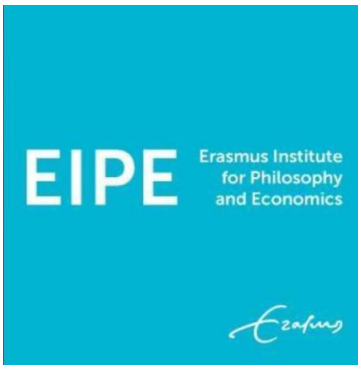

IMPLICIT BIAS - OUT OF CONTROL?

THE CULTIVATION OF RESPONSIBLE AGENCY

Research Master Thesis
Maximilian Gasser
Student number: 517289

Supervisor: Prof. Dr. Constanze Binder
Advisor: Prof. Dr. Alex Voorhoeve

Date of completion: 19.07.2021
Number of ECTS credits: 30
Word count: 28.912

Table of Contents

INTRODUCTION

Research in implicit social cognition suggests that implicit biases influence our behaviour on an everyday basis. Implicit biases are defined as evaluative tendencies (prejudices or stereotypes) we hold towards stigmatized social groups, such as women or people of African descent. Importantly, we are mostly not aware that implicit biases influence us and that we, therefore, act in a stereotyped way. Furthermore, on reflection, we strongly disagree with this behaviour. In that sense, our behaviour is 'biased' in that it contradicts our explicit judgements. (Brownstein 2019; Buckwalter 2019)

For example, analysing the behaviour of study participants in a computer simulation, Glaser and Knowles (2008) found evidence of implicit stereotypes and prejudices associating the presence of weapons more often with Black men than with White men. Furthermore, in the computer simulation, these implicit associations were positively correlated with a tendency to "shoot" armed Black men faster than armed White men. People were not aware of these unconsciously formed implicit biases concerning the social category 'race'. They stated that they did not at all intend to act in a racist manner. On reflection, they strongly disagreed with these implicit attitudes.

In this sense, implicit bias poses a puzzling problem for many theories of moral responsibility. We are largely unaware of acting in a discriminatory, implicitly biased way. Furthermore, we deliberatively strongly object to such discriminatory ways of behaving. Thus, one might conclude that we lack control over the influence of implicit bias. But then, how can it be fair to blame us? Are we in control and morally responsible for the impact of implicit bias on our behaviour? (Buckwalter 2019)

1.1 Thesis Project

This brings us to the main research question I aim to answer: *when and in what sense are we morally responsible for harmful behaviour caused by implicit bias*?

I will address this question by focussing on our ability to *control* our behaviour. As noted, implicit bias seems to be out of our control. At least in so far as we are greatly unaware of it and, on reflection, strongly disagree with this kind of conduct, the influence of implicit bias seems uncontrollable. The present piece shall encounter this challenge: are we, in some sense, in control of the influence of implicit bias and, therefore, morally responsible for our harmful behaviour?

I will approach what it means to have control over implicit bias and to be morally responsible from a 'revisionist' perspective. In light of challenging cases such as implicit bias, revisionist approaches suggest revising conventional theories and our often backwards-looking intuitions about control and moral responsibility. In this piece, I defend, supplement, and scrutinize one particular revisionist and forward-looking control account applied to implicitly biased conduct, that of Manuel Vargas (2013). In a nutshell, for Vargas, whether we are morally responsible depends on whether we are capable of controlling implicit bias in the circumstances in which we act. Furthermore, our having such control capacities in circumstances is greatly determined by whether blaming or praising us generally contributes to making us better able to control what we do. Thus, the attribution of circumstantial control capacities depends on the valuable effects of our responsibility practices, our blaming or praising each other. This is the simplified, short story of his forward-looking account.

But can such a revisionist and forward-looking control account be promising for addressing the research question and moral responsibility for implicitly biased behaviour? I shall argue that Vargas` (2013) view is promising, although it also faces limitations and requires supplementation.

Let me now briefly outline three caveats related to the present approach to the research question. After that, I will summarize my answer to the research question and the argumentative strategy through which I will arrive at it.

1.2 Three Caveats

First, I will be concerned with moral responsibility for the harmful *impact and influence* of implicit bias on behaviour. This is different from the question of moral responsibility for merely having/ inhibiting implicit biases while never acting upon them (Brownstein 2019; Dominguez 2020, 155).

Second, this piece takes an ontological, desert-oriented approach and focuses on *being* morally responsible for implicit bias. With that, I mean whether we deserve a moral response due to behaviour that can be seen as our fault or attributable to us as persons. More specifically, with the question of 'are we morally responsible?', I focus on whether we can be thought of as having a certain sense of control over the influences of our implicit biases and *therefore* deserve moral responses towards us for our failure to control them. This is different from whether we *should be* praised or blamed irrespective of whether we are able to control what we do and, therefore, morally responsible (Sie 2018). I will come back to this in Part Five.

Third, in this piece until Part Five, being morally responsible is understood as being liable to praise or blame. I focus on the negative influence of implicit bias and the harm caused. Therefore, I bracket praiseworthiness. How it can be justified to be a target of other moral responses than praise or blame crucially figures in Part Five.

1.3 Argumentative Strategy and Chapter Overview

Research question: when and in what sense are we morally responsible for harmful behaviour caused by implicit bias? I will give the following, roughly summarized answer:

> I shall defend a revisionist, forward-looking view which implies that in many cases, we are morally responsible for the influence of implicit bias in the sense of being blameworthy for it. We have the required capacities to control implicit bias, given the context in which we find ourselves. Thereby, whether we have such circumstantial capacities is understood as greatly dependent on whether blaming us (e.g., reacting with indignation) can generally help us improve these capacities and become better able to control our conduct (Vargas 2013). On the other hand, though, I shall also establish that if blaming does not generally help us become better persons who can better control harmful behaviour induced by implicit bias, we are not morally responsible. But what can we do then? I will argue that in that case, other moral responses, such as an appeal to universal values, can still be justified because of their coordinative function of settling the content of normative expectations. And crucially, this second function can be separated from the cultivation of valuable control for the addressees through blame (Sie 2018). Furthermore, even when we lack the requisite control, I shall concede that there might be other forward-looking and not control-related reasons for which we should be held morally responsible for the harm our behaviour caused (Ciurria 2019).

I will develop this answer arguing that the revisionist and forward-looking control account of Vargas (2013) is, to a large extent, suitable to address the research question. My strategy is the following:

In **Part Two**, I will clarify what a control account of moral responsibility should accommodate if it aims to successfully answer the research question. I will establish that this includes the peculiar nature of implicit bias and the challenge it poses to some theories of moral responsibility. I will formulate three conditions a control account of moral responsibility should accommodate:

Condition (1): it should be flexible enough to accommodate the instability and context-dependency of the influence of implicit bias.

Condition (2): it should accommodate the fact that implicit biases can cause harm and contribute to discriminatory practices.

Condition (3): it should accommodate that we might commonly expect people to be, in some way, aware of the harmful influence of implicit bias on behaviour and to do something about it dependent on the context and the roles they have.

In **Part Three**, I will show that Vargas` (2013) revisionist and forward-looking control account of moral responsibility, which entails a 'circumstantial' view on morally responsible agency, is a promising one for addressing the research question above. I will show that Vargas` account can accommodate condition (1), the context-dependency of implicit bias`s influence. This is because our capacities to control implicit bias that ground our responsibility for failing to do something are conceived of as highly context-dependent. Furthermore, the forward-looking account accommodates condition (2), the necessity to avoid harm. Whether a person is morally responsible for implicitly biased behaviour depends on whether holding her blame- or praiseworthy contributes to the cultivation of her capacity to control and avoid these harmful influences. Finally, it is promising to accommodate condition (3), ascribing a decisive role to our context-dependent expectations when determining responsibility for implicit bias. Based on this, I shall propose the following *provisional* answer to the research question: people

are morally responsible for implicit bias`s harmful influence in the sense of blameworthy if moral responsibility contributes to cultivating their circumstantial capacity to control it and, thus, to avoid harm.

In **Part Four**, I will defend Vargas` (2013) 'circumstantial' account of responsible agency against objections from Jules Holroyd (2018). Holroyd objects that McGeer`s (2015) revisionist scaffolded responsiveness view better contributes to the forward-looking aim of cultivating control than Vargas` circumstantial view. In addition, she objects that McGeer`s account is more in line with both conventional theories and our existent responsibility practices. An argument crucial to these objections is that Vargas` circumstantial account is bound to our current circumstances. It neglects the potential significance of what Holroyd coins as 'indirect reasons'. These can be understood as expectations from people who we encounter not in our current contexts but potentially in the future. In answering these objections, I will argue that Vargas` concept of 'circumstantial' capacities is not strictly bound to current circumstances narrowly understood. Crucially, I will show in what sense circumstantial capacities are concerned with a person`s ability to learn and with expectations, including those that stem from potential future encounters. This shall help me amend Vargas` view, outline the role expectations play in it, and thus, explain how it can accommodate condition (3). Based on this, I will *specify* the provisional answer developed in Part Three by complementing it. Namely, I assert that people might be morally responsible for implicit bias also dependent on what they might be able to expect from themselves and others in the future in a specific context. Furthermore, I will describe what blaming and non-blaming amount to in practice.

In **Part Five**, I will consider a limitation of Vargas` view. For Vargas, people are not morally responsible (understood as liable to blame) if blaming does not generally contribute to avoiding harm. Now, in some situations, people do not expect each other to control implicit bias. This lack of expectations might undermine the effectiveness of blaming and moral responsibility. But if people are sometimes not morally responsible in Vargas` account, what can be done to avoid the harm induced by implicit bias? Addressing this worry, I will outline an essential conceptual limitation Vargas` view faces, which stems from its exclusive concern with praise- and blameworthiness. To overcome this limitation, I will propose to supplement Vargas` account by drawing on Sie`s (2014, 2018) conversational view on human agency and the different social functions of moral responses. Furthermore, I shall make evident that the present focus on control as grounding a person`s deserving moral responses is strictly limited in accounting for the harm produced by implicit bias. Finally, I will arrive at the following *supplementation* to the answer, which can address conditions (1), (2), (3) of this piece more fully than Vargas` account alone: even if people are not morally responsible and blameworthy, they can still be addressed by *other moral responses than blaming*. This is justified because it can contribute to collectively determine the content of normative expectations present in a context. And crucially, this second function of moral responses can be separated from another function blaming someone for exhibiting implicit bias has, which is the cultivation of control capacities and responsible agency for the addressees.

PART TWO: OUT OF CONTROL? – THE RELEVANCE AND CHALLENGE OF IMPLICIT BIAS

The following scenario forms an example of a behavioural influence that stems from implicit bias and contributes to discrimination. I will continuously come back to it in the present piece (inspired by Holroyd 2018, 146):

> *Suppose Tom owns a chain of fitness studios and will soon open a new one. Therefore, he is looking for new trainers and evaluates résumés. Most résumés are from racialized White male and female applicants. Tom explicitly believes that men and women and racialized Black and White people are equally suited for the job of a trainer. Though, he has implicit biases that negatively influence his evaluation of résumés both of female trainers and of not racialized White ones. Due to this influence of stereotypes and despite comparable qualities of all applicants, Tom ends up with a list of only male and only White applicants who he deems suitable for the job. Tom is not aware of the existence of the phenomenon of implicit biases and their influence on his judgement, nor about his potential ability to control or monitor implicit biases when evaluating the résumés. And indeed, he is still convinced of having chosen the "best" trainers for his studio.*

When and in what sense are people (such as Tom) morally responsible for their harmful, stereotyped behaviour caused by implicit bias? As noted, this research question regarding moral responsibility for the influence of implicit bias is the one I aim to answer in this piece. Thereby, I focus on people`s capacity to control implicitly biased behaviour. Many conventional (mostly backwards-looking) theories of moral responsibility regard control as a necessary condition for being

morally responsible. Only if we are, in some sense, in control of what we do, we are also morally responsible and deserve to be blamed or praised.

In Part Two, I formulate three conditions (condition (1) in Section One, conditions (2) and (3) in Section Two) that determine what a *control-based* account of moral responsibility for implicitly biased behaviour should achieve for successfully addressing the research question.

2.1 Implicit Bias - What It Is and Why It Matters

The term “implicit bias” was coined in 1995 by psychologists Mahzarin Banaji and Anthony Greenwald (Greenwald and Banaji 1995). In this piece, I shall endorse the following working definition of implicit bias (cf. Brownstein 2019, Buckwalter 2019):

(a) Implicit bias refers to an involuntary tendency to evaluate an individual (mostly negatively) based on the perceived membership in a social group or category. These implicit attitudes contradict a person`s explicit judgements.
(b) This evaluation can influence people`s behaviour and be harmful to individuals who belong to certain social groups or are associated with certain social categories.
(c) And the behavioural influence of this evaluation is at least partly uncontrollable and unconscious.

In Section One, I will critically assess the empirical foundations of this working condition. I will first ground the relevance of feature (a) by briefly discussing evidence on implicit tendencies of negative evaluations dependent on perceived group membership. Then, I will assess (b) whether such implicit biases matter: Can implicit biases influence behaviour? And can these influences contribute to harm? By shedding light on some arguments concerning the current state of empirical research, I will emphasize that whether implicit biases are more or less likely to influence behaviour grounds condition (1) for this piece. An account of moral responsibility for the influence of implicit bias must be able to accommodate that this influence is highly dependent on behavioural, personal and situational factors and on how they interact. Finally, I will critically discuss (c) evidence on the unawareness and the uncontrollability of implicit bias. This shall provide the grounds for Section Two, where I formulate two other conditions

engaging with the challenge implicit bias poses to some theories of moral responsibility.

2.1.a Do Implicit Biases Exist?

Note that what makes implicit biases ‘implicit’ is that they are typically thought of as unconscious (I will closer assess this in 2.1.c). But why can we assume that such implicit biases exist in the first place? Do individuals sometimes (involuntarily) tend to evaluate others negatively based on perceived group membership (feature (a) of the working definition)?

The implicit association test (IAT) is widely used to measure the existence and one`s endorsement of implicit attitudes that may contradict one´s explicit judgements. In a nutshell, the IAT measures the reaction time of individuals who are asked to sort words and pictures in categories without making mistakes and as fast as possible. Thus, the IAT aims at activating implicit associations of words and pictures with social categories (such as socialized “race”) that coincide with commonly held social stereotypes and prejudices. But what does it mean if the IAT delivers that one “has” an implicit bias? While implicit measures should not be understood in the sense that one is ‘racist’, they represent momentary pictures, how a person is implicitly biased at a certain moment in time and in a certain context. They are momentary, situational-dependent pictures of a person`s implicit attitudes. (Brownstein et al. 2020, 288-289)

If such implicit evaluations such as stereotypes and prejudices do not coincide with explicit judgements or convictions (such as what has been measured in self-reports), these measured implicit attitudes form instances of implicit ‘biases’ (Brownstein 2019). For example, Nosek et al.`s (2007) findings indicate that racial bias is widespread. Many who explicitly state not to have any preference of

Whites over African Americans are still likely to be implicitly biased against African Americans.

This is how the existence of an individual`s momentary implicit attitudes and biases is commonly assessed. Indeed, a mountain of evidence demonstrates the existence of such implicit biases against women, Muslims, the elderly, the obese, and persons with mental illnesses[1].

The scenario of Tom describes a relevant kind of situation in which someone is influenced by an implicit bias. It will be frequently referred to in this piece. Although the relevance of this piece is not restricted to implicit biases in hiring processes, the scenario, or a similar one, is relevant given that research suggests that implicit stereotypes exist and operate in the context of hiring processes and employment decisions as well. For example, Moss-Racusin et al.`s (2012) study found that participating members of science faculties from research-intensive universities rated identical applicants for laboratory manager positions significantly more favourable (whereby the effects were moderate to large) if they were randomly assigned a male rather than a female name. (Bendick Jr. and Nunes 2012)

Thus, it is plausible to assume (a) the existence and operation of implicit biases in many domains, including contexts of hiring processes. However, could Tom`s implicit biases influence his behaviour in a *causal* way producing harm? Assessing the relevance of implicit bias, I will now scrutinize feature (b) of the working definition.

[1] As cited in Buckwalter 2019: for against women, see Dasgupta and Asgari 2004; for Muslims, see Park et al. 2007; for the elderly, see Castelli et al. 2005; for the obese, see O'Brien et al. 2007; for persons with mental illnesses, see Rüsch et al. 2010.

2.1.b Do Implicit Biases Contribute to Harm?

Do implicitly operating stereotypes and prejudices influence behaviour in a causal way that contributes to harm (assumed in feature (b) of the working definition)? Some, such as Buckwalter, are sceptical about the validity of findings in the field of implicit social cognition: "*[…] it appears that assuming causation is premature*" (Buckwalter 2019, 2971). Shedding light on the empirical debate regarding the relevance of implicit bias, I shall emphasize that rather than asking whether implicit bias causes harmful behaviour, one should ask *when* it can do so. This shall serve as condition (1) of this piece. (Buckwalter 2019; Brownstein et al. 2020)

Meta-analyses found only small positive average correlations between implicit attitudes and individual behaviour, while results are mixed. Thus, do implicit biases even matter when it comes to behaviour? Kurdi and Banaji (2017) draw the conclusion that they don`t. They conclude that only between 1% and 8% of the variation in actual discriminating behaviour between groups can be explained based on the variation between individuals` measures of implicit bias. (Buckwalter 2019, 2969-72)

Brownstein et al.`s interpretation of the same findings is different. They assert that the small but positive average correlation still counts as a "predictive success" (Brownstein et al. 2020, 279). This is due to the high dependence of implicit bias's influence on other factors, such as the context or the person. And due to this ignorance of other factors, mixed results in average correlations are plausible as well. (Brownstein et al. 2020, 279-82)

Let us have a closer look at person-specific and context-specific factors that have been found to influence the impact of an implicit attitude on behaviour. For example, a person`s capacity for working memory (short-term memory) can affect the influence of and capacity for control of implicit attitudes (Friese et al. 2008).

The higher an individual`s ability for working memory, the weaker the association between her implicit attitudes and behaviour. Thus, Tom`s capacity for working memory might affect whether implicit bias influences him and whether he can control this influence. Contextual factors play a role in the influence of implicit bias as well. For example, the stability of implicit bias over contexts can be influenced by what types of images are used or whether relevant contextual cues are made salient. For instance, Gschwendner et al. 2008 found that when measuring implicit evaluations of German versus Turkish faces, varying the background on the computer screens from a garden to a Mosque significantly increased the accessibility and stability of implicit biases. In Part Three, I will come back to how an account of moral responsibility can accommodate such personal and contextual factors. (Brownstein et al. 2020, 278, 284-87)

Some conclude that high dependence on context undermines the relevance of implicit bias as the causal source of structural harm and, thus, the relevance of this piece. Doesn`t this mean that only the context matters? Indeed, some (e.g., Payne et al. 2017; Hehman et al. 2017) propose to shift attention from individuals to situational factors and aggregated implicit measures[2].

While researchers disagree on the importance of situational versus personal factors, most do not neglect that implicit biases still matter for discriminatory outcomes. Instead, viewing individual factors in *interaction* with situational factors is taken to be a more promising perspective on the matter. For example, evidence suggests that sometimes weaker implicit attitudes are associated with stronger effects of situational factors on behaviour (Granados Samayoa and Fazio 2017). Other studies explicitly test for interactions of implicit biases with both personal and situational factors in producing discriminatory outcomes. Cesario et al. (2010) analysed such "personality-by-implicit-bias-by-situation interactions".

[2] This also addresses the critique of some that research on implicit bias would primarily reflect an individual focus (e.g., Haslanger, 2015).

They found that differences in confrontational personalities still influenced discriminatory behaviour among subjects who were implicitly biased. This suggests that it matters not only what personal *or* contextual factors are present when Tom evaluates resumes for whether implicit bias influences him. But it also seems to matter how Tom`s personality (e.g., his confrontational attitudes) interacts with and *relates* to the context. As I will show in Part Three, the account of moral responsibility I propose shall be able to account for this peculiarity about the influence of implicit bias. (Brownstein et al. 2020, 291)

Furthermore, note that implicit bias`s influences on seemingly trivial individual actions can accumulate and become systemic disadvantages for particular social groups. For example, if a female professor was continuously approached by her students as a secretary, being approached in that way every day can contribute to an environment of systemic disadvantage. (Brennan 2016)

In sum, while the degree of harmful influence of implicit attitudes remains a matter of debate, their general relevance for harmful outcomes is mostly not neglected. What seemingly determines the influence of implicit bias on behaviour and its strength shall form condition (1) of the present piece: an account of moral responsibility should be able to flexibly accommodate the nature of implicit bias`s influence, its dependence on the person, the context, and on how all factors interact. (Brownstein et al. 2020)

2.1.c Are Implicit Biases Unconscious and Uncontrollable?

Are agents (c) unaware and unable to control implicit biases? I will now specify what I will start out from in Section Two. Namely, that evidence suggests that people lack awareness of the influence of implicit bias on their behaviour and the ability to directly control this influence. In Section Two, these assertions will

partly ground the challenge implicit bias poses for some traditional accounts of moral responsibility.

2.1.c.i Are Implicit Biases Unconscious, and What Does That Even Mean?

What makes implicit attitudes and biases 'implicit' is typically taken to be their unconscious nature (Gawronski et al. 2006). But is this so? And what does 'unawareness' mean in the first place? Unconsciousness or unawareness of implicit attitudes and biases can have different forms. Gawronski et al. (2006) distinguish between two variants of unconsciousness which are relevant in this piece. First, a person might lack *content* awareness. This means that Tom might be unaware about inhibiting (having) an implicit prejudice, associating perceived members of socially stigmatized groups to common, negative stereotypes. Second, a person might lack *impact* awareness. Tom might be unaware that he is influenced by implicit bias while evaluating the resumes, leading him to act in a stereotyped, harmful way.

Evidence suggests that it might be the case that someone like Tom has, in fact, and contrary to the scenario sketched above, *content* awareness. He might be aware of himself potentially having implicit biases or something like it[3]. In this

[3] Gawronski et al. (2006) assert that while agents often lack impact awareness, they often dispose of content awareness. This is supported by Madva (2018) as well, who suggests that many are indeed aware that they inhibit implicit biases, or at least something similar to it. In fact, Hahn et al.`s (2014) and Hahn and Gawronski`s (2019) findings support this empirical picture. They indicate that many people can well predict their own IAT scores, and this, regardless of the description or explanation of the test or their previous experiences with it.

Some have questioned whether content awareness is a 'sufficient' form of awareness for constituting self-knowledge about one`s implicit biases. This might be questionable at least in so far as self-knowledge is to be based on justified, true beliefs. It might matter how an agent achieves his content awareness of inhibiting an implicit bias. If it is achieved by making inferences e.g., from her every-day behaviour, then it might well be the case that her inferences on her inhibiting implicit biases or not are unjustified and false. Whether this content awareness

piece, for simplicity, I shall largely grant Tom`s unawareness. Though, I will also come back to the possibility of him having some content awareness and the implications of this for his moral responsibility in Part Four.

On the other hand, recall that this piece is not concerned with a person`s moral responsibility for having an implicit bias but for the harmful *impacts/ influences* of implicit bias on her behaviour. Evidence suggests that agents often lack awareness over these impacts (Gawronski et al. 2006). Thus, and most importantly for the present piece, it is plausible to assume, as done in our scenario, that Tom is not aware of the *impact* of implicit biases on his behaviour. He is not aware of his acting in a stereotyped, harmful way, being influenced by implicit bias while evaluating resumes. As I will show in Section Two, this can matter crucially for his moral responsibility.

2.1.c.ii Are Implicit Biases Uncontrollable?

'Direct' control is understood as being able to control one`s behaviour right here and now, based on reasons one evokes right here and now (as roughly defined by Vargas (2020). I will come back to this in Section Two).

On the one hand, for some (e.g., Saul 2013), a lack of awareness about the impacts of implicit bias can constitute a lack of such direct control. If Tom is not aware of being influenced right here and now, he is also not able to directly/ immediately control the influence of implicit bias.

On the other hand, as established, many are aware of having something like implicit bias. Many have content awareness. Saul, who is sceptical about moral

---

can therefore be sufficient to constitute self-knowledge about one`s implicit biases will ultimately depend on the epistemic capacities of an agent to interpret her own mind. E.g., Levy (2014) asserts that people have these capacities. (Brownstein 2019)

responsibility for implicit bias, argues that even when being made aware of inhibiting implicit biases, we are still not instantly able to control them directly. Tom cannot simply direct the impact of implicit bias on his behaviour in the precise moment in time in which it affects him. Although being (content-wise) aware of him possibly inhibiting implicit biases, he would still lack direct control over his behaviour in so far as it is affected by implicit bias. (Saul 2013, 55)

For the case of implicit bias, Holroyd (2012) and Sie and Vader-Bours (2016) differentiate a person`s 'indirect' ability to control her behaviour from her 'direct' ability to do so. Holroyd plausibly asserts that there are many attitudes or abilities over which we do not have immediate, direct control. And still, we can employ what she coins as an 'indirect, long-range control' over these things.

For example, if Tom were informed about his potential implicit bias before evaluating the resumes, he could have employed strategies to avoid its influence later in time. He could, for instance, have asked for anonymized resumes (Madva 2020). Or, as some evidence suggests (Dasgupta and Greenwald 2001), Tom might have been able to effectively weaken the influence of his implicit racial stereotypes by exposing himself, while evaluating the resumes, to counter-stereotypical images or admired Black celebrities. Then, when evaluating the resumes, Tom would still lack direct control over his implicit biases` unintended influence. Though, he would be able to exercise indirect control over them, as he was informed of them prior in time and thus able to employ some strategies of monitoring their influence. (Holroyd and Kelly 2016)

Thus, implicit bias`s influences can be seen as uncontrollable only if we are exclusively concerned with a direct, immediate sense of control. Though, we can indeed be able to control implicit bias in an indirect way. I will come back to this in Section Two.

2.1.d Conclusion - Section One

In Section One, I scrutinized the working condition for implicit bias I employ in this piece step by step.

Implicit biases are usually measured by the IAT and refer to those implicit attitudes that contradict a person`s explicit judgements. I discussed how evidence strongly supports (a) that such implicit biases exist in various domains in the sense that people inhibit them.

Feature (b) entails that implicit biases can influence individual behaviour and contribute to harmful discrimination. This feature is crucial for the present piece. It implies that implicit biases can cause harmful behaviour and therefore matter morally. But it also implies that implicit biases can influence behaviour in the first place, which is relevant for this project in so far as it is concerned with moral responsibility for precisely the negative, harmful *influence* of implicit biases on behaviour. Shedding light on the empirical debate over whether implicit biases can cause behaviour and contribute to harm, I emphasized that whether implicit biases are likely to do so in a certain situation is strongly dependent on different factors and on how they interact: e.g., on the person, such as her capacity for working memory, or on contextual factors, such as the salience of contextual cues. Based on this, I formulated the first condition for the piece: (1) an account of moral responsibility should be flexible enough to accommodate the instability of implicit bias`s influence, which depends on personal and contextual factors and on how they interact.

Feature (c) of the employed working definition concerns a person`s unconsciousness about and inability to control the impact of implicit bias. I asserted that people are mostly unaware of the influence. Furthermore, they are unable to exercise direct, immediate control over the impact of their implicit biases but might well be able to exercise indirect forms of control. Awareness and

control over implicit bias are closely related to the challenge implicit bias poses to traditional control accounts of moral responsibility. I will turn to this challenge in Section Two.

2.2 Implicit Bias - A Challenge for Theories of Moral Responsibility

The second condition and the third condition I propose for this piece are grounded in the challenge implicit bias poses for some accounts in the literature on moral responsibility. More specifically, I argue that implicit bias can challenge 'reason style control accounts' of moral responsibility. This challenge shall ground two additional conditions an account of moral responsibility should accommodate: (2) implicit biases cause harm, and (3) we commonly expect people in certain contexts to be or to become aware of implicit bias and to control their harmful behaviour.

Let us now first have a closer look at what reason style control accounts are (2.a) before I explicate the challenge implicit bias poses to them (2.b) and different potential solutions to it (2.c).

2.2.a What are Reason Versions of Control Accounts of Moral Responsibility?

In the literature on moral responsibility for implicit bias and its influence (I focus on the latter), one can distinguish between three main strands of arguments. Each of them regards one condition that constitutes moral responsibility. The two main strands of arguments are concerned with how a person`s 'awareness' and a person`s 'control' over implicit bias are necessary pre-conditions for her moral responsibility. In this piece, I will concentrate on the control condition (and propose a 'revisionist' approach to it, as I will explain in 2.c). Though, I shall partly relate the discussion to arguments from awareness in so far as awareness is regarded as necessary for control. 'Attributionist' or deep self-views on moral

responsibility for implicit bias form a third main strand of arguments in the literature. I bracket those and focus on control[4]. (Brownstein, 2019)

Why should we focus on control in the first place? As noted by Levy (2017), it is a widely accepted idea in the literature that the exercise of a certain form of control is a requirement for moral responsibility (Levy 2017, 5)[5]. Often, approaches that put the freedom condition of control to the forefront reflect the foundational principle in moral philosophy of "*ought implies can*" (Copp 2008). If an agent failed to do something she ought to do, to be morally responsible for her failure, she must have control over her behaviour. If I ought to attend an online seminar, for me to be culpable for failing to do so, I must also be able to control my laptop, which I need to be able to attend, for example, I must be able to charge it and to switch it on. (Buckwalter 2019)

But importantly, what does it mean to have control over an action? As mentioned, this piece is primarily concerned with reason variants of control-based accounts. Reason style control accounts of moral responsibility include a broad variety of very different views (Fischer and Ravizza 1998; McKenna 2013; Nelkin 2011; Wallace 1994), among which Fischer and Ravizza`s (1998) account is widely regarded as the most influential one (Stout 2016). These very different views have in common the idea that our rational powers are an essential form of capacity for

[4] In brief, deep self-views, also denoted as 'attributionist' views, entail that for an agent to be morally responsible for an action (in the sense of being blameworthy), the action needs to reflect who the agent really is, an agent`s "deep self", and must, in that sense, be attributable to her. For example, Sripada (2015) refers to a person`s deep self as her fundamental evaluative stance, which can be roughly understood as her fundamental character and values. Only if one endorses the fundamental evaluative stance of a racist, which could amount to inhibiting the character of a racist, she can also be responsible for racist actions. Focusing on the control condition, I bracket those views. See Zheng (2016) for a 'revisionist' deep self-view. (Brownstein 2019; Dominguez 2020, 161f)

[5] Levy (2017, 5): "the claim that an agent is morally responsible for an action or for the consequences of an action only if she exercised 'freedom-level' control over that action or that consequence is a condition on responsibility that almost every prominent theorist accepts, in some form or another".

moral responsibility. Whether an agent acted for reasons and whether reasons that are the agent`s own ones are the causes of her behaviour form issues that are central to these approaches.

More precisely, for reason accounts, being able to control an action means disposing of a psychological mechanism that permits one to *recognize and respond to reasons that are relevant for this action*. What does this mean? First, for a person to be in control of her actions, she must be able to *recognize* relevant reasons. Recall the scenario mentioned in the introduction. For example, Tom would need to be able to recognize (e.g., be aware of) the fact that he is influenced by implicit bias, a reason to control his behaviour. Secondly, for Tom to be able to control his behaviour, he must also be able to *respond* to reasons he recognizes. For example, he must be able to control his behaviour and be able to actually act upon his recognition of the fact that he is influenced by implicit bias. Thus, in a reasons account, only if Tom is able to recognize and respond to relevant reasons, he can control his behaviour. (Vargas 2020)

2.2.b The Challenge of Implicit Bias for Reason Style Control Accounts

I argue that implicit bias can challenge such reason style control accounts of moral responsibility. Let me first illustrate this concentrating on a person`s 'direct' ability to control, her ability to exercise control right here and now[6]. In reason style control accounts, 'direct' control does not consist in at least two different categories of cases: first, if a person is unable to *recognize* reasons, such as in a

[6] Later (section 2.b.ii), I will elaborate why accounts that focus on a person`s 'indirect' control ability, if understood as requiring a person`s ability to anticipate, face difficulties as well.

case of lack of information; and second, if a person is unable to *respond* to reasons, such as in a case of volitional impairment. (Vargas 2020)

For brevity, let me illustrate Tom`s lack of direct control as constituted by his lack of *recognition* of reasons[7]. In a reason style control account, an agent is not in direct control if she did not dispose of the information necessary to recognize the relevant consideration. As mentioned, in the scenario of Tom, I grant that he lacks content awareness[8]. Furthermore, as has been established, evidence suggests that it is plausible to assume that people in general, including Tom, lack impact awareness, awareness over the influence of implicit bias on behaviour (which, as suggested by Holroyd (2012), should be our primary concern given that our research question regards moral responsibility for the influence of implicit bias). Thus, we might conclude that Tom`s discriminating is not under his 'direct' control. He lacks the recognition of relevant reasons, as he is simply not aware of his harmful, stereotyped behaviour being influenced by implicit bias while evaluating resumes.

(I ask for the reader`s patience. One might think: yes, but Tom might have avoided the influence by implicit bias before being influenced in a manner out of his direct control. I will return soon to such responses that focus not only on the present but also on the past, in 2.2.c.)

[7] Note that a person`s lack of direct control over implicit bias can also plausibly be derived from her inability to *respond* to reasons. This is because even if Tom was made aware about inhibiting implicit biases, one might as well regard Tom as not having direct, immediate control over his implicit bias`s influence (compare Saul 2013, 55, and see section 2.1.c.ii). Thus, in a reason style control account, Tom would lack his ability to 'respond' to relevant reasons and would be seen as not morally responsible for what he does.

[8] As briefly noted in the previous Section, this must not always be plausible. Someone like Tom might well have sometimes something like awareness over his having/ inhibiting implicit biases (although whether this can constitute self-knowledge can be disputed as well, compare footnote 3). While for the present purposes, a lack of impact awareness seems more decisive, in this piece, for simplicity and illustrative purposes, I greatly grant this lack of content awareness.

Let us now address the challenge which, I argue, implicit bias poses to such reason style control accounts. Should Tom be excused? On the one hand, I argue that it seems so. Namely, what can make implicit bias highly problematic for some control accounts of moral responsibility is that the failure to recognize relevant reasons (interpreted as unawareness) at that moment in time is likely *unintended, non-volitional* (Vargas, 2020). Tom did not intend to be unaware. He simply evaluated the resumes and did not think about the possibility of implicit biases potentially influencing him while doing so.

Why might this be the case, more precisely? For example, because Tom did not make plans to be unaware of it while evaluating resumes. Furthermore, and more crucially, behaviour derived from implicit bias is defined as contradicting a person`s explicit judgements and values. It is defined as behaviour a person disagrees with after deliberating (feature (a) of the working definition). Thus, let me assert that we can plausibly suppose that it was not Tom`s intention to be unaware of and to not have control over the negative influence of implicit bias while evaluating the resumes. He simply was unaware.[9]

Some have taken a similar route of argumentation, attributing importance to impact awareness. As suggested by Jennifer Saul (2013), Tom`s lacking awareness of an implicit bias`s influence can indeed be seen as possibly undermining his ability to directly control his actions and, therefore, his moral responsibility: "[…] a person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture." (Saul 2013, 55)[10].

[9] This is not to say that unawareness about the impact of implicit bias is indeed in all cases non-volitional. For the moment it suffices to assert that in the case of implicit bias, as I argue, we have reasons to assume that it *can* be non-volitional. And these cases form a source of the challenge for some accounts of moral responsibility.

[10] In fact, Cameron et al. (2010) suggest that such a view reflects psychological attitudes of the general public.

On the one hand, it seems implausible to *always* exculpate agents like Tom for the influence of implicit bias due to their lack of awareness. Why?

I argue that this is due to two points that form two conditions which an account of moral responsibility for implicit bias must accommodate. Condition (2), the fact that implicit bias`s influences can cause harm. The influence of Tom`s implicit bias likely contributes to a structural disadvantage for female or racialized Black applicants. I established this in Section One in detail. And condition (3), expectations we commonly have about what other people should know and do about their harmful behaviour given the context they are in. Let me now elaborate on what I mean by condition (3).

Often, we think that people should be aware of something they do not know. We might think that they should recognize relevant reasons, although they don`t. For example, if I forget an important meeting, one might think I fail in a meaningful sense. Similarly, we might commonly think that Tom should be aware of the possibility of being influenced by implicit bias while evaluating the resumes, although he is not. Furthermore, it seems implausible to assert moral responsibility only for those attitudes whose influence falls within our conscious awareness. Any case of negligence, inattention or forgetting would exculpate agents. (Holroyd et al. 2016)

As an alternative, George Sher (2009) proposes that "when someone acts wrongly or foolishly, the question on which his responsibility depends is not whether he is aware that his act is wrong or foolish, but rather whether he should be" (Sher 2009, 20, as cited in Holroyd et al. 2016, 6). Such an approach concentrates not on an individual`s actual awareness but on what is commonly expected from her, such as duties or obligations to be aware.

As mentioned, I grant that Tom is not aware of inhibiting implicit bias (content awareness) and, more importantly, also not aware of being influenced by implicit

bias at the moment he acts (impact awareness). This constitutes his potential lack of direct control. Though, Holroyd (2014, 2016) outlines different ways in which people might (not be but) *become* aware of implicit bias and thus, be *expected to be or become* aware. For instance, Tom could become aware of being influenced by implicit bias by drawing inferences. He might draw conclusions on his probable psychological properties and behaviour by consulting scientific studies. Or he might become aware through the potential harm his actions caused for the neglected applicants, e.g., by questioning his short-list (Levy 2014).

In these or similar senses, I maintain that we might *expect* others to become or to have become aware of the harmful influences by implicit bias they have been or will be exposed to. Why might we expect someone to become aware? As one possible answer, Washington and Kelly (2016) relate the requirement of awareness of implicit bias to the environment. For example, suppose Tom was working on a hiring committee. In his role in this context, he would bear the responsibility not to act in a discriminative manner towards applicants. Washington and Kelly argue that this person should be inferentially aware of implicit bias in this role or environment. Tom should be aware of scientific evidence on implicit bias and able to infer whether his judgements are likely to be biased. Furthermore, given his role, he should be aware of different methods to mitigate implicit biases. Similarly, Holroyd (2016) argues that if scientific evidence about implicit bias is readily available in a context, then a person can also be expected to draw inferences about her own potentially biased behaviour. Therefore, whether non-awareness about implicit bias`s impact can excuse a person such as Tom can depend on the context.

The revisionist account of the responsibility-relevant capacities at the core of this piece shall be related to this idea. In short, it does not seem that a mere lack of awareness about implicit bias and its influences always exculpates agents. Rather, we can think of awareness as something an agent should have, something we can

expect from her. And this obligation to awareness can be context-dependent.[11] (Vargas 2013)

So, let me summarize this challenge which, I argue, implicit bias poses to reason style control accounts, at least in so far as we are concerned with 'direct' control. On the one hand, in the framework of reason accounts, it seems that Tom is not under direct control over his behaviour. We can grant that Tom was unaware of his having and, more importantly, being influenced by implicit biases. In that sense, one might conclude that Tom did not recognize relevant reasons at that moment. Furthermore, Tom *did not intend* this failure to recognize relevant reasons. He did not intend to be unaware of being influenced by implicit bias while evaluating resumes. He simply did not realize it at that moment. On this line, Tom`s lack of direct control being unintended would exculpate him in the framework of reason style (direct) control accounts.

On the other hand, as established in Section One, (2) Tom`s behaviour derived from implicit bias still contributed to harm. And I argued that (3) we might still expect that Tom should know about implicit bias influencing his evaluation of the résumés given the context in which he was. Although he did not know at that moment in time, we might still say that he should have known or found out about implicit bias. And we might still think that he should have exercised *some* sort of control over his harmful behaviour.

This allows me to develop conditions (2) and (3) of this piece. I maintain that a control account of moral responsibility for implicit bias should accommodate that (2) implicit biases cause harm; and that (3) we commonly expect people in certain contexts to be or to become aware of implicit bias and to control their harmful behaviour.

[11] Compare Jay Wallace (1994, 21) for a similar account of "holding a person [responsible] to expectations".

How can this challenge be addressed and conditions (2) and (3) be accommodated while grounding moral responsibility for implicit bias`s influence? Let us now shed light on possible responses to the challenge I have just formulated.

2.2.c People`s Indirect Control - Responses to the Challenge of Implicit Bias

Note that so far, we have been exclusively concerned with a person`s 'direct' control ability. Though, we know that people can 'indirectly' control the influences of implicit bias. I will now briefly sketch two other lines of possible responses which ground Tom`s moral responsibility relating to his ability to *indirectly* control his behaviour. Both responses are variants of reason style control accounts and, in different ways, imply that momentary unawareness and lack of control being unintended must not excuse an agent. Rather than focusing on the present time, the responses look to the past and to the future. The first response takes a more conventional, backwards-looking approach. The second line of response is the one I shall endorse in this piece.

2.2.c.i The Tracing Response

Let us begin with the first, backwards-looking approach and a limitation it faces. An attentive reader might have asked herself continuously: couldn`t Tom have known that implicit bias would influence him and prevented this influence *prior in time* through indirectly controlling it? This is the so-called "tracing response". Following a tracing response, one might insist that in some sense, unawareness being unintended at that moment in time is irrelevant. It is irrelevant if his momentary unawareness and resulting lack of direct control is in a certain sense

still his fault because it can be ‘traced’ back to a *prior* failure of direct control for which he is morally responsible. (Vargas, 2020)

And, as I have elaborated in 2.1.c.ii, a person can in different ways ‘indirectly’ control her implicit biases (Holroyd 2012; Sie & Vader-Bours 2016). On this line, in the framework of a reasons account, the effectiveness of indirect control strategies shows that a person`s implicitly biased behaviour can indeed be responsive to reasons and under control despite the involvement of unconscious cognitive processes during the display of implicitly biased behaviour (Holroyd 2012, 295-96). For illustration, consider the penalty kick of a football player (compare Arpaly's (2002, 52) similar argument and description of a fast-paced tennis match). Just as in Tom`s case, it is true that conscious, reflective cognitive processes are unlikely to direct the football player`s kick. And if so, then she would probably miss the goal. However, the cognitive processes involved in the production of the shot are clearly responsive to reasons and are under control. For example, how to approach the ball, in what corner to look right before making the shot etc., are reasons the player can recognize and respond to before making the shot. Similarly, in the case of Tom`s implicit bias, he could have adopted indirect control strategies before evaluating the resumes. For example, he could have anonymized the resumes before evaluating them. Indirect control strategies would have allowed Tom to respond to reasons before being influenced and despite being unaware while being influenced.

Importantly, I will not neglect that indirect control strategies can be effective for preventing harmful behaviour. Instead, let me restrict my scepticism to how we aim to ground moral responsibility for implicit bias`s influence. If we ground being morally responsible in a more conventional, backwards-looking sense through such tracing responses, such an approach can, I assert, be limited. The limitation is set by what Michael McKenna denotes as our “epistemic radar” (McKenna 2012, 191). Namely, a tracing response requires that in a state in which

Tom is under direct control over his behaviour and thus morally responsible, he is, in fact, able to anticipate implicit bias`s future influence. Though, someone like Tom might not have been able to anticipate and avoid every instance of such not directly controllable influence. For example, he might not expect to be evaluating resumes in a state of direct control, e.g., earlier the month when a friend told him during dinner about implicit bias. And later that month, Tom might have simply forgotten about it. He, who has never evaluated resumes before, might have simply found some lying on his desk and started evaluating them. He did not intend to forget. He simply forgot. (Vargas, 2020)

Let me establish that these epistemic limits seem even more relevant for a theory of moral responsibility for implicit bias if we consider everyday encounters that cannot always be anticipated, such as a person passing by on the street. For grounding moral responsibility despite a lack of direct control, a tracing response would need to show precisely Tom`s ability to anticipate every later situation in which he is negatively influenced by implicit bias.

This piece shall explicitly defend the effectiveness of indirect strategies of control over implicit bias. Concerning the question of how a person`s moral responsibility can be grounded, though, backwards-looking tracing responses can be limited by our epistemic radar and excuse agents pro tanto for exhibiting implicit bias. This, I assert, can be problematic for a theory of moral responsibility. It leaves unaddressed the challenge of implicit bias I formulated. Tom`s behaviour (2) causes harm, and (3) we might still expect him to do something about it or to be aware of it, given his role as someone who hires others. Given the context in which Tom acts, the roles he takes up, we have expectations towards him.

This piece aims to employ an alternative, forward-looking strategy to ground a person`s ability to control the influence of implicit bias and, therefore, her moral responsibility for it.

2.2.c.ii Vargas` Revisionist, Forward-Looking Response – a Perspective on What Will Follow

If we do not intend to lack control and not to recognize that we act in a stereotyped, harmful way, caused by implicit bias; if it might also not *always* make sense to regress to our ability to anticipate stereotypic, implicitly biased behaviour in a previous state of control: why and when are we morally responsible?

For solving this question, this piece proposes a revisionist approach to control and moral responsibility. While revisionist approaches start out from and do not aim to completely abandon our existing and prevalent concepts and intuitions about moral responsibility, they also entail revisions of these mostly backwards-looking intuitions. It is argued that phenomena such as implicit bias reveal that our backwards-looking intuitions about moral responsibility, as well as more conventional theories which often endorse them, are inadequate to provide a satisfying account of moral responsibility. Thus, traditional accounts should be revised and pre-philosophical common-sense intuitions about moral responsibility partly abandoned. (Dominguez 2020, 164f)

For example, Glasgow (2016) and Faucher (2016) provide such revisionist accounts of moral responsibility for implicit bias. Suggesting partial revisions of more conventional accounts, they propose that conditions of moral responsibility should not be stable but instead vary across contexts and persons. For instance, according to Faucher (ibid.), conditions of moral responsibility should vary in their significance with how an agent relates to the harm she causes. For a victim of perpetuated harm, these conditions (such as whether she has a certain degree

of awareness or acts intentionally) can have a different significance for the assessment of her moral responsibility than for someone else.[12]

This piece scrutinizes Vargas` (2013) forward-looking control account as one of these recent revisionist approaches[13]. As mentioned, I focus on what it means to have control and to be, therefore, morally responsible for implicitly biased behaviour. And Vargas explicitly engages with control. For Vargas, it is not a person`s intention to lack direct control or to be influenced by implicit bias at that moment in time that grounds her moral responsibility. Instead, he frames his account as 'capacitarian' (Vargas 2020). A person`s failure to exercise a capacity she has is what grounds her being a responsible agent and her potential blameworthiness. Thus, Vargas` forward-looking response avoids the epistemic limits backwards-looking tracing responses face. Furthermore, and unlike what I described in 2.2.b, it entails that Tom can have the control capacity to recognize the influence of implicit bias even while being unaware of this capacity or the influence of implicit bias while evaluating resumes.

Now, one might ask: can such a response be defensible? Why should Tom be morally responsible for the influence of his implicit bias even if he is unaware of the influence at that moment in time and even if he did not intend it nor his failure to recognize it? The short answer to this refers to the future and to what makes the approach forward-looking: because moral responsibility is valuable in its effects. Blaming will generally and in a forward-looking sense support Tom in cultivating his capacity to not discriminate and thus, to avoid harm in the future. The long answer to what it means to have circumstantial control that grounds one`s moral

[12] Other revisionist approaches to moral responsibility for implicit bias are, e.g., the proposals of Zheng (2016) or Mason (2018) (compare Part Five).

[13] For a more precise classification, although I shall focus on the implications of Vargas` view in accounting for (1), (2), and (3), Vargas` approach considers incompatibilist elements and arguments due to which revisions about our ordinary thinking about moral responsibility and free will can be necessary. On the other hand, his approach is compatibilist, as it implies moral responsibility/ free will to be compatible with causal determinism. (Timpe 2014, 927)

responsibility and how such an account can accommodate the three conditions I formulated will follow in Part Three. (Vargas 2013)

2.3 Conclusion - Section Two and Part Two

When and in what sense are people morally responsible for implicit bias`s influence? In this Part, while limiting the project's scope, I formulated three conditions that specify what a control account of moral responsibility for implicit bias should accommodate when addressing this research question successfully.

First, while evidence strongly suggests that implicit biases exist, their influence on behaviour is volatile. Their influence depends on the behaviour in question, on personal and contextual factors, and on how they interact. Thus, I established (1) that an account of moral responsibility for implicit bias should be flexible enough to accommodate the high instability and context-dependency of the influence of implicit bias.

Second, I argued that implicit bias challenges reason style control accounts of moral responsibility. These accounts conclude that an agent is not in direct control over her behaviour if she is unable to *recognize* relevant reasons. I granted that Tom is unaware of his having implicit biases, and even if he was, he is plausibly unaware of his being influenced at the moment he acted. Thus, he can be seen as not being in direct control over his behaviour derived from implicit bias. Furthermore, Tom`s unawareness and lack of direct control were not intended while evaluating resumes. I maintained that for reason style control accounts, in so far as concerned with direct control only, these are reasons to excuse Tom.

On the other hand, I established that always excusing someone like Tom seems implausible. The reasons for this formed condition (2) and (3) of this piece. An account of moral responsibility for implicit bias should accommodate: (2) implicit biases cause harm; and (3) we commonly expect people in certain contexts to be or to become aware of implicit bias and to control their harmful behaviour.

Anticipating how control accounts of moral responsibility for implicit bias accommodate this through referring to an indirect control ability, I argued that

instead of trying to trace back behaviour to a previous state of direct control, we should adopt a less conventional, forward-looking strategy. In Part Three, I will argue that Vargas' forward-looking, circumstantial view on responsible agency capable to control implicit bias is promising to accommodate conditions (1), (2), and (3) and can thus largely address the research question of this piece.

PART THREE: WHAT CONTROL CAPACITIES MAKE US RESPONSIBLE AGENTS? - VARGAS´ CIRCUMSTANTIALISM

Is Tom morally responsible for his implicitly biased behaviour? Tom is unaware of acting in a stereotyped way derived from implicit bias. He did not intend or plan to be unaware. He simply was. And on reflective deliberation, he would strongly disagree with the stereotyped behaviour. On the other hand, he still caused harm. And we might expect him to act differently in his role of engaging in hiring others. When and in what sense are people such as Tom morally responsible for behaviour that is caused by implicit bias?

As anticipated, I will propose a forward-looking response to this research question and to the challenge I formulated in the previous Part. The purpose of this Part is twofold. First, it introduces Manuel Vargas` (2013) view on what capacities for control morally responsible agents must have, and the crucial role circumstances and expectations play in that view. Second, by doing so, and more crucially for the purposes of the present piece, I will argue that Vargas` account is a promising one for addressing the research question 'when and in what sense are people morally responsible for the influence of implicit bias'. I will establish this by showing that Vargas` circumstantialism can accommodate the conditions I developed in Part Two: (1) the context-dependency of implicit bias`s influence; (2) the necessity to avoid harm; and (3) our expectations regarding this avoidance of harmful behaviour.

In Section One, I will introduce Vargas` (2013) general approach and argue that it accommodates condition (2) in that it recognizes the necessity to avoid harm induced by implicit bias`s influence. Subsequently, I will argue that Vargas`

circumstantial view accommodates condition (1), the instability and context-dependency of implicit bias's influence. As I will show, this is because it conceives of morally responsible agency as a function of the agent and her circumstances. In Section Two, I will argue that Vargas` circumstantial view is promising to accommodate condition (3) as well. I will establish that in two different ways, it crucially recognizes that we often expect people to be aware of harmful behaviour and to act differently in certain contexts.

3.1 Circumstantial Capacities as a Function of the Agent and Her Circumstances

As I will establish in this Section, whether an agent has circumstantial control-capacities does not depend only on the agent in isolation from her circumstances. Crucially, having them also depends on how the agent relates to the circumstances. I will argue that this makes Vargas` circumstantialism a suitable candidate to accommodate condition (1), the peculiarity of implicit bias as highly unstable and dependent on personal and contextual factors. But before I explain why this is so in greater detail (1.b), I will first (1.a) set the stage for this piece, introducing Vargas` general forward-looking Agency Cultivation Model, which entails his circumstantial view (henceforth: CIV). This shall serve to establish how Vargas accounts for condition (2) in that it moves the necessity to avoid harm induced by implicit bias to the foreground.

3.1.a The Justification Thesis and Vargas` Agency Cultivation Model

Before I begin: what is Vargas concerned with in the first place? For Vargas (2013, 309), to say that people *are* morally responsible exclusively means that they are liable to praise or blame, whereby I bracket praiseworthiness[14]. Vargas`

[14] Some additional restrictions of the scope of this piece are in place. Whenever I mean responsibility, I mean moral responsibility. Also, note that Vargas (2013, 309f) is concerned with *being* morally responsible, and our *being* morally responsible depends on whether our *holding* each other responsible is justifiable. In this sense, Vargas combines the notions: our holding each other responsible (in the sense of blaming) requires our being responsible agents (see Part Four for more details on what blaming means). The distinction between the notions shall figure in Part Five as well: I will propose that some of our *holding* each other morally responsible might be justifiable even though we *are not* morally responsible agents (in the sense of blameworthy) in a circumstantial view.

exclusive focus on (praise-and) blameworthiness leaves a gap which I will explain and fill in Part Five.

Let me now start introducing Vargas` (2013) forward-looking view while establishing how it can account for condition (2), the necessity to avoid harm induced by implicit bias. The first important aspect of it: it is based on the acceptance of the justification thesis. This thesis boils down to the idea that our social practices of responsibility, our blaming each other, need to be generally justified. Thus, for Vargas, we need to ask: why is it generally justified to blame each other, or a person such as Tom, for implicitly biased behaviour?

Importantly for the present piece, note that for Vargas (2013), the justification of our *general* practice of holding each other morally responsible should be thought of as on a different level than the justification of our holding each other responsible *in specific instances*. In specific cases, we justify our blaming or praising Tom not necessarily on forward-looking grounds (Vargas 2013, 172: "I think [this view] is dead wrong"). Instead, we justify specific instances of blaming or praising each other based on the content of our norms which can even be backward-looking. For example, on the grounds of what Tom failed to do in the past. This underlies Vargas` (2013) claim of his forward-looking approach being compatible with our backward-looking intuitions in our practice.

For an illustration of this claim, Vargas (2018, 118) refers to foul calls in a football game. The system of norms and the practice of foul calls are generally justified by their overall contribution to a forward-looking aim, such as to protect the players and to provide the fans with an enjoyable game. Though, a particular instance of foul calling in the last minute might not make the game more enjoyable or protect the players. Specific instances are not directly justified through their specific contribution to the forward-looking aim. Rather, a particular foul call is justified on another level, referring to the content of norms. And this content might well be backward-looking, e.g., concerning mistakes a player made.

In this piece, we are concerned with the general, forward-looking justification of responsibility practices. Though, as explained, Vargas claims that this does not rule out the possibility of a justification of a particular instance through the (potentially) backward-looking content of norms.

It is worth emphasising this focus. In this piece, we are concerned with *blaming in general someone like Tom*. Decisive for the justification of an overall system of praising and blaming each other in a certain context, our blaming *someone like Tom*, is the overall accumulated (forward-looking) effectiveness of such a system of holding each other morally responsible. This is important to keep in mind for fully grasping what it means for a system of responsibility practices (whereby its norms can be backwards-looking in content as well) to be justifiable and effective in a forward-looking sense.[15]

Now, for Vargas, the short answer to what makes blaming generally justifiable is its general effectiveness in making us "better" agents who are morally responsible and thereby capable of avoiding harm (condition (2)). But let us have a closer look. This brings us to the first feature of his general forward-looking approach and his Agency Cultivation Model (comp. Vargas 2018, 119f.):

(1) our social practices of holding one another morally responsible are *justifiable* if they cultivate a valuable kind of agency, one that is sensitive to moral considerations.

Vargas defines a "moral consideration" as "a consideration with moral significance such that, were one to deliberate about what to do, it ought to play a role in those deliberations." (Vargas 2013, 203). Thus, it can be understood as a

[15] While I endorse but do not explicitly defend this distinction (indeed, it has been criticised by McGeer (2015) or Harland (2020)), it is worth to keep in mind that we are concerned with this more general level of justification.

consideration in the form of a reason that, on reflection, has normative significance (depending on the underlying normative ethical theory).

To be sensitive to such moral considerations means to be capable of both recognizing and responding to reasons that have normative significance. As mentioned in Part Two, for example, it means that Tom can recognize, in some way or the other, that implicit bias influences him in a manner that causes him to harm others. And it means that he can respond to reasons in the sense that he can direct his behaviour in light of this consideration. Therefore, as explained, adopting a reason style control account, saying that an agent is sensitive to moral considerations, simply refers to the idea of her having a capacity to recognize and respond to reasons, and thus, to control her behaviour. (comp. Vargas 2013, 203f.)

Why should cultivating such an agency be *valuable*? Very briefly, Vargas (2013) specifies different reasons for this. People experience that they can align their behaviour to what they or others value and expect from them, and in that sense, experience control and can avoid harm. They regard themselves and are seen by others as responsible agents, as trustworthy and reliable, and as responsible for their potentially harmful behaviour (I will come back to this in 3.2.b).

Let us now make the discussion more concrete by applying all this to the case of Tom: (1) whether the general social practice of holding people like Tom morally responsible for their implicit biases in this or a comparable situation in which someone like him would be evaluating resumes is *justifiable* depends on its overall contribution to a forward-looking aim. That is, would blaming someone like Tom generally contribute to making people like him "better" agents who are overall capable of discriminating less? In other words: would blaming someone like Tom contribute to the aim of cultivating a valuable sort of agency that is sensitive to moral considerations? Does it make people like him more able to recognize and direct their behaviour according to the consideration that discriminating against others on the grounds of, e.g., their gender, is unjust?

This is the first feature of Vargas` forward-looking agency cultivation model. And I argue that this is how such a view accommodates (2) and contributes to the avoidance of harm caused by implicit bias`s influences. For Vargas (2013), the justification of our social moral responsibility practices and blaming each other stems from their overall contribution to a forward-looking aim. Simply put, blaming is justified because it is generally effective in cultivating sensitive agency, in making people like Tom better beings who are capable to control, e.g., the harmful behaviour caused by implicit bias.

But crucially, what does it mean to be a responsible agent? What does it mean to have a capacity for control to, e.g., avoid harmful behaviour caused by implicit bias? The specification of these responsibility-relevant capacities forms the second feature of Vargas` forward-looking Agency Cultivation Model. At this point, his CIV and how he accepts the justification thesis come in (Vargas 2018, 119f.):

(2) Vargas upholds a circumstantial view on what the capacities to be sensitive to moral considerations, which are relevant for responsible agency, consist of.

In Vargas` view, for a person to be a responsible agent in a specific situation, she must dispose of 'circumstantial' control capacities (I will soon, in 3.1.b, explain what they are). And someone like Tom must possess these circumstantial capacities *above a certain level* to count as a responsible agent while evaluating resumes. The specification of this level is where (1) and (2) connect (this point shall figure later in 3.1.c and 3.2 as well). The level of capacities someone like Tom must have to count as a responsible agent is determined by the forward-looking aim to cultivate sensitive agency. (Vargas 2013, 219)

Again, what does all this mean more concretely? Back to Tom: (2) The CIV specifies the sort of capacities that determine whether Tom is a responsible agent

able to control implicit bias. Thereby, the necessary level of reason-responsiveness that people like Tom must possess to count as responsible agents in such a situation is determined by what level overall contributes most effectively to (1) the forward-looking, justifying aim of the practice. And this aim is to make someone like Tom more able to control his conduct, to recognize and act according to the moral consideration of, e.g., not discriminating against others and to avoid harm.

This is Vargas` (2013) forward-looking approach and his Agency Cultivation Model. It moves the forward-looking aim of cultivating responsible agency capable to control harmful behaviour caused by implicit bias to the fore. I argue that this is how it accommodates condition (2), the necessity to prevent harm induced by implicit bias`s influence. The cultivation of capacities to control harmful behaviour is what generally justifies our responsibility practices of blaming each other and, thus, whether we can be thought of as responsible agents in the first place (having these capacities to control at a minimal level). While I will come back to how Vargas does not ignore the role of expectations for our avoiding harmful behaviour in certain contexts in Section Two, we have now set the stage.

But what is at the core of the present piece is still unclear yet: what are ‘circumstantial’ control capacities for Vargas?

3.1.b Vargas` Circumstantial View on Responsible Agency

I will now make clear how precisely Vargas (2013, 2018) can account for circumstances as being a direct part of an agent`s ability. Thereby, I will show how his CIV can accommodate condition (1), applying Vargas` understanding of circumstantial capacities to empirical evidence for the instability and dependence

of implicit bias`s influence on personal and circumstantial factors (as discussed in Part Two).

What are circumstantial capacities more in detail? Vargas (2013 204f) maintains that our rational capabilities are not content-neutral or cross-situationally stable. Instead, he upholds that whether psychological processes function well is context-dependent (in that sense, "ecological"), and different capacities function better or worse in different contexts. Thus, for Vargas, relevant capacities for responsible agency are relational and plural. They are relational as they form a "function of whether the agent (with the relevant features in the considered context of action) stands in a particular relationship to the normative practice" (Vargas 2015, 2622). They are plural, as a plurality of mental capacities or psychological structures constitute whether an agent is suitably able to respond to reasons. Noteworthy is Vargas concern not with idealized agents, but with the abilities of agents in the actual world, given their imperfect cognition and psychologies. Furthermore, having a circumstantial capacity does not necessarily require a person`s awareness of it. She can still have it while not being aware of that. (Vargas, 2013, 2018)

In sum, Tom`s (scenario 1) circumstantial capacity to control what he does (in the sense of being able to recognize and act upon the moral consideration not to discriminate) is unstable in different contexts and depends on various factors. A circumstantial capacity (a person must not be aware of having) can be understood as a function of features internal to the real agent and features of her circumstances. Such a control capacity shifts if epistemic environments shift. (Vargas, 2013)

And this, I argue, suitably accommodates condition (1) empirical findings on the instability and dependence of the influence of implicit bias on other factors, such as personal and circumstantial ones.

First, having a circumstantial capacity depends on personal factors. These can include Tom`s intrinsic commitments and other attitudes, such as his (context-independent) values or character. And as we have seen in Part Two, whether implicit biases influence behaviour in a specific situation and whether people might be able to control it can indeed depend on personal factors. For example, as mentioned, it has been found that a person`s capacity for working memory is associated with her ability to control the influence of implicit attitudes on her behaviour (Friese et al. 2008).

Second, a circumstantial capacity is relational. Having it can depend on contextual factors that are present while Tom evaluates resumes. And, as has been found, contextual factors such as the presence of contextual cues and how a person relates to them can indeed play a role in whether someone is influenced by implicit bias or not and in whether she might be able to exercise control. For example, as mentioned, Gschwendner et al. (2008) indicate that changing the background from a Mosque to a garden on a computer screen can affect how faces commonly associated with German or Turkish stereotypes are implicitly evaluated. Master et al. (2016) found that the decoration of computer-science classrooms can impact whether female students state to be interested in computer science. In neutrally decorated rooms, female students were up to three times more likely to express interest than in rooms decorated with objects commonly associated with science fiction or video games.

In addition to this, a circumstantial capacity is plural as Tom might need different capacities or psychological functions to be able to avoid discrimination when he evaluates the resumes. And this, I argue, is again in line with empirical evidence on the influence of implicit bias. As mentioned, evidence suggests that these influences and an ability to control them can depend on multiple factors such as a person`s capacities (e.g., her working memory (Friese et al. 2008)), but also psychological functions such as the condition of a person to control her impulsive

behaviour (Cameron et al.`s 2012). Furthermore, recall Cesario et al. (2010), who analysed "personality-by-implicit-bias-by-situation interactions". They found that differences in individuals` confrontational personalities affected the influence of situational factors on the manifestation of implicit biases.

On this line, let me establish that Vargas` (2013) circumstantial view on control capacities accommodates condition (1), the empirical evidence on the dependence of implicit bias`s influence and people`s control on personal and circumstantial factors and their relation. Precisely like evidence on implicit bias`s influence, having a circumstantial capacity depends on agential and circumstantial characteristics and on how they relate. This is one reason why Vargas` circumstantial view on capacities, which morally responsible agents must have, can fruitfully address the research question and the issue of moral responsibility for implicit bias.

Now, let me add to this. For Vargas (2013, 219), how the relevant threshold of circumstantial capacities is determined again accommodates condition (1). The threshold of necessary reason responsiveness must be so set that it enables a system of blame and praise to function well to achieve its effects. Here, Vargas (2013, 214f.) refers to the idea of an 'ideal observer' who picks the *context-specific* degree of reason-responsiveness that is best able to support us in doing generally the right thing. There cannot be an answer about what level of capacities an agent must possess to be a responsible agent that is general and independent of the context or agents. I argue that this again reflects how Vargas' account can accommodate condition (1), the dependence of implicit bias`s influence on personal and contextual factors. Given that, simply put, the relevant capacities are a function of the agent and her circumstances, the verdict on the context-specific degree of reason-responsiveness will consequently vary depending on the circumstance or agent`s general characteristics. (Vargas 2015, 2622)

3.1.c Conclusion - Section One

Whether Tom is a responsible agent depends on his circumstantial capacities. Blaming him is justified if, in general, it fosters such circumstantial capacities to recognize and respond to moral considerations as a form of reasons, such as the consideration to avoid discrimination through implicit bias`s influence. I argued that Vargas` circumstantial view is promising in accommodating condition (1) of this piece, the empirical evidence on the dependence of the influence of and control over implicit bias on personal and contextual factors, and on how they relate. This is because Vargas` circumstantial capacities consist in a function of how Tom`s internal abilities (such as his capacity for working memory) relate to the circumstance (such as the presence of contextual cues that make stereotypes salient or not). Furthermore, this is because the responsibility relevant threshold of reason-responsiveness varies over contexts and persons, dependent on the forward-looking aim. (Vargas, 2013, 2018)

3.2 The Role of Expectations in Vargas` Circumstantial View

I am now going to establish that Vargas` (2013) circumstantial view can, to some extent, accommodate condition (3) as well. His account attributes a crucial role to (3) the fact that we often expect others to know and do something about stereotyped harmful behaviour caused by implicit bias. I will establish this in two different ways: first (in Section 3.2.a), I argue that (i) expectations of others in a context that concern our harmful behaviour are forms of reasons. They are, therefore, what circumstantial capacities are concerned with in the first place. Second (in Section 3.2.b), I argue that (ii) these expectations play a decisive role in avoiding harm. They partially constitute whether blaming is generally effective in the avoidance of harm and, thus, generally justified. In Part Four, defending Vargas` view, I will come back to (i) the role that expectations play in Vargas` account and to empirical evidence on the effectiveness of blaming in the context of implicit bias.

3.2.a Why Circumstantial Capacities are Sensitive to Expectations

I will now specify that what we expect others to do in certain situations crucially matters for circumstantial capacities to be in place. Namely, as a first way in which Vargas accommodates condition (3), (i) expectations can be what circumstantial capacities are concerned with. Circumstantial capacities are "relational" (Vargas 2013, 206) in that they include not only Tom himself in a narrow, atomistic sense, but also his relation to his environment, including present expectations towards him. More specifically, circumstantial capacities are relational also in the sense that they are 'interest-sensitive', sensitive to expectations. We can think of whether someone has the capacity to be sensitive to moral considerations as depending on the interests/ expectations and their content, which are in place in

the relevant circumstance. For example (compare Vargas 2020, 408), if Tom was sleeping and one expects him to avoid the manifestation of his implicit bias right here and now, while still being asleep (e.g., through deleting the pictures on the resumes), Vargas regards Tom as not being able to do so. But if the content of one`s expectation of what Tom should do was different, e.g., entailing that Tom avoids implicit bias as soon as he wakes up, Tom might well possess this ability. Following this approach, it becomes evident how circumstantial capacities are 'interest-sensitive'/ sensitive to expectations: a circumstantial capacity is concerned with one`s ability to act in line with reasons such as the interests of others and their expectations present in a context. Thus, having such a capacity is 'sensitive' to/ depends on the content of expectations. (Vargas, 2013, 2017)

At this point, recall Washington and Kelly`s (2016) argument I sketched in the previous Part. Coupling the requirement of awareness of implicit bias with the relevant context, they argue that the responsibility-relevant question is not whether Tom is aware of implicit bias`s harmful influence, but whether someone like him should be. Given the role one endorses in a context, for example, Tom`s role of hiring trainers as an owner of gyms, or an employee's role of being on a hiring committee, it might well be that one should be or have become (e.g., inferentially) aware of the potential harmful influence of implicit bias. In their view, the content of expectations that are present in a context and associated with relevant roles can greatly determine whether someone should be aware of and, therefore, morally responsible for implicit bias`s harmful influence. On this line, I formulated condition (3) a control account should accommodate: we might have strong expectations towards people to be aware of and to control implicit biases (even if they were unaware).

Now, I hope to have shown that Vargas` (2013) account of circumstantial capacities is promising to accommodate condition (3). (i) Conceptually, circumstantial capacities do not ignore the presence of our expectations about

implicit bias`s harmful influences. Instead, these expectations are precisely instances of reasons with which relational, interest-sensitive, circumstantial capacities are concerned with. In a certain context, we can have circumstantial capacities to be more or less sensitive to these expectations and their content. Therefore, we can account for Washington and Kelly`s (2016) argument. I established that in Vargas` view, the presence of expectations in contexts and their content (such as that one should be aware of implicit bias) can partly determine whether we can have circumstantial, interest-sensitive control-capacities in the first place and can, therefore, be morally responsible. I will come back to this in Part Four as well.

3.2.b Expectations Determine the Effectiveness of Blaming

Another reason that makes Vargas` (2013) account a promising candidate for accommodating condition (3) to a considerable extent is (ii) the role Vargas ascribes to expectations and the internalization of norms for whether blaming can be effective and thus, justified.

To show that for Vargas, expectations matter for the effectiveness of blaming, let us slightly modify the scenario of Tom: suppose now that Tom is minimally aware that he has "something like" an implicit bias. Should someone like Tom be blamed? As anticipated in Part Two, some empirical evidence (e.g., Madva 2018) suggests that this modification is plausible. Content awareness about the possibility of inhibiting/ having something like implicit bias might, in some contexts, indeed be widespread. Though, even then, as Vargas stresses, the effectiveness of blaming is far from guaranteed. Importantly, Vargas emphasizes the role that norm internalization plays. 'Internalizing norms' can be understood as having a feeling of what is expected from someone, more or less consciously (Vargas, 2020). While awareness about the phenomenon of implicit bias might be

in place, social norms that entail expectations about implicit biases and one`s ability to control and self-monitor them are, according to Vargas, still not widely internalized. Therefore, blaming is often likely to not be effective in reducing discrimination. (Vargas, 2017) This is the reason for which, according to Vargas, someone like Tom might not always be a responsible agent[16]. As Vargas argues elsewhere: "Internalization of norms of praise and blame is key. Without internalization of such norms, it is hard to see how actual practices could be suitably stable and reliable enough to yield the relevant result." (Vargas 2013, 175). (Vargas, 2017)

Why is the internalization of expectations, having a feeling of what is expected from someone given her context-specific role, so crucial for making blaming effective? Some argue that this is because if one has a feeling for what is expected, she then understands the blaming, and will not respond defensively or with hostility (Saul 2013, 55). Similarly, one might suggest that the presence and sufficient awareness/ internalization of relevant expectations of others is what is required to direct one`s behaviour in a social context in the long term and in a more stable manner. Otherwise, one simply cannot know what is expected and, importantly, does not know how to interpret the behaviour of others.

Vargas` (2020) conception of social self-governance serves as another possible explanation that further specifies (ii) why expectations in contexts can matter for blaming to be generally effective in fostering control and avoiding harm. Namely, Vargas argues that it is valuable to us to be regarded as a reliable, self-governing, and self-controlling person in what we do, and regarded as trustworthy, given the roles we take up in specific contexts. Whether Tom is regarded as a trustworthy

[16] Recall: not a responsible agent in the sense that he lacks the required circumstantial control-capacities that would make him one. And the ineffectiveness of blaming in contributing to the forward-looking aim of the practice determines whether someone like him lacks such a minimal level of required control-capacities, the required level of reason-responsiveness (again, this is how (1) and (2) in Vargas` Agency Cultivation Model connect). (Vargas 2013)

owner of gyms who hires new trainers partly depends on what expectations the role Tom has in this context implies, what people commonly expect from an owner in this business. In short, according to Vargas (ibid.), expectations coupled to the roles we have right here and now are simply what often matters to us in our behaviour. It matters for someone like Tom that others recognize him as competent to live up to relevant expectations. This can explain why sometimes, a general practice of blaming can be ineffective in fostering control and reducing the harm caused by implicit bias. According to Vargas, in some contexts, (indirectly) controlling implicit bias might not be something that someone like Tom relates to being competent and reliable in what he does in this context. (Vargas, 2017)

Drawing on Vargas (2020), I have now specified (ii) why expectations can make blaming effective or ineffective to avoid implicit bias. While people value being seen as competent in their context-specific roles, expectations related to their roles do not (yet) always include the avoidance of implicitly biased behaviour, which can undermine the effectiveness of blaming. (Vargas 2017, 2020)

Note that the present discussion focuses on condition (3.ii), why expectations can matter for blaming to be effective in fostering avoidance of implicit bias. Of course, taking this focus does not rule out that reasons other than her present or future circumstances, such as a person`s inner moral code, could still effectively steer her behaviour and control over implicit bias and influence the effectiveness and justifiability of blaming. Indeed, Amodio et al. (2007) suggest that guilt, which might also be affected by a person`s inner moral code, can be an effective regulator of implicit bias. Rather, I specified (ii) that a lack of internalizing expectations represents *one* explanation for a potential ineffectiveness of blaming. Again, this is because we often simply care about being seen as competent in living up to the roles we have.

3.3 Conclusion - Part Three

In this Part, I aimed to introduce Vargas` general forward-looking account and circumstantialism. How can we make sense of implicit biases, which are (1) flexible, context-dependent, and (2) induce harm, while at the same time, (3) we expect others to know about and to control this harmful implicitly biased behaviour given their roles?

In Section One, I set the stage for the piece by presenting Vargas` forward-looking Agency Cultivation Model and the prominent role condition (2), the avoidance of harm, plays in it. Vargas addresses the research question on responsibility for implicit bias by focusing on the effects our praising and blaming (whereby I focus on blaming) can generally have.

Furthermore, I specified how, in his view, control-capacities responsible agents have are 'circumstantial', a function of the agent and her circumstances, and how this accommodates condition (1). It provides a flexible account of moral responsibility that can live up to the empirical findings of implicit bias`s influences and control. Having circumstantial control capacities depends on the agent, the circumstances, and on how these factors relate.

In Section Two, I focused on condition (3), the role of expectations towards other`s avoidance of harmful behaviour in Vargas` account. I have specified two ways in which expectations matter for Vargas: (i) they are accounted for in what circumstantial capacities are, namely, having circumstantial capacities is sensitive to interests and expectations present in a context (as argued, from whomever they stem); and (ii) expectations and the internalization of norms matters for the effectiveness of blaming. It can crucially matter whether people generally feel what is expected from them in a context to make blaming *effective* against implicit bias. Thereby, one explanation Vargas provides is that we simply care about being seen as competent in fulfilling our roles.

This is why Vargas` (2013) account seems promising to address moral responsibility for implicit bias`s influence in cases such as Tom`s. Based on this, we can formulate the following *provisional* answer to the research question: people are morally responsible for implicit bias`s harmful influence, in the sense of being blameworthy, if moral responsibility practices contribute to cultivating their (potentially minimal) circumstantial capacity to control it and, thus, to avoid harm.

But how successful is this account? Let us scrutinize how, more specifically, Vargas accommodates condition (3), Jules Holroyd`s (2018) critique of Vargas (2017), and what blaming or non-blaming can even mean in practice in Part Four.

PART FOUR: DEFENDING VARGAS` CIRCUMSTANTIAL VIEW ON CONTROL

As noted, a capacitarian account tackles the question of what it means to have capacities to do things in specific contexts which make a person morally responsible if she fails to do these things. If Tom had the circumstantial capacity to control harmful influences by implicit bias, he would also be morally responsible for his failure to exercise such a capacity. And, as explained, whether someone like Tom is thought of as having such a circumstantial capacity is greatly determined on forward-looking grounds. It largely depends on whether blaming (or praising, though I focus on blaming) someone like him would make him more sensitive to implicit bias and his ability to exercise control. (Vargas, 2013, 2017)

Now, is someone like Tom morally responsible for his implicitly biased behaviour? Vargas (2017) asserts that mostly he is, but *in some situations he is not*. Blaming people like Tom might, in some contexts in which people are not yet sufficiently sensitive to implicit bias, sadly not yet contribute to cultivating capacities to discriminate less. This verdict, which I will explain in more detail soon, can seem provocative, especially since (2) the behaviour was harmful.

Jules Holroyd (2018) sees a problem in Vargas` (2013, 2017) forward-looking view being 'circumstantial'. Simply put, what if the 'current' circumstances themselves are problematic. What if Tom lived in an unjust current context that does not entail the expectation to avoid implicit bias? As I will explain, Tom might be able to optimistically expect the future to be more just and sensitive to implicit bias. For instance, he might be able to hope for a shift in social norms, for more critical future clients of his gym that judge his present discriminatory conduct while current clients are less critical. What about the future? What about the things we might be able to hope for? Thus, is Vargas` circumstantial control account

suitable for addressing moral responsibility for implicit bias and condition (3) if it is bound to our 'current' circumstances only? On this line, Holroyd argues that Victoria McGeer`s (2015) scaffolded responsiveness view (henceforth: SRV) on responsible agency should be preferred over Vargas` circumstantial view. According to Holroyd, the SRV better serves the forward-looking aim to cultivate responsible agency and is more in line with both conventional theories and our existent responsibility practices than the CIV. (Holroyd 2018; Vargas 2013; McGeer 2015; McGeer and Pettit 2015)

In this Part, I defend Vargas` Agency Cultivation Model and his CIV. This defence shall serve me as a narrative to specify how Vargas` account accommodates condition (3), (i) regarding whose expectations can matter in a circumstantial view and what blaming can even mean in practice. I will primarily focus on responding to Holroyd`s (2018) critique that the SRV likely better serves the forward-looking aim than the CIV. Then, building on this response, I will also briefly outline implications of it for Holroyd`s assessment of what view is more in line with both conventional theories and our existent responsibility practices.

Let me begin by providing an overview of the argument of Holroyd (2018, 149f.) that is at the core of this Part:

> Premise 1a: only if a view on responsible agency can account for indirect reasons does it entail that someone like Tom is responsible/blameworthy for the influence of implicit bias.
>
> Premise 1b: the CIV cannot account for indirect reasons, while the SRV can.
>
> Sub-conclusion: therefore, the CIV`s verdict is that Tom is *not* a blameworthy/ responsible agent in the case of implicit bias, while the SRV`s verdict is that he is blameworthy/ responsible.
>
> Premise 2a: a view on responsible agency should serve the aim of cultivating agency that is sensitive to implicit bias's influence.

> Premise 2b: blaming someone like Tom can be effective in serving the aim of cultivating sensitive agency.
>
> Conclusion: therefore, the SRV with its verdict of blameworthiness has better prospects to serve the aim of cultivating sensitive agency than the CIV.

In two separate steps, I will reject both Premise 1b and Holroyd`s conclusion.

First (4.1, 4.2), I argue that Premise 1b of Holroyd`s (2018) argument is false. This defence shall serve me as a narrative to interpretatively specify Vargas` view on (i) whose expectations might matter in a circumstance. I will argue that, interpretatively, the notion of a circumstance is not restricted to people we encounter in a circumstance right here and now. While Holroyd asserts that the CIV cannot account for an agent`s capacity to be sensitive to what she coins as 'indirect reasons' (Holroyd 2018, 143), I argue that, interpretatively, this does not seem too obvious. Interpretatively, the CIV extends to expectations that stem from people we will encounter in the future (indirect reasons). Thus, I will specify the first way Vargas accommodates condition (3): (i), whose expectations are at issue can matter in the CIV.

In a second step (4.3), I will show that even if Premise 1b were true (and I was wrong), Holroyd`s conclusion would still not follow from the Sub-conclusion, Premises 2a, and 2b. This is because the CIV does not preclude other forms of moral responses that differ from blaming, which can also be effective in serving the cultivation of sensitive agency. This argument shall serve me to discuss what blaming and non-blaming can amount to in practice. I will come back to this in Part Five.

Furthermore (4.4), what I have established shall allow me to briefly extend the defence of the CIV in relation to two additional criteria evoked by Holroyd.

4.1 Indirect Reasons – When the Reason Lies in the Future

Let me start by explaining what indirect reasons are and why, for Holroyd (2018), only McGeers' SRV but not Vargas` CIV can account for them (before I question this conclusion in 4.2).

What if we live in an unjust society, but we can still hope for a change? How can a circumstantial view accommodate this if it is, as the name suggests, bound to the current circumstances in which we live? According to Holroyd, one crucial difference between Vargas` and McGeer`s approaches to moral responsibility lies in how they accept the justification thesis[17]. In brief, McGeer does not share Vargas` circumstantial view on valuable capacities relevant for responsible agency sensitive to moral considerations. Instead, she proposes a scaffolded responsiveness view (SRV). (Holroyd 2018; Vargas 2013; McGeer 2015, 2637)

Holroyd (2018) argues that McGeer`s conception of scaffolded responsiveness to reasons can be 'indirect' as well, and not only 'direct', as Holroyd interprets that it is in Vargas` CIV (I will question that). Scaffolded reason-responsiveness is indirect in the sense that whether a person is sensitive to moral considerations is also a question of whether she is able "to adjust or sensitize to the reasons that there may be*"* (Holroyd 2018, 143). McGeer and Pettit (2015) specify what such an indirect form of reason responsiveness might amount to by evoking the notion of the "prospective audience" (ibid., 170). It is not only reasons and the audience we encounter directly in a specific present situation we can be responsive to. But instead, responsiveness to reasons that 'there may be' extends to the anticipation of a future audience whose judgements about our present behaviour matter to us.

Let me interpret such a sensitivity to a prospective audience and their judgements of our present conduct as an expectation. Although our current audience, whose

[17] Noteworthy are the many similarities between Vargas` and McGeer`s approaches. As Vargas, McGeer follows a revisionist approach and accepts the justification thesis.

judgements matter to us, might be uncritical and unsupportive, under a certain optimistic idealization, one might still be sensitive to reasons that stem from an audience one can (optimistically[18]) expect to encounter in the future, given the present context. (Holroyd 2018)

For a better understanding, consider again the scenario of Tom. But let us suppose now that Tom could (under an optimistic view) expect future clients, who inhibit different norms and expectations than his current audience, to judge him for discriminating the applicants. For example, Tom can expect prospective clients to be reluctant to subscribe to a gym where only White male trainers work. Tom is still not in a context where his *current* audience is critical and would have judged his behaviour in a way so that valuable, sensitive agency could develop. The current clients will not judge him negatively and are unsupportive towards his efforts to control implicit bias. But after adding this detail, it becomes clear that Tom could still, in some way, expect to be in a cultural context where norms against discrimination are more prevalent. And if Tom cared about his future clients` opinions and expectations (or any judgement from a person he will encounter in the future given his current context), these expectations could also represent reasons Tom might be more or less sensitive to. In other words, if Tom was able to expect future clients or others to judge him negatively if he did not prevent his implicit bias now, this might well be a reason for him to control his implicit bias now. (Holroyd 2018)

For Holroyd (2018), that the CIV cannot conceptually account for such indirect reasons, expectations of a future audience, is a crucial shortcoming of the CIV. In her reading, as noted, circumstantial capacities are restricted to the ability to "here and now […] register and act on certain reasons" (Holroyd 2018, 143). And this implies, in her reading, that the CIV cannot account for expectations that stem

[18] Holroyd (2018) emphasizes the role of optimism which is involved in the ability to be sensitive to indirect reasons, according to the SRV.

from people we will encounter. For Holroyd, in the CIV, only the present audience we encounter in a situation can induce reasons we can be effectively sensitive to in a context, right here and now. I will now show why I disagree with this conclusion.

4.2 Why the CIV Can Account for Indirect Reasons

Now, I will interpretatively specify a first way in which Vargas accommodates condition (3), namely, (i) whose expectations one can be sensitive to in Vargas` circumstantial view. I will show that Vargas can indeed account for indirect reasons. In my interpretation, at least in this point, his view is not as conceptually problematic as Holroyd (2018) suggests.

In short, I argue that if we can interpret indirect reasons as expectations of judgements that stem from future audiences, then they will crucially matter in Vargas` (2013) CIV. As explicated in Part Two, a circumstantial capacity is 'interest-sensitive'. And, I argue interpretatively, such sensitivity might also well extend to expectations of a future audience one can optimistically expect given one`s context. This is because, in my view, it is not convincing to interpretatively draw a strict line on how a 'context' should be interpreted on a *time dimension,* excluding the future. It seems plausible that reasons in a context can include expectations from future audiences, but which are nevertheless dependent on the contexts we are in now. If Tom would expect different clients in his gym tomorrow, in 10 days, or in a year: these expectations might still affect his present conduct in the context of him owning a gym. This must not occur in a much different manner to how someone like him can be sensitive to expectations from present clients in the context of being a gym owner. If Tom can (optimistically) expect his future clients to judge him negatively if he, as a gym owner, hired only male, White trainers now, Tom might still aim to live up to these 'indirect' expectations of future clients now. For example, if he wanted to stay in the market

and remain a gym owner. In contrast, if Tom was in a different context now, for instance, not an owner of a chain of fitness studios, he will also not be faced with expectations of present or future clients of gyms. In that sense, the presence of indirect reasons is closely coupled to the context Tom is in right now. Note that different clients are only an example of a potential future audience. Nothing precludes that Tom could have expectations concerning judgements of different future applicants or even of a different and more critical future self. A variety of future audiences might form the source of indirect reasons. For the moment, though, let me establish that these indirect expectations can still be seen as *context-dependent* reasons, which the CIV *conceptually* accommodates. (Vargas, 2013, 2015b)

Therefore, Holroyd`s (2018) Premise 1b is, in my interpretation, false. Taking stock, I argued that Vargas` (2013) concept of the CIV can account for the potential presence of indirect reasons, expectations from people who we *will* encounter in a context. This specifies (i) the first way in which Vargas accommodates condition (3), the role expectations play in his account of being in control and morally responsible. Contrary to Premise 1b, I interpretatively asserted that *whose* expectations might matter in a context in Vargas` circumstantial view can extend to people we do not currently encounter in that context. But we might well be sensitive to expectations from people who we *will* encounter, given the context we are placed in right now.

Although here, I focus on 'indirect' reasons and on the extent to which Vargas can accommodate the role of expectations in our holding each other responsible, let me add a slightly different point. McGeer (2015, 2646-48) objects that the SRV distinguishes itself from the CIV in a somewhat different sense than it is emphasized by Holroyd (2018). Namely, McGeer argues that the SRV but not the CIV implies that people can 'learn' what they could not do before. In her interpretation, only the SRV considers the case that someone can be insensitive to

a consideration now, while it would still be justified to blame (someone like) her if she could become sensitive to that consideration in the future. Saying that people have a capacity to do something can, for McGeer, also mean that "even though they don't have any competence at all, they could develop it with the requisite training and hard work" (McGeer 2015, 2646). In other words, being responsive to reasons in a scaffolded sense is the "[…] kind of capacity that is sensitive to the on-line scaffolding effects of praise and blame; in effect, it is to be sensitizable to moral considerations one failed to be sensitive before." (ibid., 2647). And McGeer asserts that Vargas` CIV cannot account for such a case of learning, this being an essential difference between the views. She classifies the CIV as "atemporal" (ibid. 2646), requiring capacities to be in place right here and now for blaming practices to be justifiable.

I argue that Vargas` (2013) Agency Cultivation Model, which entails the CIV, can likely address this additional objection as well. Indeed, McGeer`s (2015, 2646) claim that the CIV would be "atemporal" is not invulnerable. This becomes evident considering the broader picture of Vargas` account I illustrated and evaluated in Part Three. Recall that for Vargas, the forward-looking aim determines the minimal level of reason responsiveness relevant for responsible agency. If blaming someone like Tom would make him more sensitive to the harmful influence of implicit bias than before, blaming someone like him would be justified. We would regard someone like him as having a *minimal level* of circumstantial capacities to respond to reasons. Thus, we can maintain that the CIV is not clearly or explicitly 'atemporal' and could accommodate 'learning' as well. In a context, but not necessarily restricted to our current contexts only, it is the forward-looking consideration of whether it is minimally effective to treat someone as morally responsible which determines her being a morally responsible agent. And in Vargas´ view, it is not strictly ruled out that this effectiveness could not also extend well over time, but within contexts. In Vargas` framework, it does not seem precluded that this effectiveness could not also mean making someone

discriminate less or learn to discriminate less in a context over time and to be therefore thought of as disposing of a minimal level of reason-responsiveness now.

I have defended Vargas` (2013) CIV and questioned conceptual differences to the SRV concerning 'indirect reasons' or 'learning'.[19] As became evident, in Vargas` account, more decisive than these conceptual questions is a forward-looking, *empirical* one to which I will turn in the next Section (4.3). Whether blaming someone like Tom is justified crucially depends on its general effectiveness for making him an agent better able to control, which is an empirical question.

4.3 What It Means to Blame and Not to Blame, in Theory and Practice

I will now show why Holroyd`s (2018) conclusion would still not follow even if Premise 1b were true. For the sake of argument, let us assume that the SRV, due to its accounting for possible indirect reasons, would be more likely to entail the judgment that someone like Tom is blameworthy than the CIV. Thus, for the moment, we keep in mind: CIV=non-blaming; SRV=blaming (this shall be sufficient for the present purposes).

Even then, though, I argue, Holroyd`s (2018) conclusion that the SRV (blaming) can better contribute to the forward-looking aim than the CIV (non-blaming) still does not follow. As I will make evident, blaming (the SRV) must not necessarily be *more* effective in contributing to the forward-looking aim than non-blaming (the CIV). Overall, as I will argue, the evidence remains mixed.

[19] Another difference between their views seems to be generally more decisive, although not for the present purposes. In brief, and as becomes apparent in McGeer (2015) and Vargas` (2015a) reply, McGeer rejects Vargas` 2-level-justification. Given that Holroyd (2018) does not discuss this difference in her assessment and given that this discussion would go far beyond the present argument, I bracket this contrast as well.

Thereby, responding to Holroyd in this second way, my main aim is to discuss what it means for Vargas and others to approach a person with blame or another moral response *in practice,* a discussion I will come back to in Part Five.

Let us begin by discussing what 'to blame' or 'not to blame' someone even means. As noted, Vargas regards the phenomenon of holding each other morally responsible as being liable to blame. In a nutshell, according to Vargas (2013, 118f.), for blaming to be an instance of holding someone morally responsible, it requires a judgement of the other person to be (1) the right sort of agent and (2) to deserve a blame characteristic reactive attitude because she has done something of moral significance. On top of that, in the usual case, to blame someone also means (3), to express a blame characteristic reactive attitude, such as one of "resentment, indignation, and the like [, …] verbal condemnation, calls for censure or shame, and more common forms of reaction such as avoidance, emotional distance, or retractions of interpersonal warmth." (Vargas 2013, 119). (Vargas 2013, 116-21, 2017)

In a recent paper, Vargas describes how it is useful to think of evaluative attitudes as on a spectrum. On the one end, there are judgements of "better" and "worse", and on the other end, judgements of blameworthiness/ culpability (Vargas 2020, 413). For example, characterological judgements such as "Jim is bad at mathematics" (Vargas 2020, 412) must not necessarily be instances of blaming. Those judgements must not, e.g., necessarily imply that Jim would (2) deserve the blame-characteristic reactive attitude of condemnation or usually induce (3) the expression of such an attitude. (Vargas, 2020)

We have briefly discussed what it means to hold someone blameworthy for Vargas (2013). Let us get more concrete. Let me now engage with two sources of empirical evidence Holroyd (2018) refers to in support of her argument in favour of the SRV (blaming) and against the CIV (non-blaming) (recall that in the previous Section, I rejected this differentiation of Holroyd between the verdicts

of the views. Here, I adopt it merely for the sake of the present argument). Thereby, defending the CIV, I shall establish what blaming and non-blaming mean in practice. First, I will have a closer look at Czopp et al. (2006). Subjects who had manifested implicit bias were confronted with messages of two different types[20]: (Czopp et al. 2006, 788):

> *"Low threat: but maybe it would be good to think about Blacks in other ways that are a little more fair? it just seems that a lot of times Blacks don't get equal treatment in our society. you know what I mean?"*
>
> *"High threat: but you should really try to think about Blacks in other ways that are less prejudiced. it just seems that you sound like some kind of racist to me. you know what i mean?"*

Czopp et al. (2006, 799) found that accusing subjects harshly with racism (high threat message) was *just as effective in reducing implicitly biased responses* as the message that merely entailed a plea for fairness (low threat).

The high threat message clearly counts as an instance of blaming (accusing the subject of (1) being racist and of (2) deserving the moral response). Notably, though, drawing on Vargas` classification, I interpret that the low threat message does not count as blaming. Neither is it evident that (3) a blame-characteristic reactive attitude would have been involved, such as condemnation or resentment. Nor is (2) a judgement that the addressed person would personally deserve the (not involved) reactive attitude present. It seems more similar to something like "Jim is bad at mathematics" (Vargas 2020, 412). And this non-blaming has been found to be as effective as a blaming response.

[20] In the first of three experiments, which is the most relevant one for present purposes. The "confrontation" employed in this first study consisted in confronting individuals with "the fact that their egalitarian self-concept was inconsistent with their prejudiced values, attitudes, and behaviours, [and, as a result of this,] they experienced feelings of self-dissatisfaction." (Czopp et al. 2006, 785).

But if the low threat message does not count as ‘blaming’, what is it then? Vargas (2018, 31) refers to aretaic or axiological evaluations as moral responses that are different to blaming. Crucially, even in the absence of morally responsible agency, such aretaic moral responses are not precluded by Vargas` CIV. As I will argue in Part Five more in detail, the CIV is merely concerned with the justifiability of blame and praise. It does not preclude other moral responses.

In a different paper, Holroyd et al. (2017) refer to “aretaic appraisals” (Watson (1996), as cited in Holroyd et al. (2017)) as a form of judgment that is not one of blameworthiness. Instead, it is “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” (Holroyd et al. 2017, 5). And indeed, we have reasons to classify the effective low threat message as an aretaic moral response. The low threat message evaluated whether the confronted person satisfied the value of fairness she herself presumably aspired to, without asserting that her bad performance in this aspiration would have been her fault or that she is racist.

In sum, I showed in detail why the low threat message is not an instance of blaming (as defined by Vargas (2013, 118)). Instead, I argued that it can be understood as an aretaic moral response. Importantly, such an aretaic response, which invokes the universal value of fairness and promotes a person to aspire to it, has been found to be possibly as effective as a blaming response to counter implicit bias. And crucially, the CIV does not preclude such aretaic responses being merely concerned with the justifiability of blame and praise. Therefore, while I made the discussion more concrete, I showed that the conclusion of Holroyd (2018) that the SRV is more effective than the CIV does not necessarily follow from the Sub-conclusion and Premises 2a and 2b. Blaming (SRV) need not be more effective than non-blaming/ aretaic moral responses (CIV). And this is what Holroyd (ibid.) would need to show for her conclusion that the SRV serves the forward-looking aim better than the CIV to follow.

Finally, after having discussed what blaming and non-blaming can amount to in practice, what might underlie the empirical ineffectiveness of blaming? Recall what I established in Part Three (ii). Again, one explanation relates to expectations. We, and someone like Tom, might want to be seen as competent in fulfilling our roles in the contexts we act in. And a lack of norm-internalization, if we do not relate an obligation or interest to control implicit bias to such roles, can be a reason for why we might generally, in some situations, still react reluctantly (Vargas 2017). Now, note that Scaife et al. (including Holroyd herself) (2020), the second source of evidence evoked by Holroyd (2018), ascribe significant importance to norm-internalization as well. Scaife et al. (ibid.) admit that whether blaming or not blaming is more effective ultimately remains a debated empirical question[21]. It could well be the case that "self-reported changes in explicit intentions are due to social desirability effects rather than individuals internalising the relevant moral norms" (Scaife et al. 2020, 8). And only in the latter sense, they admit, the 'expressions of intentions' to control implicitly biased behaviour will actually lead to such a behaviour change. In that sense, Scaife and Holroyd et al. seem to agree that normative expectations present in contexts and their internalization can be decisive for making blaming effective in fostering control.

[21] Compare Scaife et al. (2020, 1). Scaife et al. (2020, 8): "Further analyses of the relative efficacy of communicating negative but non-moralised feedback versus negative but moralised feedback is required."
Importantly, note at this point that my aim is not to emphasize that blaming is not generally effective to combat implicit bias. Rather, my aim is to emphasize that this remains a disputed empirical question (Czopp et al., 2006; Scaife et al., 2020; Saul, 2013). I will come back to this in Part Five and to why, on this line, Vargas` (2013) account of moral responsibility is limited in not justifying such effective moral responses.

4.4 Extending the Assessment – Coherence with Conventional Theories and Social Practice

I have just critically evaluated Holroyd`s (2018) argument concerning the CIV`s contribution to the forward-looking aim of cultivating responsible agency. Based on what we have established, we are now in the position to defend Vargas` circumstantial view on control also with regards to two additional criteria evoked by Holroyd.

First, according to Holroyd (2018, 150-51), McGeer`s (2015) SRV is not only better to serve the forward-looking aim of the practice. It is also *"independently more plausible"* than the CIV. Both Vargas` (2013) and McGeers` approaches to moral responsibility belong to the revisionist camp. As I have discussed, revisionist approaches start out from our pre-philosophical intuitions while allowing for revisions. Note that revisionist approaches entail the commitment to only make 'necessary' revisions. They aim to abandon and revise our existing concepts and backwards-looking intuitions only in so far as necessary. This should ensure, as coined by Holroyd, their 'independent plausibility'. Now, most of these more conventional accounts of moral responsibility do not neglect that responsible agency might be in place. Indeed, as Holroyd asserts, most conventional theories are not even concerned with whether responsible agency would be in place but with the different question of what kind of moral response is appropriate. And according to Holroyd, only the verdict of the CIV neglects that someone like Tom would be a responsible agent, not the SRV`s verdict. Along this line, Holroyd argues that the SRV is independently more plausible than the CIV. The SRV does not entail this major departure from conventional concepts and intuitions, neglecting responsible agency to be in place, while, in her view, the CIV does. Holroyd adds to this, arguing that moral responses "directed toward the agent[s] for their failure to behave as they were expected to." (Holroyd 2018, 150) seem defensible while the CIV`s verdict rules out responsible agency.

This, she argues, makes the SRV again independently more plausible. It entails that Tom is morally responsible and can be addressed with such defensible moral responses.

Evoking a second but similar criterion, Holroyd (2018, 151-54) assesses the views` *"coherence with practice"* (rather than with conventional theories that entail common intuition, as before). Holroyd argues that the SRV better captures our existing practices of holding each other morally responsible for implicit bias than the CIV. In our holding each other morally responsible every day, most of the time, we interact with others while assuming them to be morally responsible agents. We usually occupy what Strawson (1962) coined as "the participant standpoint"[22]. Revisionist approaches attempt to safeguard such common practices of conceiving each other as responsible agents from deterministic scepticism. This attempt is justified in recognition of the value responsibility practices have for us, such as the valuable cultivation of control capacities. Therefore, according to Holroyd, the SRV, with its different verdict, coheres more with our existent responsibility practices of assuming responsible agency than the CIV, with its ruling which neglects such responsible agency to be in place.

Based on what we established before, we can now again question Holroyd`s assessment. I assert that the CIV and the SRV are likely to perform similarly for both criteria. First, I interpretatively argued that the CIV and the SRV do not vary clearly in their accommodating for indirect reasons or learning. This makes the CIV as independently plausible and coherent with existent responsibility practices as the SRV. Being there no obvious difference between the views in their accounting for indirect reasons or learning, it is not obvious that their verdicts on Tom being or not being a responsible agent would differ either. Furthermore, even if the verdicts differed, I made explicit that if Tom were not a responsible agent

[22] Compare Maureen Sie`s (2009, 2014, 2018) "Traffic Participation View on Human Agency", which is partly inspired by Strawson (1962) and which I discuss in 5.3.

in Vargas` view, the CIV would still not preclude that someone like Tom could be addressed by non-blaming moral responses that entail a person`s failure to comply with expectations towards her. This is a moral response Holroyd (2018, 150) regards as independently plausible and coherent with the practice for a case such as Tom`s. And it is a response which the CIV does not preclude. Thus, again, it does not seem that the CIV would necessarily be independently less plausible or cohere less with existent practices than the SRV.

In this Section, I extended the defence of Vargas` (2013) CIV to additional criteria. Importantly for what shall follow, let me establish a significant limitation of the CIV and Vargas` revisionist account that becomes apparent. The CIV allows for responsible agency not to be in place (which, I argued, is not obviously different for the SRV). This is a thought Holroyd describes as detached from both conventional theories and existent practices. At the same time, revisionist approaches aim at keeping their detachment from the latter two as small as possible (Holroyd 2018, 150-54). While I made explicit that the CIV does not preclude other non-blame responses towards (for Vargas) non-responsible agents, it still holds that Vargas` revisionist view does not account for these non-blame responses. This, I argue, can limit his revisionist approach given that it aims at not being wholly detached from existent practices and existing theories. It does not provide an understanding of the functions of these other moral responses while they are part of our actual, everyday practices. And Vargas does not give a justification of non-blame responses while more conventional theories of moral responsibility are partly concerned with them. In Part Five, I will further discuss and address this significant limitation of Vargas` revisionist account.

4.5 Conclusion - Part Four

In Part Two, I argued that a control account of moral responsibility for implicit bias should accommodate condition (3), the role of expectations in practices of moral responsibility. What if the 'current' circumstances are precisely part of the problem? Through defending Vargas` 'circumstantial' account of responsible agency against critique that arises out of this worry, I have now further specified how the CIV is promising to accommodate condition (3).

In a first step (4.1, 4.2), I interpretatively amended the concept of Vargas` circumstantial capacities regarding (i) *whose* expectations a person can have the circumstantial capacity to be sensitive to in a context. Someone like Tom might well have a circumstantial capacity to be sensitive to expectations that stem from people he could hope to be judged by in the future. They might, in the future, judge his present failure to control implicit bias given the context he is in currently. Thus, I defended Vargas` CIV against Holroyd`s critique and showed that Premise 1b of her argument is false. I argued that the notion of a context should not be roughly restricted to the present time.

Instead, what seems to be more decisive for a person to be responsible is the effectiveness of blaming to serve the forward-looking aim of becoming capable of avoiding discrimination through implicit bias. In 4.3, I described what blaming and non-blaming mean in practice and showed why Holroyd`s (2018) conclusion does not follow.

Based on this, in 4.4., I extended the defence of Vargas` CIV to two additional criteria evoked by Holroyd and anticipated a significant limitation of Vargas` account.

We can now answer the research question in a more *specified* manner, greatly (although, as we shall see in Part Five, still not fully) accommodating condition (3). (i) Someone like Tom can also be morally responsible for implicit bias if this

makes someone like him better able to live up to expectations he can expect to encounter at a later point in time. On the other hand, as established in Part Three, (ii) someone like Tom might be less likely to be morally responsible for implicit bias if people in his context have internalized norms that are uncritical towards implicit bias. Empirically, blaming might not be generally effective to cultivate a control capacity to avoid its harmful influence.

But if our blaming each other might sometimes not contribute to making us persons who are better able to control, and therefore not be responsible agents for Vargas, what can be done? In Part Five, I will come back to this worry, formulate a vital limitation of Vargas` view, and show how it can be addressed.

PART FIVE: DESIGNING NORMATIVE EXPECTATIONS – THE LIMITS OF VARGAS´ ACCOUNT

Our initial worry in Part Four was the following: what if the problem lies precisely in our 'current' contexts (Holroyd 2018)? Suppose Tom`s current context simply entails the 'wrong' expectations. His current clients or colleagues do not relate the expectation of avoiding implicit bias with his role of being a gym owner. Nevertheless, he might still be able to hope for a more critical audience in the future. It might be possible to cultivate his ability to prevent implicitly biased behaviour by making him more sensitive to how a future audience could judge him (to what we denoted as an 'indirect reason'). Is Vargas` (2013) circumstantial view bound to our 'current' circumstances and unable to accommodate changes we might be able to hope for?

In Part Four, I have shown that the CIV can conceptually account for this (whereby I specified how the CIV accommodates condition (3), the role of expectations in our holding each other morally responsible for implicit bias). People can have (i) circumstantial capacities to 'better' expectations that stem from the future and that are more supportive. Instead (ii), empirically, blaming can, due to the internalization of these 'wrong' expectations that matter to people in some contexts, sadly be ineffective. This empirical ineffectiveness in the cultivation of an ability to avoid harmful behaviour is what, I argued, is decisive for Vargas` verdict of someone like Tom not being a responsible agent, rather than a conceptual limitation of the view. Sometimes, someone like Tom might not be a morally responsible agent for his implicitly biased behaviour. This is because blaming him might not always make him better capable to control what he does and to avoid harm. Thus, it lacks justification. (Vargas, 2013, 2017)

This assertion might leave us again with a slightly different worry, though. Even if, in this sense, people were not morally responsible for their implicitly biased behaviour, according to Vargas (2013), what can we do then to avoid the harm they cause[23] (Ciurria 2019)? This worry shall crucially figure in what follows in this Part. As we saw in Part Four, an alternative is moral responses other than blaming, such as reminding each other of universal values we aim to achieve. We saw that such non-blaming responses might effectively lead us to avoid our harmful, implicitly biased behaviour as well. Though, as we shall see now, Vargas` Agency Cultivation Model and CIV are conceptually limited in that respect. Namely, Vargas does not provide a justification for these moral responses other than blaming. How can we justify and make sense of them, if not within Vargas` existing account?

In Section 5.1, I will argue that we should not abandon Vargas` account by justifying blaming and non-blaming responses in the same way. After defending my proposal (5.1.a), I shall critically concede (5.1.b) that the present approach is strongly limited as well. It is mainly concerned with a person`s deserving blaming responses due to having control. Thus, it cannot completely remedy harm induced by implicit bias (recall condition (2), the necessity to avoid harm).

In the next step, in Sections 5.2 and 5.3, I will propose to fill the 'gap' the Agency Cultivation Model leaves (by not justifying non-blaming responses) in a way that does not abandon Vargas` account. I will propose to *supplement* it through Maureen Sie`s (2014, 2018) view on an additional function of our moral responses, the collective design of the content of normative expectations.

[23] Let me remind the reader at this point of the two-level justification of Vargas´s (2013) approach (recall Part Three). It requires blaming to be *generally* ineffective (and thus not justifiable in a context). While it might well be impossible to cultivate the capacities of a *particular*, disrespectful and uncritical person, blaming her would still be *generally* justifiable. This means that the relevant worrisome case might be, in light of this, less frequent than if we were concerned with particular cases only.

5.1 The Liability Assumption and Possible Ways Forward

A critical limitation of Vargas` (2013) account stems from its exclusive concern with moral responsibility as blameworthiness. Vargas upholds the "liability assumption [which entails] that to be a responsible agent is to be liable to praise or blame" (Holroyd 2018, 153).[24] I argue that Vargas` exclusive concern with blameworthiness limits his Agency Cultivation Model. It simply does not provide an understanding of the function or justification of *other moral responses than blaming*. And this is relevant given that, as established in Part Four, other moral responses than blame-characteristic ones, such as aretaic responses, can be effective as well in contributing to the avoidance of harm (Czopp et al. 2006). Therefore, finding a way to justify these moral responses within an account of moral responsibility can contribute to accommodating condition (2) of this piece, recognizing the necessity to avoid harm induced by implicit bias.

Furthermore, this limitation appears even more significant given the revisionist commitment to keep the detachment from both conventional theories and existent responsibility practices small (compare Part Four). Conventional theories are greatly concerned with what moral responses are appropriate, and non-blame responses do form a part of our responsibility practices (Holroyd 2018). But still, Vargas` revisionist view does not account for non-blame responses. This, I argue, can again be seen as a limitation for a revisionist approach that aims to start out and not be totally detached from existing theories and existent social practice.

So, how can these non-blaming responses be justified within an account of moral responsibility for implicit bias's influence? What are some possible ways forward?

[24] Vargas (2013, 309): "We should reserve the phrase 'is morally responsible' for cases in which moral praise and blame can arise".

5.1.a Possibility One: Withdraw the Liability Assumption. And Why We Should Not Do So.

A first possibility would be, as suggested by Jules Holroyd (2018), to withdraw the liability assumption and not equate being morally responsible with being liable to praise or blame anymore. This would mean that we would, to a certain extent, abandon, rather than supplement, Vargas` (2013) approach. To be a responsible agent would then amount to be a proper target of not only blame or praise but also of non-blame-characteristic moral responses.

Let me begin by proposing not to go in this direction and not to withdraw the liability assumption. In brief, I argue that we should not justify aretaic moral responses in the same way as blaming responses because the former do not require the sort of justification that we developed so far for the latter. Aretaic responses do not require or entail that the addressee deserves the moral response because she has, in some sense, control. Thus, such a proposal risks, I argue, bypassing our control-focussed approach to the research question[25]. Let me explain this more in detail.

In short, our approach to the research question is 'desert-oriented'. This means that it focuses on whether we deserve a moral response because implicitly biased behaviour is, in some sense, in *our* control. More precisely, the question I aimed to address was whether, despite our unintended unawareness, our lack of direct control, and the possible limits of anticipation (compare Part Two), we can still be thought of as having a certain sense of control over implicit bias`s influence that constitutes our deserving blame and our *being* morally responsible for it.

[25] A similar worry, though, in a different context, has recently been expressed by Dominguez (2020).

Note that given this approach to the research question and the focus on the control condition, we are *not primarily* concerned with the avoidance of harm induced by implicit bias. Although the questions coincide with the present approach, it is not our primary concern whether we *should be* morally responsible because this could contribute to diminishing harm, independently of whether we *are*. (Sie 2018)

I addressed the research question through Vargas` (2013, 2017, 2020) forward-looking account. It regards our *being* morally responsible as requiring circumstantial capacities to control that are shaped (among others) by expectations. As noted, his account relates to the different question of whether we *should be* morally responsible for implicit bias. It does so, taking the presence of our expectations about what people should know and do into account in his understanding of circumstantial control capacities. Though, while doing so, Vargas` account does not completely abandon the question that motivated our piece in the first place, namely, whether we *are* morally responsible for implicit bias`s influence because we can control it.

I argue that withdrawing the liability assumption, the assumption that to be a responsible agent means to be liable to blame or praise only (Holroyd 2018, 153), risks abandoning this question. The reason for this lies in what aretaic moral responses are, and in what sets them apart from blaming responses. As discussed in Part Four, an aretaic moral response can, as has been described, merely consist in reminding the addressee of a universal value she aims to live up to. For example: “but maybe it would be good to think about Blacks in other ways that are a little more fair?” (Czopp et al. 2006, 788). An aretaic moral response does not entail (2) that an agent deserves a moral response because a failure to do what is right is, in some way, the agent`s failure, being it under her control or attributable to her as a person (such as to her racist character). This sets aretaic moral responses apart from blaming-characteristic-ones, which, in contrast, do entail that the behaviour is, in some sense, the agent`s, e.g., under her control or

attributable to her racist character (as an example, recall: “it just seems that you sound like some kind of racist to me. you know what i mean?” (Czopp et al. 2006, 788)). This is why Vargas (2013) is not concerned with such aretaic responses, given that they do not entail a judgement of desert due to a personal failure or the agent *being* morally responsible *as a person*. Aretaic moral responses are more in line with statements such as “Jim is bad at mathematics” (Vargas 2020, 412), which do not entail that this was Jim`s fault and under his control.

Summing up, I argue that we should not withdraw the liability assumption completely. We would justify non-blaming responses in the same way as blaming ones (in terms of general effectiveness for cultivating control capacities). But then, non-blaming responses would be justified based on a conception of responsible agency capable to control her behaviour that they do not require. Withdrawing the liability assumption would risk, I argue, bypassing the necessity to address the puzzling question we started with. We would risk circumventing the question of whether implicitly biased behaviour is in some sense in our control, *our fault*, and whether we, therefore, deserve to be addressed through moral responses.

5.1.b Limitations of the Present, Desert-Oriented Approach

At this point, though, I must also concede that our desert-oriented approach and control-focused account are very narrow and significantly limited in addressing condition (2), harm induced by implicit bias. This is because we focus on whether a person can be thought of as in control of her implicitly biased behaviour and as, *therefore,* morally responsible for it. With this approach, we are *not primarily* concerned with the harm induced by implicitly biased behaviour. It is true that condition (2), recognizing the necessity to avoid harm, figures prominently in our approach to moral responsibility. Nonetheless, we are primarily concerned with

the question of whether people (like him) *are* morally responsible for this harm *in the sense of having control over implicit bias* and, therefore, deserve the moral response. And I concede that this desert-oriented approach is insufficient to fully remedy and completely avoid harm induced by implicit bias in all its forms. (Vargas, 2013; Sie, 2018)

This limitation becomes even more pressing given that there may be good reasons for justifying our holding each other morally responsible not based on our deserving such a response due to a failure of ours (e.g., to exercise control we have), but merely based on the mere necessity to avoid harm induced. For example, Ciurria (2019, 57-63) specifies this line of critique. It may be that blaming contributes to the avoidance of harm even if we *are not* responsible agents for implicit bias (according to Vargas` CIV and Agency Cultivation Model). Even if blaming, in some contexts, induces backfire-effects, does not cultivate responsible agency, and is thus not justifiable according to Vargas (2018), Ciurria (2019, 57-63) argues that for different reasons, it can still be defensible that we *should be held* morally responsible, and that we should be blamed. When it comes to implicit bias, we might, as Ciurria (ibid.) argues, not be concerned with cultivating responsible agency of the addressees at all. Blaming can still be justifiable on other grounds, such as the overall avoidance of harm or the shaping of power structures.[26]

Zheng`s (2016) view on 'accountability' for implicit bias or Mason`s (2018) view on 'taking' responsibility for implicit bias are revisionist approaches to moral responsibility that follow similar strategies. On the same line, Sie (2018) proposed to generally shift our attention to the design aspect as a justification of moral responses, irrespective of our deserving the moral response. These strategies share

[26] Note that Vargas` (2013) desert-oriented approach is limited in other ways as well, due to which it does not fully accommodate condition (2) and the different forms of harm induced by implicit bias. For example, it is not concerned with other people than the addressees of moral responses, such as the ones who express them or bystanders (Ciurria 2019, 57-63).

that they ground the defensibility of our holding each other morally responsible in the sense of blameworthiness, *irrespective* of our being morally responsible agents capable of awareness or control. While these strategies risk bypassing the research question of the present piece, as similarly argued by Dominguez (2020, 164f.) who refers to Zheng`s (ibid.) and Mason`s (ibid.) approaches, they might be able to address harm induced by implicit bias more fully. Blaming or non-blaming might avoid harm *irrespective* of our research question of whether we are morally responsible in the sense of deserving the response (not exercising our control capacity).

We just saw that our approach is, in its desert-oriented focus, strictly limited. It is limited, at least in so far as it cannot fully accommodate condition (2), the necessity to avoid harm induced by implicit bias. Now, is there an alternative possibility for how to justify possibly effective non-blaming responses? One that can contribute to accommodate condition (2), at least partly addressing this limitation, while, as discussed in 5.1.a, not risking to bypass what we are concerned with, namely, what it means to have control over implicit bias`s impact and to *therefore* be morally responsible for its influence?

5.1.c Possibility Two: Designing Our Normative Expectations Of One Another

Again, recall our worry in this Part. In some situations, someone like Tom might not be a responsible agent. Due to, e.g., a lack of norm internalization, blaming someone like him would not make him able to discriminate less (which is, as I argued in Part Four, an empirical question). Though what should we do then? How can we justify alternative responses, such as aretaic ones? In the previous Subsection (5.1.a), I proposed not to approach this issue by justifying non-blaming responses equally to blaming responses.

In what shall follow in this Part, I will propose an alternative possibility. The proposal does not abandon Vargas` revisionist Agency Cultivation Model but merely supplements it while contributing to accommodating condition (2). Furthermore, it can be seen as a valuable supplementation to a revisionist approach: it reduces the detachment from both more conventional theories, which do consider blaming responses, and from social practice, of which blame responses do form a part. As a supplementation to Vargas, I shall propose Maureen Sie`s (2018) view as an example of a conversational approach to moral responsibility. My proposal boils down to conceiving of what Sie (ibid., 306) coins as the "design aspect", the additional function of moral responses to determine the content of normative expectations, as a justification of non-blame responses, which does not require responsible agency to be in place.

In Section 5.2, I will briefly sketch what sets the design aspect apart from what we have been concerned with so far, the "process aspect" (Sie 2018, 306) of holding each other morally responsible. In 5.3, I will summarize my proposal for a partial supplementation through the design aspect and how it can more fruitfully address conditions (2) and (3) of the present piece.

5.2 The Design Aspect of Our Moral Responses

Conversational approaches to moral responsibility shed light on often under-appreciated, communicative aspects of our responsibility practices[27]. Maureen Sie (2018) follows such an approach. Sie distinguishes two valuable social functions communicating through moral responses has.[28]

The first function Sie (2018) identifies is what this piece was concerned with so far. Sie denotes it as the "*process aspect*" (ibid., 306) of our moral responses: our praising and blaming are justified because they serve us to develop agential capacities to be sensitive to moral considerations (Vargas 2013; McGeer 2015). In general, it is justified to blame someone like Tom because it serves him in the ongoing 'process' of developing his moral identity, in becoming a responsible person sensitive to considerations such as what he values or what is expected from him. What we care about, how we should act in light of our values in a specific situation, how these sorts of considerations conflict or align with other people`s interests, desires and expectations… all this we realize when we are targets of moral responses such as indignation, praise or blame. And, as discussed, if praise or blame cannot generally support a person to, in this sense, develop his moral identity and become a better person who can prevent harmful behaviour, for Vargas (ibid.), blaming is not justified.

But according to Sie (2018), there is also a second function of our moral responses. It is the "*design aspect*" (ibid., 306) of our holding each other morally responsible, the 'designing' of what it is that we expect from each other in a particular situation. On this line, moral responses serve to settle the *content* of the moral considerations to which we might or might not be sensitive, such as

[27] See Michael McKenna`s (2012) for a recent and more widely recognized conversational account of responsibility.

[28] Besides these two functions that describe in what sense being held responsible can be valuable, see, e.g., Darwall (2006), McKenna (2012) or Watson (2004) for accounts that tackle a potential third way, which is to be addressed as moral persons.

normative expectations we uphold. Suppose someone blames Tom for his implicit bias. For Tom himself, for the person who expresses the blame, and even for us, supposing we were not directly involved in the matter, witnessing this form of moral feedback serves to continuously and collectively "co-determine, consolidate, and fine-tune [our] normative expectations of one another" (ibid, 300). It serves us to determine what we expect from each other or someone like Tom in specific contexts. What we achieve in observing or exchanging moral feedback, such as our reminding each other of values we strive for (e.g., to avoid discrimination whenever we can), or in our praising and blaming each other for our failure to do so, is to collectively determine how we aim to continue to communicate and to interact.

Note that Sie`s (2018) description of *how* both the process and the design aspect of our responsibility practices occur is compatible with Vargas´s view on our circumstantial capacities (compare Part Three). This is one reason why her view on the design aspect can be a suitable candidate for a supplementation of Vargas. Recall that Vargas` view on our circumstantial capacities to be sensitive to reasons, e.g., to what is expected from us, entails that these capacities are circumstantial, that is, highly dependent on how an agent relates to a context. Maureen Sie`s (2009, 2014, 2018) "Traffic Participation View on Human Agency" is similar and thus likely compatible with Vargas. Similar to how we learn to drive a car, to become a proper participant in traffic, we learn how to navigate in the contexts in which we act "as we go along" (2018, 318). Such a view entails that both our becoming responsible agents (the process aspect) as well as our adapting the content of relevant considerations (the design aspect) are highly context-dependent and not restricted to reasons we are aware of. Instead, we navigate and act upon what Sie specifies as "conditional frameworks" (Sie 2018, 303), sets of context-dependent considerations, aims, conditions, and complex combinations of those.

For example, in his role as an owner of a chain of gyms, Tom acts upon a conditional framework that includes considerations such as financial matters, his conception of what makes a good trainer, but also upon expectations he presumes others might hold of a gym owner (even expectations of future clients, compare Part Four). And in another context, such as in his role of a father, the conditional framework upon which he acts, the normative expectations present in a context, shift as well. By being blamed, Tom becomes more or less sensitive to some of these considerations (the process aspect of moral responses). And by being the target of blame or other responses, or by expressing those, or by even only witnessing them, he participates in collectively determining what it is that we expect from someone like him (the design aspect of moral responses). Thereby, most of the time, Tom is not aware of all the considerations that make him choose the applicants as he does, or of his continuous participation in the refinement of the content of these normative expectations, the framework of conditions in a context. Without Tom`s awareness of this process, a moral response, such as the blaming of him, can also only serve as a sign that others agree on norms and values he transgressed. (Sie, 2009, 2014, 2018)

Therefore, Sie`s (2018) view on human agency and on *how* both processes occur is in at least two aspects in line with Vargas` (2013) circumstantial view (compare Part Three). Both views entail that our behaviour is highly context-dependent, and that we are, most of the time, not aware of all reasons that are relevant for our actions.

5.3 The Design Aspect as a Supplementation of Vargas' Account

I have illustrated how we praise and blame each other or express other forms of moral feedback not only to cultivate responsible agency for those who are blamed, but also to design the content of the moral considerations and normative expectations we aim to live up to. Let me now formulate the following proposal:

> The 'design aspect' can form a justification of our non-blame characteristic (e.g., aretaic) moral responses, even if agents were not morally responsible in the sense of blameworthy.

This supplementation of Vargas´s (2013) account can, I argue, be supportive of addressing the limitation set by his exclusive concern with blameworthiness. At the same time, it does not bypass what it means to be morally responsible in the sense of failing to control what one does. It justifies other forms of moral feedback than praising or blaming, even if the latter responses are not justifiable in Vargas` account.

5.3.a Filling the Gap: Justifying Non-Blame Characteristic Moral Responses

To establish this, let us suppose for the moment that when applying Vargas` (2013) account, we have reason to regard someone like Tom not as a morally responsible, blameworthy agent in a specific situation. Why would we do so? Recall Part Three: blaming might, in some contexts, be ineffective if people reacted reluctantly and without understanding to blaming responses. Then, for Vargas, blaming would be unjustified and morally responsible blameworthy agency would not be in place. I argue that Sie`s (2018) design aspect of our

holding each other responsible can close the gap of justification Vargas (2013) leaves with regards to other forms of still possibly effective moral feedback (e.g., aretaic, non-blame responses). I propose, drawing on Sie, that we can think of why we address someone like Tom by reminding him of universal values we all aim to live up to in the sense that this might serve the function to fine-grain and collectively recognize what can be generally expected from someone like him in that specific situation. If, for example, a bystander observes an applicant or someone else reminding Tom of the universal value to treat all applicants equally (as an instance of a non-blaming response), this can serve as a confirmation of what it is that we can commonly expect from someone like him. And, filling the gap Vargas leaves, I propose that this design function can serve as a justification of non-blame moral responses.

Furthermore, I propose that *even if* people were not morally responsible, blameworthy agents in Vargas` account, non-blame (e.g., aretaic) moral responses (compare Part Four) could still be justified by regressing to this design function as it is sketched by Sie (2018). Let me explain.

Suppose someone like Tom was, for whatever reason, not a morally responsible, blameworthy agent (according to Vargas (2013)), and he was approached with an aretaic moral response. A colleague explains to him what should be expected from someone like him and reminds him of his own egalitarian ideals he aims to live up to. Drawing on Sie (2018), I argue that even if someone like Tom was not blameworthy (for Vargas), such an aretaic response could still contribute to consolidating and making public what can be expected from people like Tom. Imagine some people standing by. The message can manifest how we can commonly expect each other to avoid the influence of implicit bias on common hiring decisions. Aretaic responses contribute to determining the content of what we collectively expect from people like Tom in specific contexts and are thus justifiable, even if people like Tom were not blameworthy (in Vargas` account).

They can remind not only us and Tom, but also people not directly involved, of what we can expect from each other in such a situation.

Regarding the justification of our *blaming* responses through the design aspect, two caveats are in order.

5.3.b Going Beyond the Gap: Justifying Blame-Responses Through the Design Aspect as Well?

First, let me note that if someone like Tom was a *blameworthy* agent according to Vargas (2013), I do not aim to rule out that then, besides the cultivation of responsible agency, blaming someone like him might well contribute to the designing of normative expectations as well. Nothing in the present argument precludes that blaming a blameworthy agent could serve that function as well or that, in this case, the design function could not serve as an additional justification.

Nevertheless, I propose that the cultivation of, in Vargas` sense, morally responsible agency that can circumstantially control implicit bias should remain the *necessary* justification of our blaming responses. If someone like Tom was not a morally responsible agent in Vargas` account, I propose that blaming someone like him would continue to be generally unjustified despite the possible contribution of blaming to the design aspect.

The reason for my proposal of this second caveat is congruent with what I have established in the rejection of Proposal 1 (in 5.1.a). I argue that the design aspect should not be seen as completely replacing Vargas` account in justifying blaming responses. This is because taking such a direction would, I fear, again amount to risking losing track of the research question we are concerned with. Why precisely? Because, in a very similar vein as my argument in 5.1.a, the design aspect is not directly concerned with control. More specifically, in line with Sie`s

(2018) classification, the design aspect is not concerned with the different, desert-oriented question of whether someone like Tom deserves the moral response due to a failure that is his own (or attributable to him as a person), a failure to exercise control.[29]

5.3.c The Design Aspect as a Supplementation of Vargas' Account

In summary, supplementing Vargas` account, I propose that if we deem agents blameworthy and morally responsible (according to Vargas), then both blame- and non-blame-characteristic moral responses can be justified on the grounds of the design aspect. Though, if agents are not morally responsible and blameworthy, according to Vargas, I propose to only justify *non-blame-characteristic moral responses* (e.g., aretaic responses) exclusively on the grounds of the design aspect. On this line, we do not risk bypassing the desert-oriented research questions and only supplement but do not replace Vargas` account of what it means to be blameworthy and in control.

This is the supplementation of Vargas` (2013) account I propose. I argue that it is a valuable supplementation of a revisionist account because it reduces the detachment from existing theories and social practice. And, crucially for this piece, it is valuable considering the conditions an account of moral responsibility for implicit bias`s influence should accommodate: this supplementation can, I

[29] A second, potential reason for not justifying our blaming each other merely on grounds of the design aspect relates to the role of the effectiveness of blaming in Vargas` (2013) account. Being not morally responsible is greatly determined on forward-looking grounds and, with respect to the justification thesis, in general terms. If non-morally responsible agents are still blamed, this does not exclude the risk of possible, harmful backfire-effects; and, for agents not to be morally responsible, blaming would have to be ineffective or even risk to backfire in *general*, and not only in specific instances. (compare Part Three) (Saul 2013, 55; Vargas 2017)

argue, accommodate conditions (2) and (3) more fully (although, as we saw in Subsection 5.1.b, condition (2) still not completely).

First, the design aspect can contribute to condition (2), *the avoidance of harm caused by implicit bias´s influence*. This is because the design aspect can, as established, form a second justification for non-blame characteristic (e.g., aretaic) moral responses even if our blaming or praising would not be justified applying Vargas` Agency Cultivation Model. And these aretaic moral responses can, as established in Part Four, reduce the harm caused by implicit bias.

Second, the design aspect can contribute to more fully accommodating condition (3), *the roles our expectations about each other play in our holding each other morally responsible*. More specifically, supplementing an account of responsibility with the design aspect specifies a second role of how our expectations connect to our responsibility practices. Namely, our moral responses serve to continuously (and mostly unconsciously) determine the content of the normative expectations we uphold from each other. (Sie 2018)

5.4 Conclusion - Part Five

What can be done if Tom was not a morally responsible agent in the sense of being blameworthy (according to Vargas)? How can we make sense of and justify addressing someone like him with alternative moral responses?

In this Part, I addressed the conceptual limitation of Vargas` exclusive concern with the justifiability of blame (and praise). As I described in *Section 5.1*, Vargas does not account for the justifiability of other moral responses than blame-characteristic ones. At the same time, they might even be effective in avoiding harm in cases in which blame responses are not. One possibility would be to justify non-blame responses in an equal way to blame responses (Possibility One, 5.1.a). Against this, I argued that non-blaming moral responses neither entail nor require the same justification that blaming responses do, namely, that the behaviour be under our control and, in some sense, be our fault. Thus, justifying non-blaming responses based on a concept they are not concerned with nor require would, I argued, bypass the research question of this piece: namely, what it means to be a morally responsible agent who deserves to be addressed as such because she has, in some sense, control over her implicitly biased behaviour.

On the other hand (in 5.1.b), it became apparent that the present, desert-oriented approach to the research question and its focus on control is strictly limited for addressing condition (2), the harm caused by implicit bias.

As an alternative (Possibility Two, 5.1.c), I proposed to supplement Vargas` Agency Cultivation Model with a potential second justification of moral responses, namely, on the grounds of their contribution to the collective design of what the content of the normative expectations is we aim to live up to. Our moral responses do not serve only to make those addressed better agents. They also serve to collectively determine and confirm what it is that we expect from each other.

As described in *Section 5.2*, Maureen Sie`s view on how these processes occur is compatible with Vargas´s circumstantial view on human circumstantial capacities to control implicit bias. In Sie`s conversational view, this collective designing, just as our recognizing and responding to reasons, occurs "as we go along" (Sie 2018, 318) and in a context-dependent manner that we are mostly not aware of.

Based on this, I formulated the following proposal of supplementing Vargas` account in *Section 5.3*: we can justify both blame- and non-blame-characteristic moral responses towards agents we deem morally responsible/ blameworthy (according to Vargas) on the grounds of the design aspect. On the other hand, I proposed restricting the justification of responses towards *non-blameworthy* agents on the grounds of the design aspect to *non-blame or praise responses only*, such as only to aretaic responses.

This supplementation, I argued, reduces an anti-revisionist detachment from existing theory and practice. And it accommodates conditions (2) and (3) of this piece more fully. It can more fully accommodate (2) our recognition of the necessity to avoid harm induced by implicit bias. This is because non-blame-characteristic moral responses can be effective in contributing to the avoidance of harm as well. This partly addresses the worry we started out with in this Part. Also, the supplementation accommodates (3) more fully, as it considers a second function our expectations play in our responsibility practices. Namely, the content of normative expectations can be collectively shaped through our holding each other morally responsible.

6. CONCLUSION

“Ought implies can” (Copp 2008). If we cannot act better, we are also not morally responsible for our bad conduct. Implicit bias challenges this fundamental principle of moral philosophy. I illustrated this challenge drawing on reason style control accounts. We are unaware of being influenced by implicit bias while acting. We cannot directly intervene and stop its influence. Thus, one can conclude that we lack direct control over it. Furthermore, we might also not always be able to anticipate every instance of being influenced by implicit bias which is out of our direct control and intention. But how can we then be morally responsible for implicit bias in a sense that does not completely abandon this fundamental philosophical principle of control? Are we morally responsible for harmful behaviour caused by implicit bias, and if we are, when and in what sense? This formed the research question of this piece.

In **Part Two**, I formulated three conditions which an account of responsible agency should aim to accommodate for successfully addressing this research question: it should be flexible enough for what implicit biases are and for the unstable person-dependent and situation-dependent influence of implicit bias on our behaviour (condition (1)); it should recognize what implicit biases do, namely, they can cause significant harm, in single instances as well as in aggregation (condition (2)); and it should take account of what we might commonly expect people to know and to do about the harmful influence of implicit bias given the roles they have and the situations they are in (condition (3)). These three conditions limited the scope of the present piece and how it addressed the challenge of control.

In this piece, I defended and critically scrutinized an example of a revisionist and forward-looking view on what it means to be able to control our behaviour influenced by implicit bias in a sense that makes us morally responsible for it. I argued that Manuel Vargas` account can greatly, although not fully, accommodate

the three conditions of the present piece and, thus, address the research question to a big extend successfully.

In **Part Three**, I introduced and assessed Vargas` revisionist and forward-looking view on responsible agency. Vargas provides an account of what it can mean to be in control of the influence of implicit bias. Having control is understood as having 'circumstantial' capacities. I argued that this account as promising for addressing the phenomenon of implicit bias and its challenge for reason-style control accounts. First, circumstantial capacities are highly flexible, depend on a function of the context and the agent, and thus, accommodate condition (1), the unstable, context- and agent-dependent influence of implicit bias. Second, for blaming to be justifiable, it must generally contribute to making people like Tom better capable of avoiding the harmful influence of implicit bias. Therefore, Vargas` forward-looking account is promising to accommodate condition (2) as well, recognizing the necessity to avoid harm. And third, his account is promising to accommodate condition (3) for two reasons: (i) having a circumstantial control capacity is interest-sensitive and highly depends on the interests and expectations present in a context. On the other hand, empirically, (ii) having a circumstantial capacity and the effectiveness of blaming also depend on the sufficient internalization of critical and supportive norms in a context. While the effectiveness of blaming is an empirical argument, one possible influence on it is that we often care about being seen as responsible and trustworthy agents by others. Note that in Vargas` account, nothing precludes that personal, circumstance-independent values could be relevant reasons for controlling our conduct as well. Though, on top of that, his prominent acknowledgement of condition (3), that circumstantial expectations of others can strongly influence how we act and whether we can control what we do in a context, is a vital peculiarity that differentiates his approach from conventional control accounts.

Based on Vargas` account, I formulated the following *provisional* answer to the research question: someone like Tom, an owner of a gym who evaluates resumes, is seen as morally responsible for his implicitly biased behaviour, in the sense of being blameworthy, if blaming such a person generally contributes to the cultivation of her capacity to be sensitive to the influence of implicit stereotypes that make people like him discriminate between the applicants.

In **Part Four**, I addressed Holroyd`s (2018) objections that McGeer`s (2015) revisionist scaffolded responsiveness view better contributes to the forward-looking aim of cultivating control and would be more in line with both conventional theories and our existent responsibility practices than Vargas` circumstantial view. Crucially, Holroyd argues that Vargas` circumstantial view is more restricted than McGeer`s and bound to our current circumstances only. It cannot account for the potential significance of what she coins as 'indirect reasons', expectations from people we encounter not in our current contexts but potentially in the future. Answering these objections, I explained how Vargas` view accommodates condition (3) and I discussed what blaming and non-blaming mean in practice (providing the floor for Part Five). Based on my defence, I *specified* the answer to the research question as follows: people might be morally responsible for implicit bias also dependent on what they might be able to expect from themselves and others in the future in a specific context. But on the other hand, as established in Part Three, it still holds that someone like Tom might be less likely morally responsible for implicit bias if people in his context have internalized norms that are not supportive and uncritical towards implicit bias. This is because then, blaming might not be generally effective in cultivating a control capacity to avoid its harmful influence.

In **Part Five**, I addressed the worry of what we can do if we have reason to regard people not as in 'circumstantial' control over their implicit biases. Blaming would not generally contribute to their control capacities. And thus, we have reason to

regard them as lacking control and as not morally responsible in the sense of blameworthy. How can we justify alternative moral responses which, as we have seen in Part Four, can be effective as well? Addressing this worry, I proposed the following *supplemented* answer: people are not morally responsible for implicitly biased behaviour in a context, in the sense of being blameworthy, if they do not have responsibility relevant circumstantial control capacities. But although they are not responsible agents and not blameworthy, they are still proper targets of other moral responses than blame-characteristic ones (e.g., aretaic ones). To fill the gap that Vargas` mere concern with blameworthiness left, I proposed that a justification for these non-blaming responses can be their contribution to collectively determining the content of normative expectations people have from one another in a context (Sie 2018). This design function of our holding each other morally responsible supplements Vargas` account given that it can be seen as, to a certain extent, separate from the cultivation of valuable control through blame. The design function is not directly concerned with the issue of what it means to have control. Therefore, though, for not risking circumventing the research question, I also proposed not to regard it as substituting Vargas` view and as a sufficient justification of blame responses.

On the other hand, while this supplementation accommodates a revisionist approach and conditions (2) and (3) more fully, I also conceded that our desert-oriented and control-focused approach to the research question would be limited if we were *exclusively* concerned with the avoidance of harm (condition (2)). I conceded that blaming might still be justifiable on other grounds than the cultivation of responsible agency of the addressees. *Even if* the addressees reacted reluctantly and *were not blameworthy/ morally responsible* (according to Vargas), it might still be defensible to assert that we *should be held morally responsible*. For example, we might regard blaming defensible due to its effects on other people's expectations or power structures.

Despite this limitation (and the effectiveness of blaming in cultivating control capacities ultimately remaining an empirical question), we should nevertheless recognize how far the present revisionist and desert-oriented approach brought us while still addressing the puzzling question we started out with. The challenge we started out with was that often we seem to lack control over our implicitly biased behaviour in a way that risks decisively undermining our moral responsibility for it and always excusing us for the harm we caused. Now, through the supplemented forward-looking account of Vargas, we have arrived at an assertion that is very different from a lack of moral responsibility due to a lack of control. Not only what we expect from ourselves, given our personal values, plays a role in how we act and what we are capable to control. On top of that, Vargas` control account crucially acknowledges the importance of social expectations. What others expect us to know and do and what is commonly thought that we *should be morally responsible for* prominently figures in *what we are morally responsible for* because capable of doing and controlling in a context. We care about the harm induced by implicit bias and the expectations of others related to roles we have, which is why we are often able to develop control capacities through being targets of blame. Thus, contrary to the challenge, probably in most situations, we are, in this supplemented revisionist and forward-looking account, capable to control the harmful influence of implicit bias. And thus, mostly, we can be seen as morally responsible for it if we fail to exercise our circumstantial capacity to control it.

Considering the limitation of the present desert-oriented approach concerning its aim to address condition (2), the necessity to avoid harm induced by implicit bias, let me close by giving a final outlook on what else can be done. How can we still intervene in harmful circumstances which can foster our *becoming* morally responsible for implicitly biased behaviour? We have already touched upon different strategies to indirectly control implicitly biased behaviour (Holroyd 2012; Sie and Vader-Bours 2016). Examples are possibly exposing oneself to counter stereotypical images in one`s environment (Dasgupta and Greenwald

2001) or anonymising documents that should be evaluated (such as Tom`s resumes). Madva (2020, 233f) lists recent evidence on different tools and strategies that can be used for combatting implicit bias. On the individual level, for instance, “if-then” plans, simple and very concrete decision rules that can be applied to specific situations, can effectively reduce implicit bias (Mendoza et al. 2010; Stewart and Payne 2008). For example, to a reader who is worried about interrupting women more than men, Madva (ibid.) suggests learning the simple if-then plan: “If she’s talking, then I won’t!” Another example is the introduction of clear criteria for decision making such as in the context of hiring (such as in Tom`s situation) or voting (Uhlmann and Cohen 2005).

Intervening in harmful practices might, in the end, amount to making people aware of such practices on a broader basis. Such an outlook is congruent with lines of research such as Susan Hurley`s (2011) “democratic public scaffolding principle” (ibid., 193). Hurley formulates a broadly positive principle to which governments should adhere when intervening in citizen`s circumstances. It entails that governments should create an ecology in which citizens can act and reason autonomously. The government should support people in gaining control over what they do, given the high context-dependency of their behaviour.

Such lines of research are promising. As argued, according to Vargas, there might be situations in which blaming each other is generally unjustified because it cannot contribute to the cultivation of control capacities and, thus, to the avoidance of harmful behaviour. Thereby, I conceded that Vargas` desert-oriented approach is limited for addressing harm induced by implicit bias: blaming might be justifiable in other ways as well. On the other hand, lines of research, such as Hurley`s (2011), are recent and promising directions for future work. Such approaches relate the necessity of public intervention to creating circumstances in which people have control and are therefore blameworthy for their implicitly biased behaviour, something that is, as we saw, far from necessarily out of control.

7. ACKNOWLEDGEMENTS

I would like to thank my Supervisor Constanze Binder and my Advisor Alex Voorhoeve for their insightful feedback. Of course, thanks to Ina Jäntgen, Marie-Aimée Salopiata, Jakob Schönhuber, Chiara Stenico and my flatmate Mareg Marcos for the countless discussions, their fantastic emotional support, and for encouraging me to push through despite the pandemic. I would also like to thank Måns Abrahamson and Ermanno Petrocchi for their thoughtful comments on an early draft in the EIPE seminar, and Bronagh Dunne for helping me out in the very last minute. Finally, thank you to the whole EIPE community. It was an amazing time.

8. BIBLIOGRAPHY

Amodio, David M, Patricia G Devine, and Eddie Harmon-Jones. 2007. "A Dynamic Model of Guilt: Implications for Motivation and Self-Regulation in the Context of Prejudice." *Psychological Science* 18 (6): 524–30. https://doi.org/10.1111/j.1467-9280.2007.01933.x.

Arpaly, Nomy. 2002. *Unprincipled Virtue: An Inquiry Into Moral Agency*. New York: Oxford University Press. https://doi.org/10.1093/0195152042.001.0001.

Bendick Jr., Marc, and Ana P Nunes. 2012. "Developing the Research Basis for Controlling Bias in Hiring." *Journal of Social Issues* 68 (2): 238–62. https://doi.org/10.1111/j.1540-4560.2012.01747.x.

Brennan, Samantha. 2016. "The Moral Status of Micro-Inequities: In Favor of Institutional Solutions." In *Implicit Bias and Philosophy, Volume 2*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.003.0011.

Brownstein, Michael. 2019. "Implicit Bias." In *The {Stanford} Encyclopedia of Philosophy*, edited by Edward N Zalta, Fall 2019. Metaphysics Research Lab, Stanford University.

Buckwalter, Wesley. 2019. "Implicit Attitudes and the Ability Argument." *Philosophical Studies* 176 (11): 2961–90. https://doi.org/10.1007/s11098-018-1159-7.

Cameron, C Daryl, Jazmin L Brown-Iannuzzi, and B Keith Payne. 2012. "Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations with Behavior and Explicit Attitudes."

Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc 16 (4): 330–50. https://doi.org/10.1177/1088868312440047.

Castelli, Luigi, Alexia Zecchini, Leyla Deamicis, and Steven J Sherman. 2005. "The Impact of Implicit Prejudice about the Elderly on the Reaction to Stereotype Confirmation and Disconfirmation." *Current Psychology* 24 (2): 134–46. https://doi.org/10.1007/s12144-005-1012-y.

Ciurria, Michelle. 2019. *An Intersectional Feminist Theory of Moral Responsibility*. https://doi.org/10.4324/9780429327117.

Copp, David. 2008. "'Ought' Implies 'Can' and the Derivation of the Principle of Alternate Possibilities." *Analysis* 68 (297): 67–75. https://doi.org/https://doi.org/10.1111/j.1467-8284.2007.00715.x.

Czopp, Alexander M, Margo J Monteith, and Aimee Y Mark. 2006. "Standing up for a Change: Reducing Bias through Interpersonal Confrontation." *Journal of Personality and Social Psychology*. Czopp, Alexander M.: Department of Psychology, University of Toledo, 2801 West Bancroft Street, Toledo, OH, US, 43606, alexander.czopp@utoledo.edu: American Psychological Association. https://doi.org/10.1037/0022-3514.90.5.784.

Darwall, S. L. 2006. "The Second-Person Standpoint: Morality, Respect, and Accountability." *MA: Harvard University Press*.

Dasgupta, Nilanjana, and Shaki Asgari. 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): 642–58. https://doi.org/10.1016/j.jesp.2004.02.003.

Dasgupta, Nilanjana, and Anthony G Greenwald. 2001. "On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of

Admired and Disliked Individuals." *Journal of Personality and Social Psychology*. Dasgupta, Nilanjana: Dept of Psychology, New School U, 65 Fifth Avenue, New York, NY, US, 10003, dasguptn@newschool.edu: American Psychological Association. https://doi.org/10.1037/0022-3514.81.5.800.

Dominguez, Noel. 2020. "Moral Responsibility for Implicit Biases." In *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind (1st Ed.)*, edited by Erin Beeghly and Alex Madva, 310. Routledge. https://doi.org/https://doi-org.eur.idm.oclc.org/10.4324/9781315107615.

Faucher, Luc. 2016. "Revisionism and Moral Responsibility for Implicit Attitudes." In *Implicit Bias and Philosophy, Volume 2*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.003.0006.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. *Cambridge Studies in Philosophy and Law*. Cambridge: Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511814594.

Friese, Malte, Wilhelm Hofmann, and Manfred Schmitt. 2008. "When and Why Do Implicit Measures Predict Behaviour? Empirical Evidence for the Moderating Role of Opportunity, Motivation, and Process Reliance." *European Review of Social Psychology* 19 (September): 285–338. https://doi.org/10.1080/10463280802556958.

Gawronski, Bertram, Wilhelm Hofmann, and Christopher J Wilbur. 2006. "Are 'Implicit' Attitudes Unconscious?" *Consciousness and Cognition: An International Journal*. Gawronski, Bertram: Department of Psychology, University of Western Ontario, Social Science Centre, London, ON, Canada, N6A 5C2, bgawrons@uwo.ca: Elsevier Science.

https://doi.org/10.1016/j.concog.2005.11.007.

Glasgow, Joshua. 2016. “Alienation and Responsibility.” In *Implicit Bias and Philosophy, Volume 2*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.003.0003.

Granados Samayoa, Javier A, and Russell H Fazio. 2017. “Who Starts the Wave? Let’s Not Forget the Role of the Individual.” *Psychological Inquiry* 28 (4): 273–77. https://doi.org/10.1080/1047840X.2017.1373554.

Greenwald, Anthony G, and Mahzarin R Banaji. 1995. “Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes.” *Psychological Review*. US: American Psychological Association. https://doi.org/10.1037/0033-295X.102.1.4.

Gschwendner, Tobias, Wilhelm Hofmann, and Manfred Schmitt. 2008. “Differential Stability: The Effects of Acute and Chronic Construct Accessibility on the Temporal Stability of the Implicit Association Test.” *Journal of Individual Differences* 29 (2): 70–79. https://doi.org/10.1027/1614-0001.29.2.70.

Hahn, Adam, and Bertram Gawronski. 2019. “Facing One’s Implicit Biases: From Awareness to Acknowledgment.” *Journal of Personality and Social Psychology*. Hahn, Adam: Social Cognition Center Cologne, Department of Psychology, University of Cologne, Richard-Strauss-Straße 2, Köln, Germany, 50931, Adam.Hahn@uni-koeln.de: American Psychological Association. https://doi.org/10.1037/pspi0000155.

Hahn, Adam, Charles M Judd, Holen K Hirsh, and Irene V Blair. 2014. “Awareness of Implicit Attitudes.” *Journal of Experimental Psychology. General* 143 (3): 1369–92. https://doi.org/10.1037/a0035028.

Harland, Harry. 2020. “Beyond the Moral Influence Theory? A Critical

Examination of Vargas's Agency Cultivation Model of Responsibility." *The Journal of Ethics* 24 (4): 401–25. https://doi.org/10.1007/s10892-020-09328-0.

Haslanger, Sally. 2015. "Distinguished Lecture: Social Structure, Narrative and Explanation." *Canadian Journal of Philosophy* 45 (1): 1–15. https://doi.org/10.1080/00455091.2015.1019176.

Hehman, Eric, Jessica K Flake, and Jimmy Calanchini. 2017. "Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents." *Social Psychological and Personality Science* 9 (4): 393–401. https://doi.org/10.1177/1948550617711229.

Holroyd, Jules. 2012. "Responsibility for Implicit Bias." *Journal of Social Philosophy* 43 (3): 274–306. https://doi.org/https://doi.org/10.1111/j.1467-9833.2012.01565.x.

———. 2018. "Two Ways of Socializing Moral Responsibility: Circumstantialism versus Scaffolded- Responsiveness." In *Social Dimensions of Moral Responsibility*. https://doi.org/10.1093/oso/9780190609610.001.0001.

Holroyd, Jules, and Daniel Kelly. 2016. "Implicit Bias, Character, and Control." In *From Personality to Virtue*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198746812.003.0006.

Holroyd, Jules, Robin Scaife, and Tom Stafford. 2017. "Responsibility for Implicit Bias." *Philosophy Compass* 12 (3): e12410. https://doi.org/https://doi.org/10.1111/phc3.12410.

Hurley, Susan. 2011. "The Public Ecology of Responsibility 1." In *Responsibility and Distributive Justice*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199565801.003.0010.

Kurdi, Benedek, and Mahzarin R Banaji. 2017. “Reports of the Death of the Individual Difference Approach to Implicit Social Cognition May Be Greatly Exaggerated: A Commentary on Payne, Vuletich, and Lundberg.” *Psychological Inquiry* 28 (4): 281–87. https://doi.org/10.1080/1047840X.2017.1373555.

Levy, Neil. 2014. “Consciousness, Implicit Attitudes and Moral Responsibility.” *Noûs* 48 (1): 21–40. https://doi.org/https://doi.org/10.1111/j.1468-0068.2011.00853.x.

———. 2017. “Implicit Bias and Moral Responsibility: Probing the Data.” *Philosophy and Phenomenological Research* 94 (1): 3–26. https://doi.org/https://doi.org/10.1111/phpr.12352.

Madva, Alex. 2018. “Implicit Bias, Moods, and Moral Responsibility.” *Pacific Philosophical Quarterly* 99 (S1): 53–78. https://doi.org/https://doi.org/10.1111/papq.12212.

———. 2020. “Individual and Structural Interventions.” In *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind (1st Ed.)*, edited by Erin Beeghly and Alex Madva, 1st ed., 310. Routledge. https://doi.org/https://doi-org.eur.idm.oclc.org/10.4324/9781315107615.

Mason, Elinor. 2018. “Respecting Each Other and Taking Responsibility for Our Biases.” In *Social Dimensions of Moral Responsibility*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190609610.003.0007.

Master, Allison, Sapna Cheryan, and Andrew N Meltzoff. 2016. “Computing Whether She Belongs: Stereotypes Undermine Girls’ Interest and Sense of Belonging in Computer Science.” *Journal of Educational Psychology*. Master, Allison: Institute for Learning and Brain Sciences, University of

Washington, Seattle, WA, US, 98195, almaster@uw.edu: American Psychological Association. https://doi.org/10.1037/edu0000061.

McGeer, Victoria. 2015. “Building a Better Theory of Responsibility.” *Philosophical Studies* 172 (10): 2635–49. https://doi.org/10.1007/s11098-015-0478-1.

McGeer, Victoria, and Philip Pettit. 2015. “The Hard Problem of Responsibility.” In *Oxford University Press (Ed., D. Shoemaker)*. Vol. 3.

McKenna, Michael. 2012. *Conversation and Responsibility*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199740031.001.0001.

———. 2013. “Reasons-Responsiveness, Agents, and Mechanisms.” In *Oxford Studies in Agency and Responsibility Volume 1*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694853.003.0007.

Mendoza, Saaid A, Peter M Gollwitzer, and David M Amodio. 2010. “Reducing the Expression of Implicit Stereotypes: Reflexive Control through Implementation Intentions.” *Personality & Social Psychology Bulletin* 36 (4): 512–23. https://doi.org/10.1177/0146167210362789.

Moss-Racusin, Corinne A, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. “Science Faculty’s Subtle Gender Biases Favor Male Students.” *Proceedings of the National Academy of Sciences* 109 (41): 16474 LP – 16479. https://doi.org/10.1073/pnas.1211286109.

Nelkin, Dana Kay. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199608560.001.0001.

O’Brien, K S, J A Hunter, and M Banks. 2007. “Implicit Anti-Fat Bias in Physical Educators: Physical Attributes, Ideology and Socialization.”

International Journal of Obesity 31 (2): 308–14. https://doi.org/10.1038/sj.ijo.0803398.

Park, Jaihyun, Karla Felix, and Grace Lee. 2007. "Implicit Attitudes towards Arab-Muslims and the Moderating Effects of Social Information." *Basic and Applied Social Psychology* 29 (1): 35–45. https://doi.org/10.1080/01973530701330942.

Payne, B Keith, Heidi A Vuletich, and Kristjen B Lundberg. 2017. "Flipping the Script on Implicit Bias Research with the Bias of Crowds." *Psychological Inquiry* 28 (4): 306–11. https://doi.org/10.1080/1047840X.2017.1380460.

Rüsch, Nicolas, Patrick W Corrigan, Andrew R Todd, and Galen V Bodenhausen. 2010. "Implicit Self-Stigma in People with Mental Illness." *The Journal of Nervous and Mental Disease* 198 (2): 150–53. https://doi.org/10.1097/NMD.0b013e3181cc43b5.

Saul, Jennifer. 2013a. "Implicit Bias, Stereotype Threat, and Women in Philosophy." In *Women in Philosophy*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199325603.003.0003.

———. 2013b. "Unconscious Influences and Women in Philosophy." *Women in Philosophy: What Needs to Change*, 39–60.

Scaife, Robin, Tom Stafford, Andreas Bunge, and Jules Holroyd. 2020. "To Blame? The Effects of Moralized Feedback on Implicit Racial Bias." Edited by Simine Vazire and Simine Vazire. *Collabra: Psychology* 6 (1). https://doi.org/10.1525/collabra.251.

Sher, George. 2009. *Who Knew?: Responsibility Without Awareness*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195389197.001.0001.

Sie, Maureen. 2009. “Moral Agency, Conscious Control, and Deliberative Awareness.” *Inquiry* 52 (5): 516–31. https://doi.org/10.1080/00201740903302642.

———. 2014. “Self-Knowledge and the Minimal Conditions of Responsibility: A Traffic-Participation View on Human (Moral) Agency.” *The Journal of Value Inquiry* 48 (2): 271–91. https://doi.org/10.1007/s10790-014-9424-2.

———. 2018. “Sharing Responsibility: The Importance of Tokens of Appraisal to Our Moral Practices.” In *Social Dimensions of Moral Responsibility*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190609610.003.0013.

Sie, Maureen, and Nicole van Voorst Vader-Bours. 2016. “Stereotypes and Prejudices: Whose Responsibility?: Indirect Personal Responsibility for Implicit Biases.” In *Implicit Bias and Philosophy, Volume 2*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.003.0005.

Sripada, Chandra. 2015. “Self-Expression: A Deep Self Theory of Moral Responsibility.” *Philosophical Studies* 173 (August). https://doi.org/10.1007/s11098-015-0527-9.

Stewart, Brandon D, and B Keith Payne. 2008. “Bringing Automatic Stereotyping Under Control: Implementation Intentions as Efficient Means of Thought Control.” *Personality and Social Psychology Bulletin* 34 (10): 1332–45. https://doi.org/10.1177/0146167208321269.

Stout, Nathan. 2016. “Reasons-Responsiveness and Moral Responsibility: The Case of Autism.” *The Journal of Ethics* 20 (4): 401–18. https://doi.org/10.1007/s10892-016-9218-9.

Strawson, P. F. 1962. “Freedom and Resentment.” *Proceedings of the British*

Academy, no. 48: 1–25.

Timpe, Kevin. 2014. “Vargas Manuel, Building Better Beings: A Theory of Moral Responsibility.” *Ethics* 124 (4): 926–31. https://doi.org/10.1086/675870.

Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199697540.001.0001.

———. 2015a. “Desert, Responsibility, and Justification: A Reply to Doris, McGeer, and Robinson.” *Philosophical Studies* 172 (10): 2659–78. https://doi.org/10.1007/s11098-015-0480-7.

———. 2015b. “Précis of Building Better Beings: A Theory of Moral Responsibility.” *Philosophical Studies* 172 (10): 2621–23. https://doi.org/10.1007/s11098-015-0476-3.

———. 2018. “The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology.” In *Social Dimensions of Moral Responsibility*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190609610.003.0005.

———. 2020. “Negligence and Social Self-Governance.” In *Surrounding Self-Control*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780197500941.003.0021.

Wallace, R Jay. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.

Watson, Gary. 1996. “Two Faces of Responsibility.” *Philosophical Topics* 24 (2): 227–48. http://www.jstor.org.eur.idm.oclc.org/stable/43154245.

———. 2004. “Responsibility and the Limits of Evil: Variations on a

Strawsonian Theme.” In *Agency and Answerability*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199272273.003.0009.

Zheng, Robin. 2016. “Attributability, Accountability, and Implicit Bias.” In *Implicit Bias and Philosophy, Volume 2*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.003.0004.