than participants in the control group. Moreover there was a positive correlation between the strength of their intention to change future behaviour and the extent to which they felt blamed and felt guilty, as assessed with a questionnaire. This indicates that blaming people for their biases in fact motivates them (at least when they are egalitarian minded as most participants in the experiment were) to tackle their biases rather than producing resistance. Accordingly, it may in fact be a positive side effect if people feel blamed (or feel guilty) for their biases when they are called an aversive racist. [157]

The here presented pragmatic argument rests on the assumption that people can in fact do something about their troubling non-endorsed evaluative response dispositions. This may seem questionable because these response dispositions (e.g., Sarah's tendency to keep excessive distance to black interlocutors) would not be as problematic as they in fact are if they could readily be modified. However, I would like to reiterate that those mental states on which these response dispositions are based are not completely outside of our control. Recall that I argued in chapter 2 that people have at least some indirect rational control over their implicit mental states (see section 2.3.1). Sarah, for example, may decide to think more often about positive experiences

[157] In the experiment by Scaife and colleagues (2016) no significant effects of the blame intervention on indirect attitude measures that were taken shortly after the intervention could be detected. In two experiments by Czopp and colleagues (2006; experiment 1 and 2), by contrast, people in fact provided less stereotypic responses on a sentence completion task after they had been confronted for their stereotypic responses on a previously completed version of the same task. The extent of behavioural change correlated with the extent to which participants experienced negative self-directed affect such as guilt.

with black individuals to countercondition her mental associations between black individuals and negative attributes such as violence or crime. Moreover, I argued in chapter 2 that people can exert some indirect intentional control over implicit mental states (see section 2.3.2). Sarah could for example form the intention to think the word "safe" whenever she encounters a black person in a deprived neighbourhood to suppress her fear response and the activation of danger or violence related stereotypes (Stewart & Payne, 2008). If non-endorsed evaluative response disposition were not modifiable by the agent, it may seem unfair to encourage people to view them as reflective of their moral character. Yet, as there are strategies that people can adopt to mitigate these response tendencies (ecological control strategies; see section 2.3.3), it is justifiable, and indeed prudent, to foster the conception that these tendencies form part of their character in order to promote positive change. This can be done by ascribing attitudes to people that include non-endorsed response dispositions (such as aversive racist attitudes).

5.7 Conclusion

In this chapter, I have presented and refuted a possible objection against the profile model of attitudes that I introduced in the last chapter. On my proposed account, attitudes are dispositional profiles consisting of various evaluative response dispositions that are tied to particular situations. For example, an aversive racist attitude can be analysed as a profile that consists of two broad response dispositions: (1) the disposition to show favourable responses concerning black people in situations in which the agent has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments, and (2) the disposition to show negative responses concerning black people in situations in which the agent does not have sufficient time or cognitive resources to reflect on and be guided by her endorsed egalitarian commitments. On my view, both of these dispositions are reflective of the moral character of the attitude holder. This is at odds with predominant models of the "real self" in philosophy according to which dispositions that the agent does not identify with cannot be reflective of the person's moral character. On this alternative view, a person who sometimes shows negative responses towards black people (see disposition 2 above) but who endorses egalitarian values and does not want to discriminate against black people (see disposition 1 above) must be said to have a favourable attitude towards black people if we want to convey accurate information about that person's moral character (see section 5.2). As there are evaluative response dispositions that do not form part of the person's attitude on this view (see disposition 2

above), part of what would help us to explain and predict the person's behaviour is excluded from the attitude notion (see section 5.3).

I argued in this chapter, contra the real self view, that non-endorsed response dispositions can and should be seen as being reflective of a person's moral character and thus be included in our model of attitudes. I used the case of Huckleberry Finn to show that it would be deeply counterintuitive if we would base our evaluation of other people's moral character exclusively on their endorsed commitments (see section 5.4). I also showed that it is not uncommon for people to treat it as revelatory about themselves and as a call for moral self-improvement when they realise that they respond in ways that conflict with their considered judgments and values (see section 5.5). Yet, I also conceded that this is not how all people react (all of the time) when they discover response dispositions that they do not identify with. I grant that it may seem intuitive that problematic response disposition that we do not endorse do not reflect on our real self and thus on our moral character. However, we should be careful about taking this intuition at face value. This is because this intuition may be the result of a self-serving bias. The real self view allows us to see ourselves in a positive light despite our problematic response dispositions, if only we endorse the right values, and this is at least part of the appeal of this view. It should be clear, however, that the fact that the real self view helps us to keep up a positive self-image is not a good reason to believe that this view is appropriate.

Quite to the contrary, I regard it as problematic that the real self-perspective helps us to feel good about ourselves despite our problematic response dispositions because this makes us less likely to tackle these dispositions (see section 5.6). We should rather encourage people to view non-endorsed evaluative response dispositions as parts of their morally evaluable self. If they take these dispositions to reflect negatively on them, they are more likely to be motivated to do something against them. Attitude ascriptions play a crucial role in this respect because they are commonly taken to convey information about a person's moral character. By ascribing attitudes that include non-endorsed evaluative response dispositions to people, we encourage the perception that these dispositions are expressive of people's moral characters, and this may motivate people to tackle these dispositions.

In a nutshell, we have good reason to include not only endorsed but also non-endorsed evaluative response dispositions in our model of attitudes. Firstly, the real self view, according to which we should only include endorsed response dispositions in our model of attitudes (on the assumption that attitude ascriptions are to fulfil a character evaluative function), is at odds with how we would intuitively describe the attitudes of other people (see section 5.4). Secondly, the appeal that the real self view may have when we think about our own attitudes results likely from a self-serving bias

(see section 5.5). Thirdly, including non-endorsed evaluative response dispositions in our model of attitudes has pragmatic benefits in so far as it may nudge people to tackle problematic biases (see section 5.6).

To conclude, including both endorsed and non-endorsed response dispositions in our model of attitudes does not violate desideratum *D2* of a model of attitudes as the real self view would suggest. That is, there is good reason not to treat the distinction between endorsed and non-endorsed evaluative response dispositions as a distinction between aspects of a person's psychology that can be said to be constitutive of that person's moral character and aspects of a person's psychology that are not part of that person's moral character. By including both endorsed and non-endorsed response dispositions in our notion of an attitude, attitude ascriptions can fulfil both an explanatory/predictive and a character evaluative function. If we are told, for example, that Sarah has an aversive racist attitude, we can infer that Sarah is likely to behave in negative ways towards black people when she is under time pressure or engaged in an attention demanding task but in a favourable manner when she has time to deliberate. Also, we learn something important about Sarah's character that we can base our moral evaluation of Sarah on. To be sure, how exactly one evaluates the quality of Sarah's character depends on one's values, but I believe that many of us would come to the conclusion that Sarah neither is as morally corrupt as an outright racist nor as laudable as a pure egalitarian. Our moral assessment of her would be more nuanced (see Levy, 2017b, for a related argument).

Conclusion

In the introduction to this thesis, I raised three questions concerning the nature of attitudes, broadly understood as people's evaluative tendencies in regard to social groups:

(Q1) How should we individuate attitudes?
(Q2) What mental states underpin attitudes?
(Q3) What is the ontological status of attitudes?

Also, I have mentioned three desiderata for a model of attitudes:

(D1) To optimally fulfil its explanatory and predictive function, our notion of a person's attitude towards group X must pick out exactly those features of that person's psychology that drive that person's evaluative responses towards group X.
(D2) To optimally fulfil its role in character assessment, our notion of a person's attitude towards group X should be sensitive to any difference that there may be between aspects of that person's psychology that can rightly be said to be constitutive of that person's moral character and those aspects that are not part of that person's moral character.
(D3) To facilitate communication on attitudes between academic disciplines as well as between academia and the wider public, our notion of a person's attitude towards group X should ideally be a notion that psychologists, philosophers, and ordinary people can agree on.

Now it is time to summarise, with an eye on desiderata *D1*, *D2*, and *D3*, what answers I have found to questions *Q1*, *Q2*, and *Q3*.

I. Rejecting the standard view

Let us consider again the case of Sarah the aversive racist as it has been introduced in the introduction to this thesis. Sarah endorses egalitarian values and exhibits deliberate responses in regard to black people that are in line with her anti-racist commitments (e.g., she participates in rallies against the oppression of black people). Yet, she also exhibits spontaneous responses towards black people that are at odds with her anti-

racists commitments (e.g., she keeps above average spatial distance to black interlocutors).

Following an account that is popular in the philosophy and psychology of attitudes, which I have called "the standard view" (see chapter 1), we could say that Sarah's responses are the result of two different classes of attitudes (e.g., Dovidio et al., 1997; Wilson, Lindsey, & Schooler, 2000; Levy, 2014b; see section 1.2). Her deliberate responses in regard to black people are the result of explicit attitudes, such as her belief that racism is morally reprehensible, while her spontaneous responses are the result of implicit attitudes, such her association between BLACK PERSON and DANGER. On this account, attitudes are mental states (answer to *Q3*). In particular, explicit attitudes are typically described as reason-responsive, propositionally structured mental states that contribute to intentionally controlled responses, while implicit attitudes are commonly thought to be reason-insensitive, associative mental states that operate in an automatic manner (answer to *Q2*).[158] As Sarah likely harbours a range of different associative mental states and a range of different propositional mental states in regard to black people, it is implied that she harbours a range of implicit and a range of explicit attitudes in regard to black people (answer to *Q1*). Indirect measures of attitudes, such as the IAT or the affective priming task, are assumed to tap into implicit attitudes, while direct measures of attitudes, such as semantic differentials or feeling thermometers, are supposedly assessing explicit attitudes (see section 1.3.1).

I pointed out that several reasons to distinguish between implicit and explicit attitudes do not hold up to scrutiny and that, accordingly, the standard view is not all that well supported as is often suggested. To start with, the finding that results on indirect and direct measures of attitudes are often dissociated does not establish, as is often implied, that there are two different kinds of attitudes (see section 1.3.2). In fact, this finding is compatible with various different ways to individuate attitudes (including my preferred interpretation that attitudes can be construed as complex traits that are based on various mental states).

Moreover, there is evidence that sheds doubt on the claim that we need to distinguish between implicit and explicit attitudes (as measured on indirect and direct measures of attitudes) in order to optimally explain and predict people's spontaneous vs. deliberate evaluative responses (see desideratum *D1* of a model of attitudes). Oswald and colleagues' (2013) meta-analysis suggests that results on indirect measures of attitudes (which are supposedly reflective of implicit attitudes) are no

[158] Although it shall be noted, as mentioned in chapter 1 (section 1.2.6), that not every proponent of the standard view regards all of these dimensions (rational control, mental structure, and intentional control) as characteristic of the implicit-explicit distinction.

better predictors of people's spontaneous evaluative responses than results on direct measures of attitudes (which are supposedly reflective of explicit attitudes). Moreover, results on direct measures of attitudes are no better predictors of people's deliberate evaluative responses than results on indirect measures of attitudes (see section 1.3.3). If indirect and direct measures tapped into different kinds of attitudes (i.e., implicit and explicit attitudes), we would expect them to be predictive of different kinds of responses (spontaneous and deliberate responses, respectively), but this is not what we find.

I also suggested that the distinction between implicit and explicit attitudes does not mark, as it often implied, a difference between mental states that form part of a person's moral character and mental states that do not form part of a person's moral character (see desideratum *D2*). It is often claimed that implicit attitudes cannot reflect on a person's moral character because they are outside of the agent's rational and intentional control (Levy, 2014a, 2015; Glasgow, 2016). I argued, by contrast, that people can take at least indirect rational control and indirect intentional control of their so-called implicit attitudes even if these are associative mental states (see section 2.3). Sarah, for example, could take indirect rational control by engaging more frequently in positive thoughts about black individuals (Briñol, Petty, & McCaslin, 2009) or by placing photos of admired black individuals in her office environment (Holroyd & Kelly, 2016) in order to countercondition her negative associations with black people. Moreover, Sarah could take indirect intentional control by forming the intention to think the word "safe" whenever she encounters a black person in order to inhibit the activation of negative associations with black people (e.g., her association between BLACK PERSON and DANGER; Stewart & Payne, 2008). Drawing on Holroyd & Kelly (2016), I suggested that the fact that people can take indirect control (or what they call "ecological control") of their so-called implicit attitudes implies that these mental states can reflect on people's moral character. The fact that both so-called explicit and so-called implicit attitudes may reflect on a person's moral character undermines an important motivation to distinguish between implicit and explicit attitudes (see desideratum *D2*).

I emphasise that I do not claim that we cannot make sense of the standard view of attitudes at all. For example, one possible reformulation of the standard view may be that implicit attitudes are associative mental states that are subject to indirect forms of control (ecological control), while explicit attitudes are propositional mental states that are subject to direct forms of control (see section 2.4).[159] My claim is just that the standard view is not the optimal model of attitudes, provided that we want to adopt the model that accords best with desiderata *D1*, *D2*, and *D3*.

[159] Yet, one may want to reply that even paradigmatic examples of explicit attitudes, such as beliefs, are not (always) subject to direct forms of control (e.g., Hieronymi, 2008; Holroyd, 2012).

II. An alternative conception of attitudes

Rather than identifying attitudes with individual implicit or explicit mental states, I suggested in this thesis that attitudes are better conceived of as traits of people (answer to *Q3*), each of which is grounded in a cluster of different kinds of mental states (e.g., conceptual associations, affects, beliefs, desires, etc.; answer to *Q2;* see chapter 4*)*. The view that attitudes are traits is appealing because there are striking similarities between the explanatory, predictive, and character evaluative roles of trait ascriptions (e.g., ascribing arrogance to a person) and attitude ascriptions (e.g., ascribing a racist attitude to a person; see section 4.4). There is reason to assume that psychologists, philosophers, and ordinary people may possibly find common ground in a trait model of attitudes (see desideratum *D3*). The view that attitudes are traits is at the core of the folk psychological conception of attitudes with which everyone is familiar. When we say that someone has a negative attitude towards immigrants, for example, we do not normally refer to a particular implicit association or explicit belief of the agent. Rather we want to convey that the agent is generally disposed to respond in negative ways towards immigrants. In short, we refer to a general trait of the agent. Scholars in philosophy and psychology will find it immensely difficult to inform public discourse with their attitude research if their notion of an attitude is very different from this folk psychological conception. Of course, folk psychological conceptions may sometimes be misguided, in which case scholars may want to revise these conceptions (P. M. Churchland, 1981; P. S. Churchland, 1986; Stich, 1983). However, as I argued in this thesis, there is in fact a scientifically sound model of attitudes as traits available and there are good reasons to favour this model over alternative accounts of attitudes.

To start with, note that much can be said in favour of the idea that attitudes cannot be identified with individual mental states but rather have a broad psychological basis that is composed of different kinds of implicit and explicit mental states (conceptual associations, affect, beliefs, desires, etc.; answer to *Q2*). In chapter 3, I highlighted that it would be wrong to identify attitudes merely with affective mental states (Sarah's fear of black people) or merely with mental stereotypes (e.g., Sarah's association between BLACK PERSON and VIOLENCE or her propositional mental state with the content "black people are violent") because these classes of mental states tightly interact in the production of people's evaluative responses towards other people (i.e., they form "evaluative stereotypes"; Madva & Brownstein, 2016). If we want that the notion of an attitude can optimally fulfil its explanatory and predictive role, we should acknowledge that attitudes are based at the same time on affective and conceptual/stereotypic mental states. However, conceptual/stereotypic and affective mental states cannot be the only components of attitudes. I emphasised that Sarah does not only harbour

certain black person stereotypes and affective dispositions but also certain moral beliefs (e.g., her belief that it is morally reprehensible to treat people differently because of their skin colour) and desires (e.g., the desire not to discriminate against black people; see section 4.1; see also Besser-Jones, 2008). These mental states also sometimes determine the nature of her evaluative responses towards black people. If we want to optimally explain and predict Sarah's evaluative responses towards black people in the various situations in which she encounters them, we need a model of attitudes that takes all of the above mentioned mental states into account (see desideratum *D1*). I argued that the view that attitudes are traits, each of which is grounded in a variety of mental states (conceptual associations, affect, beliefs, desires, etc.) provides just such a model.

III. Attitudes as traits: dispositional profiles

However, not any characterisation of these traits will do. On Machery's (2016) trait view, attitudes are characterised in terms of an aggregate strength and valence (see section 4.2). This view implies that Sarah lacks an attitude towards black people if she is as strongly inclined to respond in a positive manner towards black people (e.g., based on her egalitarian beliefs, her desire not to behave in a racist manner, etc.) as she is inclined to respond in negative ways towards black people (e.g., based on various negative evaluative stereotypes that she harbours in regard to black people). I find this implication of Machery's account deeply problematic. Sarah's evaluative stance towards black people is clearly different to the evaluative stance of a person whose entire cognitive, affective, and behavioural responses towards black people are more or less neutral in valence (see section 4.6). Yet, Machery's model implies that both these persons lack an attitude towards black people.

To be sure, situationists would also come to the conclusion that Sarah lacks an attitude towards black people. However, they go further than Machery by claiming that no one possesses any traits, including attitudes conceived as traits. According to them people's responses are entirely dependent on situational factors and not on inner response dispositions of the kind that traits are usually identified with (e.g., Doris, 2002; see section 4.5). I disagree with this position. Mischel & Shoda (1995) have convincingly argued that some traits at least can be analysed as "distinctive and stable patterns of behavior variability across situations" (p. 246). I propose accordingly that we can identify attitudes, construed as traits, with stable patterns of evaluative response variation across situations (see section 4.7). On this view, situational variation in evaluative responding does not speak against the existence of attitudes understood as traits but is actually a crucial feature of these. Sarah, for example, can be said to

exhibit an aversive racist attitude towards black people, which is a stable profile of situation-specific evaluative response dispositions. This profile may consist of Sarah's disposition to respond in a favourable manner towards black people in situations in which she has sufficient time and cognitive resources to reflect on and be guided by her endorsed egalitarian commitments and of her disposition to respond in negative ways towards black people in situations in which she does not have sufficient time (e.g., when she has to judge quickly whether she is in danger) or cognitive resources (e.g., when she is deeply engaged in a conversation with a patient) to reflect on and be guided by her endorsed egalitarian commitments. This characterisation does justice to the evaluative complexity of Sarah's attitude and thus provides us with a good basis to explain and predict Sarah's responses towards black people (see desideratum *D1*).

Moreover, I hold that this characterisation provides us with an important insight about Sarah's moral character (see chapter 5). Sarah is neither a pure egalitarian nor is she a pure racist. Her aversive racist attitude lies somewhere in-between (see Schwitzgebel, 2010, 2013, and Levy, 2017b, for related arguments). Proponents of a real self account may object that Sarah's negative evaluative response disposition in regard to black people can hardly reflect on her moral character because she does not endorse this disposition and regrets her unfortunate responses (see section 5.2). My proposed model of attitudes may thus fail to satisfy desideratum *D2*. In response, I showed that the real self perspective (which gives priority to the agent's endorsements) is not unanimously supported by our intuitions, and might just seem appealing due to a self-serving bias: it allows us to see ourselves in a positive light despite the discriminatory responses that we often exhibit (see sections 5.4 and 5.5). Building upon this, I pointed out that there is a pragmatic reason for including non-endorsed evaluative response dispositions in our model of attitudes: it may increase people's motivation to tackle their problematic biases (see section 5.6). If we ascribe to Sarah an aversive racist attitude, this encourages her to perceive her problematic biases as part of "who she is". This may in turn motivate her to do more to get rid of these biases because she certainly does not want to be (perceived as) an aversive racist. As mentioned above, there are indeed some strategies that Sarah could adopt to tackle her biases. In a nutshell, my model of attitudes does not violate desideratum *D2* because there is good reason to treat both Sarah's endorsed evaluative response dispositions and her non-endorsed evaluative response dispositions to be reflective of her moral character. In other words, the distinction between endorsed and non-endorsed evaluative response dispositions should not be treated as a distinction between aspects of a person's psychology that can be said to be constitutive of that person's moral character and aspects of a person's psychology that are not part of that person's moral character.

IV. Attitude individuation

So far I have described Sarah as exhibiting an aversive racist attitude. However, it needs also emphasising that there are different legitimate ways to individuate attitudes (answer to *Q3*; see section 4.7.3). As I described the case of Sarah, it may be salient that the interesting aspect about her responses towards black people is that she shows different evaluative responses dependent on whether she has currently the resources to reflect on and be guided by her egalitarian commitments. However, my description of Sarah can of course only be incomplete. Sarah may encounter black people in all kinds of contexts (in her surgery, when walking home through a deprived neighbourhood, at the supermarket, at a sports club, at a parent's evening at school, etc.) and she may encounter black people with all kinds of different traits (different gender, different age, different profession, different socio-economic status, etc.). Sarah's evaluative responses towards black people may vary dependent on all these situational factors. To give a comprehensible account of her attitude towards black people, we thus need to identify salient or especially noteworthy patterns in Sarah's complex mesh of response dispositions (i.e., identify profiles of situation-specific response dispositions). This pattern detection is clearly, and legitimately, influenced by our interests and purposes as attitude ascribers (see section 4.7.3.1). For example, someone may not be particularly interested in how Sarah responds towards black people dependent on the time and the resources she has available to reflect on her egalitarian values but in how Sarah's evaluative responses towards black people vary dependent on their socio-economic status (e.g., black people who are better off than her, black people who have a comparable status as her, black people who are worse off than her, etc.). Note that on this dimension, too, a noteworthy pattern may be detectable. For example, it may turn out that Sarah tends to feel envious of black people who are better off than her, a mix of pity for and anxiety of black people who are worse off than her, and no particular affective reaction towards black people with comparable status (while there is no such pattern detectable in her responses to white people). We may say that this is Sarah's social status dependent attitude towards black people. Note that describing Sarah as an aversive racist and describing Sarah as a social status dependent racist may both be legitimate if each of these descriptions tracks actual dispositions of her. Which on we choose (or whether we want to take into account both) depends on our interests and also on our explanatory and predictive purposes.

Dependent on our purposes, we may also characterise people's attitude(s) in more or less detail (i.e., we may vary the level of situation-specificity and response-specificity; see section 4.7.3.2). Characterising profiles of situation-specific response dispositions in broad terms ("zooming out on the attitude") allows identifying

commonalities between people's attitudes. People are likely to differ in many details of their situation-specific response dispositions, but once we abstract from these specifics, we may find that different people exhibit the same broad patterns of evaluative responding. Note, for example, that the profile of an aversive racist as defined above is presumably shared by many people. Also, it is easier to convey information about a person's attitude to other people if we can denote the attitude with a simple label such as "aversive racist attitude". However, characterising a person's attitude(s) in more detail ("zooming in on the attitude") provides of course a more accurate basis for explanation and prediction of that person's responses and for an evaluation of that person's moral character.

Lastly, the scope of the attitudes that we ascribe to people is dependent on our interests and purposes (see section 4.7.3.3). Instead of being interested in Sarah's attitude towards black people, we may be interested in her attitude towards black men or her attitude towards strong black men. We can say that her attitude towards black men is "local" in relation to her attitude towards black people and that her attitude towards strong black men is "local" in relation to both her attitude towards black men and her attitude towards black people. Yet, all these attitudes can be analysed as profiles of situation-specific evaluative response dispositions.

V. Summary of key claims

In a nutshell, my answers to questions *Q1*, *Q2*, and *Q3* are as follows. In regard to the question about the ontological status of attitudes (*Q3*), I hold that attitudes are traits of people that can be analysed a profiles of situation-specific evaluative response dispositions. In regard to the question about the mental states that underpin attitudes (*Q2*), I claim that attitudes (construed as traits) are based on a variety of different mental states, which may include amongst others moral beliefs, desires, mental stereotypes (which may have associative or propositional structure), and affective mental states. In regard to the question about attitude individuation (*Q1*), I grant that there are different legitimate ways to individuate attitudes, which are contingent on the attitude ascribers interests and purposes. The described model is conducive to the explanation and prediction of people's evaluative responses (see desideratum *D1*) and to the assessment of people's moral characters (see desideratum *D2*), and may thus appeal to different parties who use the notion of an attitude (psychologists, philosophers, ordinary people; see desideratum *D3*).

VI. Future directions

I would like to conclude with some brief remarks on how the model of attitudes presented here may guide future attitude research in psychology and philosophy. On my view, it is a crucial feature of attitudes that they are composed of situation-specific response dispositions. My model of attitudes thus encourages psychologists to examine how people respond to members of a target group in different contexts (or to members of a target group with different features). All too often researchers regard these situational influences as noise that needs to be eliminated when measuring attitudes or they average across situations to estimate a person's mean evaluative tendency (Ajzen, 1988, chapter 3; Eagly & Chaiken, 2007). This masks the rich texture of people's attitudes towards members of social groups, resulting in a low predictive validity (Oswald et al., 2013; Forscher et al., 2016). To assess this rich texture of attitudes, I propose that psychologists should focus more on the situation-dependency of people's evaluative responses. For example, they could develop standardised tests to assess people's profiles of situation-specific response dispositions. The Ambivalent Sexism Inventory by Glick and Fiske (1996, 1997) is a shining example of such a test (see section 4.6). Which evaluative response patterns scientists will focus on will depend, as indicated above, partly on the scientist's interests as well as explanatory and predictive intentions. Yet, I trust that researchers can find a broad consensus concerning the relevant responses and situations that should be examined. As I have suggested in chapter 4, another valuable project would be to examine which attitudes, construed as profiles of situation-specific response dispositions, are prevalent in particular groups of people (see sections 4.7.2 and 4.7.6). For example, one may examine what patterns of evaluative responses towards immigrants are especially salient in a particular country. Relevant evidence can come from observations of people's responses in real-world settings or from lab experiments in which relevant situational factors are systematically manipulated. Another strand of psychological research could examine what responses and what situations folk psychologists take to be relevant when they assess the attitude of a person towards a social group, and which factors influence which responses and situations they find relevant.

My model of attitudes may also guide further philosophical research on attitudes. I emphasised that an agent's attitude(s) (which may include both endorsed and non-endorsed evaluative response dispositions) reflect on the agent's moral character. Building on this, one may ask what sort of moral appraisal the possession of an attitude warrants. As Holroyd and colleagues (2017a) point out, we "might invoke an evaluative judgement about the agent and her character – she is cruel, or she is racist – without taking a stance on whether this is her fault" (p. 5; see also Watson, 2004). Following

this line of reasoning, saying that someone is an aversive racist (or possesses an aversive racist attitude) expresses a character evaluation but does not yet commit us to say that the agent is blameworthy for her problematic biases or praiseworthy for her egalitarian commitments. Future research may examine accordingly what appraisals are warranted in regard to people's attitudes if these are understood as traits (that can be analysed as profiles of situation-specific evaluative response dispositions). Note that one possible conclusion could be that different attitudes (aversive racist attitudes, ambivalent sexist attitudes, etc.) warrant different kinds of appraisals due to their different structure. Another question worth pursuing is how we should conceive of self-knowledge in relation to attitudes if these are understood as outlined above. It seems that we can be mistaken about our own attitudes. After all, the evidence that we can gather about how we respond to members of a target group in different situations will always be limited (considering that there may be countless relevant kinds of responses and situations) and some parts of the psychological bases of our attitudes may be easier to introspect than others. This again may have implications for the question of moral responsibility for our attitudes and the question of what moral appraisals are appropriate in relation to these attitudes.

References

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790-805.

Ajzen, I. (1988). *Attitudes, personality, and behavior*. Milton Keynes: Open University Press.

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (reissued in 1967, pp. 798–844). New York: Russel & Russel.

Amodio, D. M. (2008). The social neuroscience of intergroup relations. *European Review of Social Psychology*, *19*(1), 1-54.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670-682.

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

Amodio, D. M., & Hamilton, H. K. (2012). Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion*, *12*(6), 1273-1280.

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*(3), 143-148.

Anderson, E. (2010). *The imperative of integration*. Princeton: Princeton University Press.

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies*, *93*(2), 161-188.

Ayduk, Ö., & Gyurak, A. (2008). Applying the cognitive-affective processing systems approach to conceptualizing rejection sensitivity. *Social and Personality Psychology Compass*, *2*(5), 2016-2033.

Baker, L. R. (2009). Non-reductive materialism. In A. Beckermann, B. McLaughlin, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (online edition, pp. 1–14). Oxford: Oxford University Press.

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

Bargh, J. (1994). The four horsemen of automaticity: Intention, awareness, efficiency, and control as separate issues. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 1–40). Hillsdale: Lawrence Erlbaum.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, *43*(3), 437-449.

Beeghly, E. (2015). What is a stereotype? What is stereotyping? *Hypatia*, *30*(4), 675-691.

Bennett, J. (1974). The conscience of Huckleberry Finn. *Philosophy*, *49*(188), 123-134.

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition*, *18*(4), 329-353.

Besser-Jones, L. (2008). Social psychology, moral character, and moral fallibility. *Philosophy and Phenomenological Research*, *76*(2), 310-332.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*(3), 242-261.

Blum, L. (2004). Stereotypes and stereotyping: A moral analysis. *Philosophical Papers*, *33*(3), 251-289.

Blum, L. (2009). Prejudice. In H. Siegel (Ed.), *The oxford handbook of philosophy of education* (pp. 451–468). Oxford: Oxford University Press.

Borgoni, C. (2015). On knowing one's own resistant beliefs. *Philosophical Explorations*, *18*(2), 212-225.

Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review*, *80*(5), 307-336.

Bratman, M. E. (2003). A desire of one's own. *The Journal of Philosophy*, *100*(5), 221-242.

Briñol, P., Petty, R. E., & Horcajo, J. (2008). *Automatic change through deliberative processing of persuasive messages*. Unpublished manuscript.

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures: What is the difference? In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 285-326). New York: Psychology Press.

Brownstein, M. (2017). Implicit bias. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, spring 2017 edition*. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/

Brownstein, M., & Madva, A. (2012a). Ethical automaticity. *Philosophy of the Social Sciences*, *42*(1), 68-98.

Brownstein, M., & Madva, A. (2012b). The normativity of automaticity. *Mind & Language*, *27*(4), 410-434.

Brownstein, M., & Saul, J. (Eds.). (2016a). *Implicit bias and philosophy, volume 1: Metaphysics and epistemology*. Oxford: Oxford University Press.

Brownstein, M., & Saul, J. (Eds.). (2016b). *Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics*. Oxford: Oxford University Press.

Carruthers, P. (2009a). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*(2), 121–182.

Carruthers, P. (2009b). An architecture for dual reasoning. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 109–127). Oxford: Oxford University Press.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. *Psychological Bulletin*, *135*(2), 183-204.

Chen, M., & Bargh, J. A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology*, *33*(5), 541-560.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*(2), 67-90.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified Science of the mind/brain*. Cambridge: MIT Press.

Clark, A. (2007). Soft selves and ecological control. In D. Spurrett, D. Ross, H. Kincaid, & G. L. Stephens (Eds.), *Distributed Cognition and the Will* (pp. 101-121). Cambridge: MIT Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-428.

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

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006-1023.

Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: a sociofunctional threat-based approach to "prejudice". *Journal of Personality and Social Psychology*, *88*(5), 770-789.

Cox, W. T., & Devine, P. G. (2015). Stereotypes possess heterogeneous directionality: A theoretical and empirical exploration of stereotype structure and content. *PloS ONE*, *10*(3), e0122292.

Currie, G., & Ichino, A. (2012). Aliefs don't exist, though some of their relatives do. *Analysis, 72*(4), 788–798.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784-803.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*(1), 100-108.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology*, *47*, 233-279.

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

Dasgupta, N., DeSteno, D., Williams, L. A., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion*, *9*(4), 585-591.

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

Davey, G. C. (1992). Classical conditioning and the acquisition of human fears and phobias: A review and synthesis of the literature. *Advances in Behaviour Research and Therapy*, *14*(1), 29-66.

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *Handbook of implicit cognition and addiction* (pp. 11-28). Thousand Oaks: Sage Publishers.

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

Degner, J., & Wentura, D. (2011). Types of automatically activated prejudice: Assessing possessor-versus other-relevant valence in the evaluative priming task. *Social Cognition*, *29*(2), 182-209.

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

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology*, *60*(6), 817-830.

Doggett, T. (2012). Some Questions for Tamar Szabo Gendler. *Analysis, 72*(4), 764–774.

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous*, *32*(4), 504-530.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Dotsch, R., & Wigboldus, D. H. (2008). Virtual prejudice. *Journal of Experimental Social Psychology*, *44*(4), 1194-1198.

Dovidio, J. F., Brigham, J. C., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice and discrimination: Another look. In C. N. McCrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 276–319). New York: Guilford.

Dovidio, J. F., Esses, V. M., Beach, K. R., & Gaertner, S. L. (2003). The role of affect in determining intergroup behavior: The case of willingness to engage in intergroup affect. In D. M. Mackie & E. R. Smith (Eds.), *From prejudice to intergroup emotions: Differentiated reactions to social groups* (pp. 153–171). Philadelphia: Psychology Press.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*(1), 22-37.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315-319.

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

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

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

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, *5*(1), 1-35.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* New York: Times Books.

Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, *67*(3), 382-394.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360-392.

Fabrigar, L. R., MacDonald, T. K., & Wegener, D. T. (2005). The structure of attitudes. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 79–124). London: Routledge.

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

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion*, *15*(2), 115-141.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & 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), 1013-1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297-327.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229-238.

Fischer, J., & Tognazzini, N. A. (2011). The physiognomy of responsibility. *Philosophy and Phenomenological Research*, *82*(2), 381-417.

Fodor, J. (1983). *The modularity of mind*. Cambridge: MIT Press.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman M., Devine, P. G., & Nosek, B. A. (2016). The effects of moral interactions on implicit racial biases. Manuscript in preparation. Retrieved from https://www.researchgate.net/publication/308926636_A_MetaAnalysis_of_Change_in_Implicit_Bias

Frankfurt, H. G. (1988). Identification and wholeheartedness. In H. G. Frankfurt (Ed.), *The importance of what we care about. Philosophical essays* (pp. 159–167). Cambridge: Cambridge University Press.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, *68*(1), 5-20.

Frankish, K. (2016). Playing double: Implicit bias, dual levels, and self-control. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 23–46). Oxford: Oxford University Press.

Frankish, K., & Evans, J. S. B. T. (2009). The duality of mind: An historical perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–29). Oxford: Oxford University Press.

Frege, G. (1948). Sense and reference. *The philosophical review*, *57*(3), 209-230.

Frijda, N. H. (1986). *The Emotions.* Cambridge: Cambridge University Press.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692-731.

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

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

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology*, *41*(6), 618-626.

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

Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, *23*(5), 552-585.

Gendler, T. S. (2012). Between reason and reflex: Response to commentators. *Analysis*, *72*(4), 799–811.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107.

Glasgow, J. (2016). Alienation and responsibility. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics* (pp. 37–61). Oxford: Oxford University Press.

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491-512.

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, *21*(1), 119-135.

Goldie, P. (2004). *On personality*. London: Routledge.

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

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464-1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1-20.

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

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society, 99*, 315-331.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character, volume 1: Studies in deceit.* New York: Macmillan.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, *34*(1), 31-55.

Haslanger, S. (2005). What are we talking about? The semantics and politics of social kinds. *Hypatia*, *20*(4), 10-26.

Hieronymi, P. (2008). Responsibility for believing. *Synthese*, *161*(3), 357-373.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & 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), 1369-1385.

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

Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition*, *33*, 511-523.

Holroyd, J. (2016). What do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society*, *116*(2), 153-179.

Holroyd, J., & Kelly, D. (2016). Implicit bias, character, and control. In A. Masala & J. Webber (Eds.), *From personality to virtue: Essays on the philosophy of character* (pp. 106–133). Oxford: Oxford University Press.

Holroyd, J., Scaife, R., & Stafford, T. (2017a). Responsibility for implicit bias. *Philosophy Compass, 12*(3), e12410.

Holroyd, J., Scaife, R., & Stafford, T. (2017b). What is implicit bias? *Philosophy Compass, 12*(10), e12437.

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

Huebner, B. (2009). Troubles with stereotypes for spinozan minds. *Philosophy of the Social Sciences*, *39*(1), 63-92.

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

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

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012). Disgusting smells cause decreased liking of gay men. *Emotion*, *12*(1), 23-27.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, *9*(3), 435-439.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, *21*(3), 384-388.

Jost, J. T., & Kay, A. C. (2005). Exposure to benevolent sexism and complementary gender stereotypes: consequences for specific and diffuse forms of system justification. *Journal of Personality and Social Psychology*, *88*(3), 498-509.

Judd, C. M., Blair, I. V., & Chapleau, K. M. (2004). Automatic stereotypes vs. automatic prejudice: Sorting out the possibilities in the weapon paradigm. *Journal of Experimental Social Psychology*, *40*(1), 75-81.

Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin Books.

Kamtekar, R. (2004). Situationism and virtue ethics on the content of our character. *Ethics*, *114*(3), 458-491.

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

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin*, *24*(4), 407-416.

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, *3*(3), 522-540.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025-2047.

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help*. New York: Appleton-Century-Croft.

Levy, N. (2011). Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytic Philosophy*, *52*(4), 243-261.

Levy, N. (2013). The importance of awareness. *Australasian Journal of Philosophy*, *91*(2), 211-229.

Levy, N. (2014a). *Consciousness and moral responsibility*. Oxford: Oxford University Press.

Levy, N. (2014b). Consciousness, implicit attitudes and moral responsibility. *Noûs*, *48*(1), 21-40.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*, *49*(4), 800-823.

Levy, N. (2017a). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research*, *94*(1), 3-26.

Levy, N. (2017b). Am I a racist? Implicit bias and the ascription of racism. *The Philosophical Quarterly*, *67*(268), 534-551.

Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to the philosophy of mind* (pp. 412–431). Oxford: Blackwell.

Lippert-Rasmussen, K. (2003). Identification and responsibility. *Ethical Theory and Moral Practice*, *6*(4), 349-376.

Machery, E. (2016). De-freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 104–129). Oxford: Oxford University Press.

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology*, *79*(4), 602-616.

Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese, 193*(8), 2659-2684.

Madva, A., & Brownstein, M. (2016). Stereotypes, prejudice, and the taxonomy of the implicit social Mind. *Noûs*. Advance online publication. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/nous.12182/full

Mandelbaum, E. (2014). Thinking is believing. *Inquiry*, *57*(1), 55-96.

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

Margolis, E., & Laurence, S. (2007). The ontology of concepts—abstract objects or mental representations? *Noûs*, *41*(4), 561-593.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*(5), 435-442.

McIntyre, A. (1993). Is akratic action always irrational? In O. J. Flanagan & A. Oksenberg Rorty (Eds.), *Identity, character, and morality: Essays in moral psychology* (pp. 379–400). Cambridge: MIT Press.

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin*, *36*(4), 512-523.

Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice*, *3*(4), 365-383.

Milgram, S. (1974). *Obedience to authority*. New York: Harper and Row.

Mischel, W. (1968). *Personality and assessment.* New York: John J. Wiley and Sons.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246-268.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*(3), 455-469.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*(3), 469-485.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395-417.

Morin, R. (2015). Exploring racial bias among biracial and single-race adults: The IAT. Pew Research Center, Washington, D.C. Retrieved from http://www.pewsocialtrends.org/2015/08/19/exploring-racial-bias-among-biracial-and-single-race-adults-the-iat/

Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, *47*(1), 103-116.

Nagel, J. (2012). Gendler on Alief. *Analysis, 72*(4), 774–788.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565-584.

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65-69.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the implicit association test. *Experimental Psychology*, *54*(1), 14-29.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36-88.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 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*(2), 181-192.

Payne, B. K., Burkley, M. A., & 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), 16-31.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277-293.

Peterson, D. R. (1968). *The clinical study of social behavior.* New York: Appleton-Century-Crofts.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*(10), 1915-1926.

Petty, R. E., Harkins, S. G., & Williams, K. D. (1980). The effects of group diffusion of cognitive effort on attitudes: An information-processing view. *Journal of Personality and Social Psychology*, *38*(1), 81-92.

Pickering, M. (2001). *Stereotypying: The politics of representation*. New York: Palgrave Macmillan.

Purkiss, S. L. S., Perrewé, P. L., Gillespie, T. L., Mayes, B. T., & Ferris, G. R. (2006). Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes*, *101*(2), 152-167.

Rachman, S. (1991). Neo-conditioning and the classical theory of fear acquisition. *Clinical Psychology Review*, *11*(2), 155-173.

Richeson, J. A., & Ambady, N. (2001). Who's in charge? Effects of situational roles on automatic gender bias. *Sex Roles*, *44*(9), 493-512.

Rokeach, M., & Cochkane, R. (1972). Self-confrontation and confrontation with another as determinants of long-term value change. *Journal of Applied Social Psychology*, *2*(4), 283-292.

Rokeach, M., & McLellan, D. D. (1972). Feedback of information about the values and attitudes of self and others as determinants of long-term cognitive and behavioral change. *Journal of Applied Social Psychology*, *2*(3), 236-251.

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523-534.

Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective, and behavioral components of attitudes. In C. I. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change* (pp. 1–14). New Haven: Yale University Press.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes & Intergroup Relations*, *10*(3), 359-372.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin*, *26*(11), 1315-1328.

Rudman, L. A., & Lee, M. R. (2002). Implicit and explicit consequences of exposure to violent and misogynous rap music. *Group Processes & Intergroup Relations*, *5*(2), 133-150.

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

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954-958.

Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change?* (pp. 39–60). Oxford: Oxford University Press.

Scaife, R., Stafford, T., Bunge, A., & Holroyd, J. (2016). The effects of moral interactions on implicit racial biases. Manuscript submitted for publication. Retrieved from https://osf.io/d69uv/

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*(5), 638-656.

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 436–457). Malden: Blackwell Publishers.

Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly*, *51*(202), 76-82.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, *36*(2), 249-275.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*(4), 531-553.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New essays on belief: Constitution, content and structure* (pp. 75–99). Basingstoke: Palgrave Macmillan.

Schwitzgebel, E. (2015). Belief. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, summer 2015 edition*. Retrieved from https://plato.stanford.edu/archives/sum2015/entries/belief/

Sechrist, G. B., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. *Journal of Personality and Social Psychology*, *80*(4), 645-654.

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring causes of the self-serving bias. *Social and Personality Psychology Compass*, *2*(2), 895-908.

Sherman, S. J., Rose, J. S., Koch, K., Presson, C. C., & Chassin, L. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, *22*(1), 13-39.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*(4), 674-687.

Sie, M., & van Voorst Vader-Bours, N. (2016). Stereotypes and prejudices: Whose responsibility? In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics* (pp. 90-114). Oxford: Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3-22.

Smith, A. M. (2004). Conflicting attitudes, moral agency, and conceptions of the self. *Philosophical Topics*, *32*(1/2), 331-352.

Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, *115*(2), 236-271.

Smith, E. R., & Conrey, F. R. (2007). Mental representations are states, not things: Implications for implicit and explicit measurement. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 247–264). New York: Guilford Press.

Smith, E. R., & 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), 108-131.

Smith, E. R., & Semin, G. R. (2004). Socially situated cognition: Cognition in its social context. *Advances in Experimental Social Psychology*, *36*, 53-117.

Smith, E. R., & Semin, G. R. (2007). Situated social cognition. *Current Directions in Psychological Science*, *16*(3), 132–135.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, *173*(5), 1203-1232.

Stammers, S. (2016). Awareness, control and responsibility for implicit bias: The continuum thesis (Doctoral dissertation). Retrieved from https://kclpu re.kcl.ac.uk/portal/en/theses/awareness-control-andresponsibility-for-impli cit-bias(95558ff7-44cb-4d8c-802a-92ecd8200eae).html

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahway: Lawrence Erlbaum Associates.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), 1332-1345.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge: MIT Press.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220-247.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662.

Stump, E. (1988). Sanctification, hardening of the heart, and Frankfurt's concept of free will. *The Journal of Philosophy*, *85*(8), 395-420.

Tapias, M. P., Glaser, J., Keltner, D., Vasquez, K., & Wickens, T. (2007). Emotion and prejudice: Specific emotions toward outgroups. *Group Processes & Intergroup Relations*, *10*(1), 27-39.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35-39.

Twain, M. (1884). *Adventures of Huckleberry Finn*. London: Chatto & Windus.

Valian, V. (1999). The cognitive bases of gender bias. *Brooklyn Law Review*, *65*(4), 1037-1062.

Valian, V. (2005). Beyond gender schemas: Improving the advancement of women in academia. *Hypatia*, *20*(3), 198-213.

Velleman, J. D. (1992). What happens when someone acts? *Mind*, *101*(403), 461-481.

Washington, N., & Kelly, D. (2016). Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics* (pp. 10–36). Oxford: Oxford University Press.

Watson, G. (1975). Free agency. *The Journal of Philosophy*, *72*(8), 205-220.

Watson, G. (2004). Two faces of responsibility. In G. Watson (Ed.), *Agency and answerability: Selected essays* (pp. 260–286). Oxford: Oxford University Press.

Webber, J. (2006). Virtue, character and situation. *Journal of Moral Philosophy*, *3*(2), 193-213.

Webber, J. (2013). Character, attitude and disposition. *European Journal of Philosophy*, *23*(4), 1082-1096.

Webber, J. (2016a). Habituation and first-person authority. In R. Altshuler & M. J. Sigrist (Eds.), *Time and the philosophy of action* (pp. 189–204). New York: Routledge.

Webber, J. (2016b). Instilling Virtue. In A. Masala & J. Webber (Eds.), *From personality to virtue* (pp. 134–154). Oxford: Oxford University Press.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41-78.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101-126.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262-274.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of Experimental Social Psychology*, *37*(3), 244-252.

Zheng, R. (2016). Attributability, accountability, and implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics* (pp. 62–89). Oxford: Oxford University Press.