Journal of the American Philosophical Association (2023) 1–20 
DOI:10.1017/apa.2021.45

# The Nurturing Stance, Moral Responsibility, and the (Implicit) Bias Blind Spot

ABSTRACT: *Can we hold agents responsible for their implicitly biased behavior? The aim of this text is to show that, from the nurturing stance, holding subjects responsible for their implicitly biased behavior is justified, even though they are not blameworthy. First, I will introduce the nurturing stance as Daphne Brandenburg originally developed it. Second, I will specify what holding somebody responsible from the nurturing stance amounts to. Third, I show how and why holding responsible can help a subject develop an impaired capacity. Fourth, I analyze empirical data about holding prejudiced subjects responsible and highlight that the internal motivation to control prejudiced reactions decreases implicit attitudes' influences. Furthermore, the data show that in order to be appropriate moral demands have to acknowledge the target's autonomy and competence. In sum, from the nurturing stance, holding implicitly biased subjects responsible is appropriate if they can adequately respond to the moral demands.*

KEYWORDS: self-control, implicit bias, forward-looking responsibility, abilities, self-image, discrimination, accountability

## Introduction

Social psychologists and philosophers currently debate the implications of implicit biases. While there are different definitions of implicit biases (Holroyd et al., 2017), the most natural understanding is that implicit biases are the subject's unintended behavioral patterns that contradict the subject's acknowledged goals or values, such as egalitarianism. For example, in one experiment that illustrates implicit biases subjects evaluate job applications and unintentionally engage in discriminatory behavior as they assess the job applications from Black (Dovidio and Gaertner 2000) or Arab-Muslim (Rooth 2010) applicants worse than those of others just because of the applicants' social groups.

If implicit prejudices or implicit stereotypes, the roots of implicit biases, contribute to unintentional discrimination, the question arises whether subjects are responsible for these actions (see section 4.2 for a short discussion; for an overview, see Brownstein 2015; in this text, I am not concerned with the question of whether we are morally responsible for *having* implicit prejudices and

This work was supported by the German Research Foundation [grant number BA 6721/1-1]. The author thanks Josh Cangelosi for assistance with improvements to this article.

stereotypes. I am interested in the behavior that might issue from them in a range of cases.). To give a reasonable answer, philosophers concentrated especially on control as the decisive prerequisite for a subject's responsibility for her behavior. Many philosophers (for example, see Fischer and Ravizza 1998) claim, if a subject acts unfreely, in the sense of having neither sufficient direct nor indirect control, she cannot be morally responsible for the action. Accordingly, some philosophers (for example, Levy 2014) assume that agents are not morally responsible for implicitly biased behavior because implicit biases undermine the agents' control.

In what follows, I will focus on the nurturing stance (Brandenburg 2018) and its implications for implicit biases. In the first section, I will introduce the nurturing stance, which Daphne Brandenburg (2018) developed to explain a phenomenon of our responsibility practices: sometimes we hold agents responsible even though they are not blameworthy. This practice of holding responsible is justified if the agent lacks an ability (Brandenburg calls it 'underdeveloped capacity'), which the agent can develop. From the nurturing stance, an agent is held responsible for a lack of effort in developing her ability and not for the past action. Based on a conversational understanding of responsibility practices (McKenna 2012), I shall argue that holding responsible through the nurturing stance helps to improve an impaired capacity if the moral demand is structured in a way that entails a meaningful responsibility exchange. The appropriateness of the moral demand depends on whether an agent can respond adequately. I will show with empirical data that, often, biased agents can respond adequately to the demand.

I conclude that holding an agent responsible because of implicitly biased behavior can be justified from the nurturing stance. This means that the person who holds somebody responsible assumes the target is not fully blameworthy for her behavior because she was not in control. Simultaneously, however, the moral demand targets the agent's lack of effort in developing her weak ability that led her to undertake the bad action. Thereby, I shed light on the mixed intuitions philosophers have about moral responsibility for implicitly biased behavior (for instance, see Levy [2014] versus Madva [2017]). The nurturing stance clarifies our otherwise mixed intuitions: Although implicit biases undermine an agent's control, there is some sense in which the agent could have done better. It will turn out that the sense in which the agent could have done better can only be realized by holding the agent in question responsible.

In section 2, I focus on questions neglected in Brandenburg's description of the nurturing stance: how does holding responsible contribute to the development of an ability? Here, I will separate two questions: (a) Why does holding responsible help an agent, and (b) *how* does holding responsible contribute? Regarding (a), I will rely on literature from social psychology that shows that we use other agents as information sources about ourselves. A part of our self-image develops through the reactions of others to our behavior. With regard to (b), I will suggest that if the moral demand is structured in the right way, it comprises valid reasons that an agent needs in order to adjust her self-image. Only if a subject adjusts her self-image, can her potential to develop the ability be realized. Accordingly, social interaction is essential for developing the ability, because the external response plays a crucial role in moral development.

In section 4, I will apply the developed thoughts to the case of implicitly biased behavior. I will show with empirical data that holding nonblameworthy agents

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

responsible indeed helps them to develop an impaired but improvable self-control ability. This only works, as I will highlight in section 4.3, if the autonomy of the agent is acknowledged.

## 1. Nurturing Stance

Brandenburg (2018) introduced the nurturing stance to explain a certain nursing practice in psychiatry. The nurturing stance, an addition to a Strawsonian (1962) understanding of holding responsible, solves the following problem: clinicians sometimes hold patients responsible for their actions, while the clinicians assume that the subject is not blameworthy. For example, a patient throws a chair across the room, and a clinician holds her responsible even though the clinician thinks that the agent was not in control. Given that a patient in psychiatric care has some mental limitation that can impair her ability to behave according to norms, she is not blameworthy. This phenomenon is puzzling if holding responsible means, as Strawsonians suggest, judging that the responsible agent is a valid target for blame and therefore blameworthy (this challenge of the Strawsonian picture was introduced and discussed in Pickard 2013).

Originally, Strawson offered two stances we can take toward other persons. First, holding somebody morally responsible presupposes a full interpersonal stance toward that person. This stance entails reactive attitudes, which are attitudes that are a response to an attitude we notice in others. A reactive attitude entails reactive feelings, such as resentment. The experience of such emotional reactions motivates people to express demands or expectations, which is essential for the practice of holding morally responsible. Second, taking an objective attitude toward an agent involves refraining from blaming and distancing oneself from an interpersonal relation to avoid the strains of involvement. People who adopt the objective attitude view the target agent as an object of manipulation or as a subject of treatment.

It seems that, for Strawson, if the clinician has to depart from seeing the agent as morally responsible and therefore as not blameworthy, then only the objective stance remains. Brandenburg (2018) suggests that there is a middle-ground stance here: the nurturing stance.

The nurturing stance tries to solve the puzzle without claiming that the clinicians' practice of holding responsible is unjustified. Three elements are crucial for this solution: First, the patient is not responsible for her behavior because she is incapable to act well and other agents do not expect her to act correctly. Second, the patient is, however, responsible for developing or restoring a certain ability. Third, holding responsible in terms of the nurturing stance does not involve blame for the poor behavior.

According to Brandenburg (2018), concerning the first element, the word 'capacity' has an ambivalent meaning. On the one hand, it means that an individual is currently not able to do something, like speaking Chinese (*Type-1 sense*). A subject who did not learn Chinese cannot start speaking Chinese. On the other hand, it can mean that an individual is able to do something if she develops the capacity (*Type-2 sense*). Although a person is not able to start speaking Chinese now, she can learn the language. Analogously, the patient who threw a

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

chair across the room was not in control of her behavior given the psychiatric diagnosis, and therefore she lacked the first sense of capacity. Simultaneously, concerning the second element, she has the ability, in the Type-2 sense, to develop a degree of self-control that is sufficient to be in control in such situations in the future. On this basis, Brandenburg concludes that 'patients are then not responsible for what they have done, but they are responsible for altering these types of doings' (Brandenburg 2018: 14).

Regarding the lack of blame, Brandenburg (2018: 16) claims that 'the patient is then indeed held responsible because he *can* do otherwise [Type-2 sense]. But he is not blamed because he *can't* do otherwise [Type-1 sense]. Both are true because *can* here refers to different *types* of capacities'. She argues that patients are not really blamed although it appears as if they are. Clinicians, Brandenburg claims, do not consider patients to be blameworthy, and clinicians do not have feelings of resentment toward them. Thus, the act of holding responsible is considered neutral in terms of praising and blaming—this way of holding responsible is possible due to the nurturing stance. This stance is not an artificial addition to our moral responsibility practices, but an actual part of it, which Strawson neglected. In this sense, the illustrated clinicians' reactions are natural and not the result of extensive deliberation.

The original puzzle vanishes by claiming that the patient is not responsible for an act but for developing an underdeveloped capacity. She is not blameworthy for executing the act because she did not have control. Furthermore, Brandenburg assumes that the agent is not blameworthy for having an underdeveloped capacity because the agent is neither in direct nor indirect control of developing the capacity, which is why the agent is not blamed. (If the reason why a person does not have a Type-1 capacity traces to prior negligence, then the person would be indirectly responsible for the bad actions.)

According to Brandenburg, in the illustrated case, a clinician holds the patient accountable for not putting enough effort in developing the relevant impaired self-control ability. If, however, being responsible means being an appropriate target for blaming or praising attitudes, how could this work with a patient who is not blameworthy? In this specific situation, the patient is not accountable for throwing the chair because, in this situation, the patient cannot adequately respond to the clinician's demands. Brandenburg thinks, however, that the patient can adequately respond to the demand to develop the impaired self-control ability. Nevertheless, blame is not in order because the patient in psychiatric care does not have sufficient indirect control over her ability's development: To develop the impaired ability, the patient relies on the clinician's social cues that are given through moral demands (in section 3, I will explain this exchange more detail). Brandenburg suggests that if the patient disregards the demand, the patient becomes a potential target for blaming attitudes. Thereby, so Brandenburg, the nurturing stance does not undermine the Strawsonian picture.

## 2. The Nurturing Stance and Moral Demands

What is the nature of responsibility? As suggested by McKenna (2012), responsibility is like a special kind of conversation, and praise and blame are only a part of that

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

unfolding interaction. From this perspective, how does holding somebody responsible from the nurturing stance actually work?

First, a clarification is in order. Even though Brandenburg prefers to talk about capacities, I think that the term *abilities* is better suited for the current purpose. In contrast to capacities, as I will understand them in what follows, abilities apply only to agents who can act. Of course, sometimes we speak of a person who is able to digest, but on a more restricted way of thinking about abilities, they bear on exercises of agency and in most cases (barring deviance) pertain to voluntary undertakings. Abilities imply something about an agent's powers, while capacities apply to objects as well (Mele 2017). An agent who has learned to play the guitar has the general ability to play the guitar. Broadly speaking, having a general ability presupposes that the agent's brain has the required neuronal synapses. A regular agent who does not have the general ability to play the guitar has the ability to acquire this general ability. Once the general ability is available to an agent, engaging in playing guitar depends on having the opportunity to play (in other words, being in a position to play the guitar). If an agent has a general ability and has the opportunity to execute the ability, the agent has the specific ability. For example, even though an agent can play the guitar, an agent cannot play if there is no guitar available or if the agent's hands are broken. Although an agent has a specific ability, it does not follow that the agent will always succeed in these cases. If an agent fails to execute an ability while having the specific ability, the agent shows a performance error, while the specific ability retains.

To illustrate how holding responsible through the nurturing stance works, Brandenburg (2018) introduces the case of Toddy. Toddy is a member of a fishing community. Even though Toddy can fish, he fears the dark water. He has the general ability to fish but lacks the specific ability to fish in deep, dark water. Individuals in the society hold Toddy responsible for his fishing inability even though he is not blameworthy. How, then, does holding responsible look, according to Brandenburg?

> They will perhaps encouragingly tell him, 'You have the capacity to do this, Toddy, we know you do!' or say more serious things to him, such as, 'You know things can't go on like this. We have many mouths to feed.' . . . In such scenarios we typically hold one another responsible without considering the other to be culpable and without responding (experiencing or expressing) with feelings of resentment, contempt or agent-directed anger. (2018: 15–16)

From the nurturing stance, holding responsible does not involve blame, but it involves a moral demand that has some of blame's properties although it does not fully qualify as blame.

This moral demand—like blame—involves the expression of moral disapproval of a person's action. Moral disapproval presupposes that the disapprover recognizes an event that violates a norm. All moral judgments require a norm system against which an event is detected as a violation. As Brandenburg (2018: 15)

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

highlights, it is unfair that Toddy benefits from his communities' fishing abilities while he does not develop his own skills.

In addition, the moral demand—like blame—involves the demand of behavior change. The fishing society wishes that Toddy would not have failed to work on his fishing skills in the past, and they want him to change his future behavior. The latter is indicated by what the fishing society is actually saying.

Finally, the moral demand—like blame—has some special force in contrast to mere descriptive statements, which potentially makes it unwelcome (costly, unsatisfying) for the receiver (Hieronymi 2004). Blame indicates that the blamed action has negative effects on social relationships. When Toddy receives his community's disapproval, he realizes the importance of standing in good relationships with mutual regard. Although some of the essential characteristics of blame are given in the Toddy case, Toddy is not blamed as the community does not show feelings of resentment or agent-directed anger.[1]

## 3. When Is It Justified to Hold a Nonblameworthy Agent Accountable?

If subjects adopt the nurturing stance, they assume that the person is not fully blameworthy because the person was not in control. However, from the nurturing stance, a moral demand targets the lack of effort to develop abilities. In the following, I want to show how such a demand contributes to the development of abilities (as in the case of rearing children in their adolescent years).

Under what conditions is it justified to hold somebody accountable from the nurturing stance although the person is not blameworthy? Brandenburg (2018) claims that blaming a patient in a psychiatry setting for uncontrollable behavior is unjustified. Assuming that the patient cannot control herself in certain contexts (specific inability), blaming is not an adequate response, according to Brandenburg. However, there is another demand that is valid:

> The demand 'to develop the skills necessary for norm compliance in stressful contexts' is a norm that the patient can understand when it is being communicated to him and he can 'communicatively participate' within this aspect of our normative practice, i.e. he can respond to this demand. (Brandenburg 2018: 19)

Even though the agent is not blameworthy, holding morally responsible can be appropriate if it serves the conversational meaningfulness of the moral-responsibility exchange. As Brandenburg describes, an apt target for the nurturing stance can understand the demand and can respond adequately. What are the preconditions for this?

[1] Arguably, the outlined blame-like moral demand qualifies as a type of blame. For example, Michael McKenna (especially, see 2012: 64–74, see also McKenna, forthcoming) argues that negative feelings are blame's characteristic vehicles, but they are not essential to it— we can blame each other with a smile on our faces. Additionally, McKenna argues that blaming does not imply an intention to hurt the target. However, I am not making this claim.

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

The target must be able to appreciate the received moral demand. Basically, the receiver has to understand the conversational rules and moral expectations. A young child, for example, does not understand how morality and moral responsibility works. In contrast to an adult, a child does not herself engage in the whole sophisticated practice of the moral community and does not understand it appropriately. A person who is held responsible executed a bad action, while the expectations and reactions of others were—at least implicitly—recognized.

For such cases, from the nurturing stance, holding a nonblameworthy agent responsible can be an appropriate response. In a similar sense, McKenna (2012) argues that blaming an agent for involuntary behavior is justified 'if the blamed agent has the ability to perform the free act of deciding or choosing to evaluate her own moral standpoint in regard to the received moral demand' (2012: 195). If the agent understands the moral demand toward her, which involves evaluating the demand, and the agent can, for instance, reply with an acknowledgment, then there is a meaningful responsibility exchange. Accordingly, from the nurturing stance, one agent justifiably holds another agent responsible if the latter can respond to the moral demand.

One way to show that an agent can respond to the moral demand is to rely on data (see section 4.3). If data show that an agent understands the moral demand and, in turn, changes her behavior in similar situations in the future, we have a strong reason to assume that the blamed agent can adequately respond to the moral demand.

For the case of implicitly biased behavior, I will show that holding responsible through the nurturing stance helps an agent to develop abilities without taking the objective attitude toward that agent. Brandenburg (2018) did not explain why and how it is possible that holding responsible helps to develop an ability. In the following, I will address both issues.

## 4.1 Why Does Holding Responsible Help to Develop an Ability?

The nurturing stance presupposes that the agent is not directly or even indirectly responsible for her inability. That means the person did not freely decide to impair an ability, such as when deciding to get drunk, otherwise she would be indirectly responsible and blameworthy. Nor did she intentionally omit to act in ways that resulted in a failure to have an ability in such a manner that would render her blameworthy for her omission. Instead, the nurturing stance demands that it is not the agent's own fault to have an inability. I will argue that agents can have an inability that they are not aware of. For instance, an agent knows that she has the general ability to control herself, but this does not mean that she has the specific ability to evaluate a job-application fairly. Moral demands have the power to correct the lack of self-awareness. Thereby, external responses can be crucial to help an agent overcome an inability.

Holding responsible and making certain moral demands serve as reliable social information for the agent who is held responsible. While self-insight is to perceive oneself accurately, we can gain self-insight via external or internal information. Thoughts, beliefs, values, and feelings count as internal information because agents can introspectively access them, and they partly reveal the agent's

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

personality traits. For example, if I owe a good friend a slight amount of money but feel reluctant about paying it back, I might conclude that I am stingy.

Social psychologists assume that we also learn about ourselves by observing how other persons react towards us ('reflected self'; see, for example, Tice and Wallace 2003). One problem for self-insight is that our self-perception is often at odds with other individuals' evaluations about us (Shrauger and Schoeneman 1979). Other persons can have privileged access to observable behavioral patterns. This explains why our self-conception does not always correlate with other persons' evaluations. An agent can be unaware or self-deceived about her traits (Johnson 1997). In contrast, other agents observe behavior that clearly indicates traits, for example, stinginess.

The famous Johari-window (Luft 1969), a psychological model from psychology of personality, illustrates this possible asymmetry in self-insight: there does exist information that a person herself cannot grasp but others do. For example, I can be ignorant about how jealous I act, even though my friends notice it. The Johari-window illustrates, furthermore, that this information may be delivered externally. Other persons can inform somebody about personality traits she is not aware of.

From the nurturing stance, an agent can receive a flow of information through the practice of moral responsibility. An agent who is held responsible through the nurturing stance lacks self-insight, the missing piece of information. For instance, an agent cannot recognize which ability is crucially underdeveloped or under which circumstances she is systematically failing. However, other persons, like clinicians, can recognize and communicate it.[2]

I want to end this section by remarking a parallel to proleptic blame (Bagley 2017; De Mesel 2020; Tsai 2017) although the moral demand from the nurturing stance does not fully qualify as blame. Briefly speaking, proleptic blame solves a dilemma for addressed blame: Would offenders appreciate a moral demand through a deliberative route from their existing motivations? 'If they would, their offense reflects a deliberative mistake, and blame's hostility seems unnecessary. If they would not, addressing them is futile, and blame's emotional engagement seems unwarranted' (Bagley 2017: 852). As Fricker (2016) explains, blame can remind of a reason whose force was already recognized, or blame can be proleptic and treat the target *as if* she recognized the reason, while the negative attitude that is directed at the target brings her to recognize the reason. In contrast, the nurturing stance does not involve a negative attitude, nor does it treat the target as if she is fully blameworthy. However, the illocutionary point of the moral demand, as for proleptic blame, is to move the target by new, shared reasons.

## 4.2 How Does Holding Responsible Help?

In the Toddy-case, Toddy can fish, but he cannot fish in dark waters. When Brandenburg speaks about underdeveloped capacities, this can mean different things in terms of abilities. I suggest that Toddy has the general ability to fish with

[2] Calhoun (1989) and Rini (2018, see also 2020) argue that although an agent is not blameworthy, there can be reasons for blame besides the practice of moral responsibility. For example, victims of discrimination in a sexist society have reasons to blame to raise awareness for injustice practices although offenders from a sexist society could be excused for epistemic reasons: reasons for blaming are not limited to fitting reasons.

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

average skills. However, the fishing society consists of agents who have advanced fishing skills, which is a distinct general ability. Agents can train and gain these abilities and, thereby, overcome specific inabilities. Similarly, I want to argue that there are average self-control abilities and high self-control abilities. Having a better self-control ability puts an agent in a position to control herself in certain situations (specific ability).

Analogously, suppose René is an average drummer. Because of his mediocre skills, he has problems with playing the drums when he sits in front of a different drum set. For instance, if the cymbals are in positions that René is not used to, he fails to play some rhythms. Accordingly, René has a specific inability to play drums on a different drum set. Arne, however, is a skilled drummer and does not have this specific inability. No matter where he plays the drums, he plays with high accuracy. Because of his distinct general ability, the ability to play the drums very well, he has more distinct specific abilities than René.

According to the self-determination theory (SDT; Deci and Ryan 2000), whether we are good or bad at an ability depends on our motivations. The SDT distinguishes between autonomous and controlled motivations. For this distinction, the reasons why we are actually doing something are important. Autonomous motivated actions are those that we fully identify with: Even though nobody appreciates our performance, we do it because we identify with the importance of these behaviors if we are autonomously motivated. This identification incorporates other aspects of the self as well. Internal values and goals are in harmony and cohere with an activity that is autonomously motivated. On the other end of the spectrum are (externally) controlled and motivated behaviors. Control motivated activities are alien to the self and its values, goals, and beliefs. People show these kinds of behaviors only because the environment demands them, including other persons' expectations and reactions. Importantly, data indicate that autonomously motivated behaviors are associated with high performance, consistency, and psychological well-being. In contrast, behaviors that are control-motivated induce poor performances and worse psychological health.

According to the SDT, an ability that is not internalized tends to be weak. In contrast, if an agent has autonomous motivation, then the reasons why she wants to use a skill align with her self-image, including her values, goals, and beliefs (see the subsection 'The Self in SDT' in Deci and Ryan 2000). In these cases, an agent masters the acquired general ability. An agent can simply be unaware of her shortcomings when executing an ability even though other people notice those shortcomings. In cases like these, other people know something about an agent that she does not know about herself (blind spot in the Johari-window). In fact, a study by Pronin, Lin, and Ross (2002) showed that individuals see the existence of cognitive and motivational biases much better in others than in themselves. Accordingly, the authors called this phenomenon the bias blind spot.

From the nurturing stance, nonblameworthy agents are held responsible to help them internalize a general ability through social cues. The weak performances so far, seen by other agents, can integrate within the agent's self-representation. If a moral demand leads to an internalization of a motivation for an ability, then we can expect better and more consistent performance. However, an agent needs the

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

right reasons to internalize the ability. A moral demand from the nurturing stance possibly contains and delivers these reasons.

## 5. Does Holding Responsible Help to Develop Self-control Abilities?

In the following, I want to suggest that holding implicitly biased agents responsible through the nurturing stance can be justified. I argue that nonblameworthy subjects are an apt target for the nurturing stance if the subjects can respond to the demand. The data shows that if subjects are held responsible for unintentional discrimination, they understand the demand and directly or indirectly acknowledge it. This is shown by actual behavior change in the future that is caused by the demand. Note that the nurturing stance is not a purely forward-looking account of responsibility, as suggested by Schlick (1972). From the nurturing stance, holding responsible is always anchored in past behavior. Accordingly, it is not justified to randomly hold people responsible to possibly improve their future behavior.

Before I turn to the data, some questions call for answers. First, how is implicitly biased behavior related to self-control abilities? Second, why is the nurturing stance relevant for the case of implicitly biased behavior? Third, does data show that holding biased agents morally responsible contributes to the development of self-control abilities?

### 5.1 Self-control and Implicit Biases

Having implicit prejudices or stereotypes does not guarantee that an individual will show discriminatory behavior. Meta-studies show that the overall effects-size of implicit attitudes is not overwhelming (Greenwald et al., 2009; Oswald et al., 2015) but small to medium. This does not show that indirect measurements lack predictive powers. In contrast, it shows something that is well known for explicit attitudes but is sometimes neglected for implicit attitudes: a single mental state usually fails to predict distinct behaviors under different circumstances (for example, see Wicker 1969). For instance, just because an individual has a positive affective explicit attitude toward meat, it does not follow that the individual eats meat—she can still be a vegetarian. Accordingly, behavior is the result of various mental states and environmental circumstances. Recently, Brownstein, Madva, and Gawronski (2020) argued that implicit attitudes' effects are strongly regulated by different mental states, such as motivations (Fazio and Olson 2014). However, metastudies usually do not take such moderating variables into account because then there would be too few studies available for the metastudy.

In fact, data (Butz and Plant 2009; Devine et al. 2002; LaCosse and Plant 2019; Plant and Devine 1998) shows that *intrinsically* motivated individuals have improved self-control and show less biased behavior, in contrast to extrinsically motivated individuals who show more implicitly biased behavior, as it is suggested by the SDT. Being intrinsically motivated to control prejudiced behavior means to value egalitarianism and to adjust beliefs, goals, and intentions accordingly to this acknowledged value. Extrinsically motivated individuals try only to avoid social

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

sanctioning; they control their behavior because of external social pressure. Having implicit prejudices does not automatically lead to implicitly biased behavior. If individuals are intrinsically motivated to control their prejudices, then the predictive powers of implicit prejudices decrease. On the other hand, subjects who lack any motivation to control their prejudiced reactions are more often influenced by implicit prejudices (Fazio and Olson 2014).

There are two distinct general abilities at play here: first, the ability to control for prejudices when externally motivated (EA) and, second, the ability to control for prejudices when internally motivated (IA). In principle, an agent can acquire both general abilities. With EA, an agent sometimes successfully controls her prejudiced reactions. Analogously, this general ability is like playing drums with basic skills. In contrast, an agent with IA is better and more consistent with demonstrating her ability—the agent has an advanced ability. A better general ability leads to more specific abilities. This means that the predictive powers of implicit prejudices decrease if an agent has the ability to control for prejudices when internally motivated (IA).

An agent can show implicitly biased behavior without noticing her self-control deficit (see section 3.2). The agent knows that she has the general ability (EA) to control herself, but she unintentionally discriminates against individuals because she was not self-aware of her weak abilities. She might be able to control her behavior when being surrounded by friends, but she does not have the specific ability, say, to act fairly toward people of color in the subway. Against the background of the nurturing stance, holding responsible is justified in these cases if it leads to improvement of self-control. In the next section, I want to address the question of why the nurturing stance is a valid option for the case of implicitly biased behavior.

## 5.2 Making the Case for the Nurturing Stance

There are cases of implicitly biased behavior for which agents are neither directly nor indirectly responsible, but nevertheless agents have the (type-2) capacity (as Brandenburg calls it) to do better. For these cases of implicitly biased behavior, the nurturing stance is a valid and natural option.

In a characteristic case of implicitly biased behavior, an agent acts unintentionally and is unaware of the action's moral characteristics. For example, while an agent intentionally evaluates a job application, she unintentionally discriminates against, say, Arab-Muslim individuals because of implicit prejudices (Rooth 2010). There are, however, other cases of implicitly biased behavior that are not characteristic cases. For instance, if an agent undergoes an Implicit Association Test, the agent recognizes that some sorting tasks, like sorting positive words to the category 'African-American', take more effort and create reaction-time deficits. If the sorting behavior automatically shows time delays, the behavior is implicitly biased, although the agent is *aware* of the effect. In the following, I will focus on characteristic cases of implicit biases.

In characteristic cases of implicit bias, agents are not directly morally responsible. Here, I agree with Levy (2014): If an agent is unaware of the actions' morally relevant

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

aspects, the agent cannot engage in self-control to act according to her acknowledged values, goals, and attitudes. In these cases, implicit attitudes undermine an agent's self-control (see also Holroyd 2012: 283). On the other hand, Madva (2017) argues that agents are aware of the implicit attitude's content (see, for example Gawronski, Hofmann, and Wilbur 2006). Arguably, some agents introspectively perceive a negative automatic gut reaction if they engage with an individual from a certain social group. For such cases, Madva argues, agents can be partly blameworthy for implicitly biased behavior. This, however, only holds if agents can access relevant aspects of implicit attitudes. In fact, implicit attitudes influence all kinds of perceptions (Xiao, Coppin, and Bavel 2016) and, for instance, lead agents to misclassify facial expressions (Hugenberg and Bodenhausen 2003). Although an agent can be aware of her implicit attitude's content, there is no reasonable ground for control if attitudes change relevant perceptual features of the situation. For such a case, if an agent discriminates against persons because the agent misperceived a facial expression, implicit attitudes undermined the agent's self-control regardless of whether the agent knows about her implicit attitude's content.

For some characteristic cases of implicit bias, agents are indirectly morally responsible. In much the same way as a doctor has the special duty to inform herself about types of cancer, so a committee that evaluates job applications has the special duty to be aware of possible biasing effects (Washington and Kelly 2016). Sometimes, although agents are unaware of their implicit biases, agents are responsible for acquiring relevant knowledge. If they fail to do so, they become blameworthy. For such cases, however, indirect moral responsibility depends on social roles and controllable situational aspects. In their absence, indirect control and indirect responsibility does not hold.

Holroyd (2012) argues that agents are indirectly responsible for the manifestations of implicit biases. For instance, she argues on the basis of psychological studies that agents can use the strategy of intention-implementation to control themselves indirectly. Agents can implement intentions that have built-in conditionals, such as 'When I am in the train, I will be friendly'. Furthermore, she argues that a lack of an internal motivation to control prejudiced reactions is an agent's own fault because agents can indirectly decide about it. However, this long-range control supposes that indirect control depends on direct control insofar that an agent was free to decide what to do at some point. This freedom is limited by some epistemic constraints (McKenna and Vadakin 2008). For instance, if a person made a free decision but had little reason to believe that this decision would lead to a bad deed or a bad character, the person is not blameworthy. I think that the indirect control strategies Holroyd refers to are undermined by these epistemic constraints. Agents are not psychology experts who know recent studies on fighting biases, nor do agents freely choose whether they develop an internal motivation to control prejudiced reactions. However, as some agents' blameworthiness is undermined for epistemic reasons, holding responsible from the nurturing stance is suited because hereby reasons are exchanged due to the meaningful moral responsibility exchange (see section 3).

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

Accordingly, there are characteristic cases of implicit bias for which agents are neither directly nor indirectly responsible. Nevertheless, agents can decrease implicit attitude's influences if they are internally motivated to control their prejudiced reactions (see section 4.1). That is, in case somebody is only externally motivated to control her prejudiced reactions and could improve self-control, the nurturing stance is a valid and natural option although the agent is neither directly nor indirectly responsible. In such a case, we do not have to decide between blaming or exempting, because 'when they [the biased agents] can become able to meet our norms, our attitude towards them in response to this transgression is not properly described as exempting' (Brandenburg 2018: 7). Now, does holding a non-blameworthy individual responsible for implicitly biased behavior contribute to overcome specific self-control inabilities?

## 5.3 Holding Responsible and How It Contributes to Self-control

Some authors worry that holding implicitly biased agents responsible might lead to backlash-effects and thereby could make things even worse (in psychology, see Baumeister and Campbell 1999; in philosophy, see Saul 2013; Vargas 2017). If holding responsible is ineffective, then the data would suggest that agents cannot respond to the interpersonal demand from the nurturing stance. If agents cannot respond to demands, holding them responsible is not justified (see section 3). However, if there is a meaningful responsibility exchange, then this should improve the subjects' motivation to control their prejudices.

One important study (Czopp, Monteith, and Mark 2006) shows that holding subjects accountable for using stereotypes leads to more behavior monitoring. In this study, participants solved a task on a computer while they had to chat with their task-partner, who was, in fact, an experimenter. When a participant solved the task and applied stereotypes to proceed, she received moral disapproval via the chat from the fake task-partner. Regardless of whether the message was high or low in threat, data shows that participants used fewer stereotypes in the following task. These are the responses the participants received:

> *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 Monteith, and Mark 2006: 788)

After receiving these messages, the participant had to solve a similar task (Experiment 1 & 2) and relied less on stereotypes. In Experiment 3, instead of solving a stereotype task, the participants had to express their explicit prejudices via the Attitudes Towards Blacks Scale (Brigham 1993). Here, participants indicated fewer prejudices, which is more evidence for their behavior monitoring

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

and their capability of controlling their behavior in a different situation (i.e., the skill is coherent across different tasks).

Besides, the study reveals important details about context. For example, Experiment 1 shows that for changing future behavior it does not matter whether the person is socially categorized as Black or White. It makes a difference regarding how much subjects dislike the one who is holding accountable: if that person is Black, subjects indicate more negative emotions toward the Black person than toward the White person. Most important, the study shows that the feeling of guilt is correlated with changing future behavior. However, one downside of this study is that it did not monitor how long the blaming effects last.

Chaney and Sanchez (2018) remedied this deficiency in another study, which successfully replicated the positive effect of holding accountable and, furthermore, measured the long-term effects of it. First, participants had to solve a stereotype task and afterward received the following message: 'I thought some of your answers seemed a little offensive. The Black guy wandering the streets could be a lost tourist. People shouldn't use stereotypes, you know?' (Chaney and Sanchez, 2018: 3) The study shows that recipients paid more attention to make stereotype-free judgments one week after the social confrontation. Accordingly, being held accountable has long-lasting effects on prejudiced agents.

In another study (Parker et al. 2018), participants were held accountable for unintentional sexism. Other studies showed that people are more defensive and insensitive for sexism than for racism (Gulker, Mark, and Monteith 2013). In this study, however, the researchers held subjects responsible with an evidence-based approach: The response came with data that clearly indicated that the participants behaved in a sexist manner. Thus, there was no way to rationalize the biased judgment post hoc, and this caused feelings of guilt. The study indicates that there is a positive correlation from guilt to the participants' intentions to monitor future behavior more carefully.

The studies mentioned so far did not involve holding agents morally responsible for implicitly biased behaviors (neither implicit nor explicit attitudes were measured before the participants showed biased behavior). In one study (Scaife et al. 2020), participants got blamed for having implicit prejudices. After the participants' implicit prejudices were measured, they received the following face-to-face feedback from an experimenter:

> You have just taken the shooter bias test, which is intended to measure differences in attitudes towards racial groups that you might not explicitly endorse. I'm afraid that the differences in your reaction times and shooting choices indicate you have negative implicit attitudes towards black people. Morally speaking, we would hope people don't have these kinds of attitudes. People who have these kinds of attitudes tend to behave in discriminatory ways, even if it is so subtle that you don't notice it. Overall, you are blameworthy for having these discriminatory attitudes and behaviours. As you probably know, it is morally unacceptable to have biased attitudes and behaviours; it would be quite normal to feel guilty about this; and to think about

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

> how to change these attitudes, or your behaviours to bring them in line with moral expectations. Later, in the debrief, we can talk more about techniques people have used to try to eliminate these bad attitudes. There'll also be the chance to ask any questions you may have. Now that you've got the results of this part of the study, we'll give you a moment to reflect on that, and then move on to the next part of the study. (Scaife et al. 2020: 4–5)

A follow-up indirect measurement showed that the blamed subjects did not have higher implicit prejudices. Furthermore, the study shows that subjects' motivation to monitor their future behavior more carefully improved, and that the subjects wanted to know more about implicit biases and how to overcome them. Accordingly, there is no backlash-effect, neither for implicit attitudes nor for motivations, when individuals are blamed for implicit prejudices.

The only study that potentially indicates backlash-effects for being blamed was conducted by Legault, Gutsell, and Inzlicht (2011). The researchers concluded that receiving a specific message leads to higher explicit and implicit prejudices. However, it is important to note that this study does not involve holding somebody morally responsible. In contrast, participants read different brochures, and afterwards the experimenters verified possible effects. In this study, two different groups received a distinct message, both based on the self-determination theory (SDT; Ryan and Deci, 2017). According to SDT, individuals have the tendency to react positively toward messages that acknowledge an individual's basic psychological needs (autonomy, competence, and social-relatedness). A message should acknowledge the personal freedom an individual has, it should inform, and it should rely on the existing social-relation appropriately. Legault, Gutsell, and Inzlicht (2011) created an antiprejudice message that contradicted the suggestions that follow from the SDT (excerpts):

> In today's society, you must control prejudice. In other words, being Canadian means having an anti-prejudiced attitude. For instance, The Human Rights, Citizenship and Multiculturalism Act prohibits discrimination in employment based on the grounds of race, color, ancestry, place of origin, religious beliefs. . . . Employers have an obligation to create a 'no prejudice' workplace, and companies face legal liability for workplace prejudice or discrimination. . . . The better we are at reducing prejudice, the more we are likely to fit in with today's anti-prejudice norms. . . . In today's multicultural society, we should all be less prejudiced. We should all refrain from negative stereotyping. It is, after all, the politically and socially correct thing to do, and it's something that society demands of us. (2011: supplemental material)

This message, which highlights external forces and denies freedom of choice, led to negative effects: After receiving this message, participants had more measurable explicit and implicit prejudices than before. However, the study does not show

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

that antiprejudice messages lead to negative effects in general. In contrast, the authors constructed a message that follows the basic assumptions of the SDT resulting in a message with the opposite effect. The following message led to lower explicit and implicit prejudices (excerpts):

> As a society, we hold the virtues of tolerance and nonprejudice in a very special place—they are important because they increase open-mindedness and social justice. Social justice is the vital ingredient in a free, fair, and peaceful society. When equality and equity among human beings are achieved, there is less reason for any group or individual to be unhappy. . . . It is also important to be nonprejudiced because it is so interesting to interact with and learn about people from other cultural and social groups. We live in a wonderful and diverse cultural community. That diversity makes our society great because it brings a wealth of knowledge and experience together. When we let go of prejudice, the rich diversity of society is ours to enjoy. . . . Not to mention, being open-minded is a real advantage to our mood and well-being. When there is less racial and cultural tension, people are happier and healthier, and better able to do the things they enjoy. . . . You are free to choose to value nonprejudice. Only you can decide to be an egalitarian person. . . . In today's increasingly diverse and multicultural society, such a personal choice is likely to help you feel connected to yourself and your social world.

This message delivers possible personal reasons for controlling prejudiced reactions. It does not rely on fear and pressure from external sources. Therefore, agents can identify with their behavior as originating from a personal source. In contrast, if an agent's autonomy is threatened, hostility toward the source of pressure arises, which explains the first message's negative effects.

Arguably, the SDT is biased toward Western civilization and its acknowledgement of autonomy. Influential psychological studies often rely on data from WEIRD (White, Educated, Industrialized, Rich, Democratic) people (Henrich, Heine, and Norenzayan 2010). While caution is in order, intercultural research shows that the more autonomous individuals are while realizing their own cultural values, the greater their psychological health and integrity (see Ryan and Deci 2017: ch. 22).

In sum, when holding a biased agent accountable, it does not matter whether the one who holds responsible is part of the target group if there is sufficient evidence for discrimination (Chaney and Sanchez, 2018; Czopp, Monteith, and Mark 2006; Parker et al., 2018; Scaife et al., 2020). Although data show that targets of prejudices have higher social costs (being disliked) if they hold agents accountable who harbor prejudices (Chaney and Sanchez, 2018; Czopp, Monteith, and Mark 2006), data also indicates that the targets monitor their future behavior more carefully. This holds as long as the agent does acknowledge that she actually showed prejudiced behavior. This speaks in favor of saying that agents sometimes lack self-awareness in terms of inabilities (see section 3.1). It was shown (Parker et al. 2018) that prejudiced agents can have a different degree of resistance to the

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

assumption that they behaved with prejudice. However, if there is sufficient evidence, prejudiced agents cannot rationalize their behavior and, therefore, feel guilty (Chaney and Sanchez, 2018; Czopp, Monteith, and Mark 2006; Parker et al. 2018). The feeling of guilt was positively correlated with the motivation to monitor behavior more carefully for prejudiced reactions.

Furthermore, it was shown that if the moral demand is structured in the right way, it can contribute to improve self-control abilities. To make somebody self-aware of her self-control shortcomings, the moral demand should acknowledge the person's autonomy and competence, as suggested by the SDT. The study by Legault, Gutsell, and Inzlicht (2011) supports this claim because it showed that messages that follow the SDT guidelines are the most efficient ones. They have the power to influence the internal motivation to control prejudiced reactions (Plant and Devine 1998), which is the most important dimension of self-control for prejudiced behavior (see section 4.1).

## 6. Conclusion: Holding Implicitly Biased Agents Responsible for Unintentional Discrimination Can Be Justified

The nurturing stance is a distinct part of our responsibility practices, one that must be distinguished from the reactive and the objective attitude. By taking the nurturing stance and analyzing its justifications, we understand philosophers' mixed intuitions regarding cases of implicit biases. Yes, characteristic cases of implicit biases undermine an agent's direct control abilities. However, in some sense an agent could have done better, and by taking the nurturing stance towards this agent, this sense could be realized *if certain conditions hold*:

1. The agent acted badly because of a specific inability.
2. The agent is neither directly nor indirectly responsible for the bad action, and the agent is neither directly nor indirectly responsible for having inabilities.
3. The agent who holds responsible assumes that the target is not fully blameworthy for the bad action.
4. The agent is not held responsible for the action but for her lack of effort in developing the weak ability.
5. The agent can respond to the moral demand.

Holding implicitly biased agents responsible through the nurturing stance can be justified. Let me recapture my argument against the background of the nurturing stance's characteristics.

- Regarding (1), it was shown that if subjects are not internally motivated to control their prejudiced reactions, they have specific self-control inabilities (section 4.1), which can lead to (more) unintentional discrimination.
- Regarding (2), subjects do not freely decide to be internally or externally motivated for self-control, which is why subjects are not indirectly responsible for having a specific self-control inability.

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

- Regarding (3), it was shown that there are characteristic cases of implicit biases, in which agents are neither directly nor indirectly morally responsible because implicit influences undermined the agents' control. Simultaneously, however, all agents who are not internally motivated could improve their self-control ability (see section 4.2).
- Regarding (4), holding implicitly biased agents responsible is justified if the moral demand targets the lack of effort to overcome self-control deficits.
- Regarding (5), it was shown that holding responsible for prejudiced behavior does not lead to backlash effects. Furthermore, it was shown that moral demands should be communicated in the right way because then they improve the internal motivation to control prejudiced reactions (section 4.1). Accordingly, moral demands can help to develop the self-control ability. On that basis, it was shown that subjects understood the moral demands and, furthermore, adjusted their behavior for relevant contexts in the future. This indicates that a meaningful responsibility exchange happened.

RENÉ BASTON
HEINRICH-HEINE-UNIVERSITÄT DÜSSELDORF
*rene.baston@hhu.de*

## References

Bagley, B. (2017) 'Properly Proleptic Blame'. *Ethics,* 127, 852–82.
Baumeister, R. F., and W. K. Campbell. (1999) 'The Intrinsic Appeal of Evil: Sadism, Sensational Thrills, and Threatened Egotism'. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology,* 3, 210–21.
Brandenburg, D. (2018) 'The Nurturing Stance: Making Sense of Responsibility without Blame'. *Pacific Philosophical Quarterly,* 99, 5–22.
Brigham, J. C. (1993) 'College Students' Racial Attitudes'. *Journal of Applied Social Psychology,* 23, 1933–67.
Brownstein, M. (2015) 'Implicit Bias'. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (spring 2015). http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/.
Brownstein, M., A. Madva, and B. Gawronski. (2020) 'Understanding Implicit Bias: Putting the Criticism into Perspective'. *Pacific Philosophical Quarterly,* 101, 276–307.
Butz, D. A., and E. A. Plant. (2009) 'Prejudice Control and Interracial Relations: The Role of Motivation to Respond Without Prejudice'. *Journal of Personality,* 77, 1311–41.
Calhoun, C. (1989) 'Responsibility and Reproach'. *Ethics,* 99, 389–406.
Chaney, K. E., and D. T. Sanchez (2018) 'The Endurance of Interpersonal Confrontations as a Prejudice Reduction Strategy'. *Personality & Social Psychology Bulletin*, 44, 418–29.
Czopp, A. M., M. J. Monteith, and A. Y. Mark. (2006) 'Standing up for a Change: Reducing Bias through Interpersonal Confrontation'. *Journal of Personality and Social Psychology,* 90, 784–803.
De Mesel, B. (2020) 'Addressed Blame and Hostility'. *Journal of Ethics and Social Philosophy,* 18, 111–19.
Deci, E. L., and R. M. Ryan. (2000) 'The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-determination of Behavior'. *Psychological Inquiry,* 11, 227–68.

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press

Devine, P. G., E. A. Plant, D. M. Amodio, E. Harmon-Jones, and S. L. Vance. (2002) 'The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond without Prejudice'. *Journal of Personality and Social Psychology,* 82, 835–48.

Dovidio, J. F., and S. L. Gaertner. (2000) 'Aversive Racism and Selection Decisions: 1989 and 1999'. *Psychological Science,* 11, 315–19.

Fazio, R. H., and M. A. Olson. (2014) 'The MODE Model: Attitude-Behavior Processes as a Function of Motivation and Opportunity'. In J. W. Sherman, B. Gawronski, and Y. Trope (eds.), *Dual Process Theories of the Social Mind*. Guilford Press.

Fischer, J. M., and M. Ravizza. (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Fricker, M. (2016). 'What's the Point of Blame? A Paradigm Based Explanation'. *Noûs*, 50, 165–83.

Gawronski, B., W. Hofmann, and C. J. Wilbur. (2006) 'Are "Implicit" Attitudes Unconscious?' *Consciousness and Cognition,* 15, 485–99.

Greenwald, A. G., T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji. (2009) 'Understanding and Using the Implicit Association Test: III. Meta-analysis of Predictive Validity'. *Journal of Personality and Social Psychology,* 97, 17–41.

Gulker, J. E., A. Y. Mark, and M. J. Monteith. (2013) 'Confronting Prejudice: The Who, What, and Why of Confrontation Effectiveness'. *Social Influence,* 8, 280–93.

Henrich, J., S. J. Heine, and A. Norenzayan. (2010) 'The Weirdest People in the World?' *Behavioral and Brain Sciences,* 33, 61–83.

Hieronymi, P. (2004) 'The Force and Fairness of Blame'. *Philosophical Perspectives,* 18, 115–48.

Holroyd, J. (2012) 'Responsibility for Implicit Bias'. *Journal of Social Philosophy,* 43, 274–306.

Holroyd, J., R. Scaife, and T. Stafford. (2017) 'What is Implicit Bias?' *Philosophy Compass*, 12, e12437.

Hugenberg, K., and G. V. Bodenhausen. (2004) 'Ambiguity in Social Categorization: The Role of Prejudice and Facial Affect in Race Categorization'. *Psychological Science,* 15, 342–45.

Hugenberg, K., and G. V. Bodenhausen. (2003) 'Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat'. *Psychological Science,* 14, 640–43.

Johnson, J. A. (1997) 'Units of Analysis for the Description and Explanation of Personality. In R. Hogan and J. A. Johnson (eds.), *Handbook of Personality Psychology* (San Diego: Academic Press), 73–93.

LaCosse, J., and E. A. Plant. (2019) 'Internal Motivation to Respond Without Prejudice Fosters Respectful Responses in Interracial Interactions'. *Journal of Personality and Social Psychology,* 119, 1037–56.

Legault, L., J. N. Gutsell, and M. Inzlicht. (2011) 'Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice'. *Psychological Science,* 22, 1472–77. https://journals.sagepub.com/doi/suppl/10.1177/0956797611427918/suppl_file/DS_10.1177_0956797611427918.pdf

Levy, N. (2014) *Consciousness and Moral Responsibility*. Oxford University Press.

Luft, J. (1969) *Of Human Interaction: The Johari Model*. Mayfield.

Madva, A. (2017). 'Implicit Bias, Moods, and Moral Responsibility'. *Pacific Philosophical Quarterly*, 99, 53–78.

McKenna, M. (2012) *Conversation and Responsibility*. Oxford University Press.

McKenna, M. (Forthcoming) 'Guilt & Self-Blame within a Conversational Theory of Moral Responsibility'. In A. Carlsson (ed.), *Self-blame and Moral Responsibility*. Cambridge University Press.

McKenna, M., and A. Vadakin. (2008) 'George Sher, In Praise of Blame: In Praise of Blame'. *Ethics,* 118, 751–56.

Mele, A. R. (2017) *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*. Oxford University Press.

Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock. (2015) 'Using the IAT to Predict Ethnic and Racial Discrimination: Small Effect Sizes of Unknown Societal Significance'. *Journal of Personality and Social Psychology,* 108, 562–71.

Parker, L. R., M. J. Monteith, C. A. Moss-Racusin, and A. R. Van Camp. (2018) 'Promoting Concern about Gender Bias with Evidence-based Confrontation'. *Journal of Experimental Social Psychology,* 74, 8–23.

Pickard, H. (2013). 'Responsibility without Blame: Philosophical Reflections on Clinical Practice'. In K. W. M. Fulford (ed.), *Oxford Handbook of Philosophy of Psychiatry* (Oxford: Oxford University Press), 1134–54.

Plant, E. A., and P. G. Devine. (1998) 'Internal and External Motivation to Respond without Prejudice'. *Journal of Personality and Social Psychology,* 75, 811–32.

Pronin, E., D. Y. Lin, and L. Ross. (2002) 'The Bias Blind Spot: Perceptions of Bias in Self versus Others. *Personality and Social Psychology Bulletin*, 28, 369–81.

Rini, R. (2018) 'How to Take Offense: Responding to Microaggression'. *Journal of the American Philosophical Association,* 4, 332–51.

Rini, R. (2020) *The Ethics of Microaggression*. Routledge.

Rooth, D.-O. (2010) 'Automatic Associations and Discrimination in Hiring: Real World Evidence'. *Labour Economics,* 17, 523–34.

Ryan, R. M., and E. L. Deci. (2017) *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. Guilford Publications.

Saul, J. (2013) 'Implicit Bias, Stereotype Threat, and Women in Philosophy'. In K. Hutchison and F. Jenkins (eds.), *Women in Philosophy: What Needs To Change?* (New York: Oxford University Press), 39–60.

Scaife, R., T. Stafford, A. Bunge, and J. Holroyd. (2020) 'To Blame? The Effects of Moralized Feedback on Implicit Racial Bias'. *Collabra: Psychology,* 6, 1–12. https://doi.org/10.1525/collabra.251.

Schlick, M. (1972) 'When Is a Man Responsible?' In P. W. Taylor (ed.), *Problems of Moral Philosophy: Introduction to Ethics* (2nd ed.) (Belmont, CA: Dickenson Publishing), 292–97.

Shrauger, J. S., and T. J. Schoeneman. (1979) 'Symbolic Interactionist View of Self-concept: Through the Looking Glass Darkly'. *Psychological Bulletin,* 86, 549–73.

Strawson, P. F. (1962) 'Freedom and Resentment'. In G. Watson (ed.), *Proceedings of the British Academy,* vol. 48 (Oxford: Oxford University Press), 1–25.

Tice, D. M., and H. M. Wallace. (2003) 'The Reflected Self: Creating Yourself as (you think) Others See You. In M. R. Leary and J. P. Tangney (eds.), *Handbook of Self and Identity* (New York: Guilford Press), 91–105.

Tsai, G. (2017) 'Respect and the Efficacy of Blame'. In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*, vol. 4 (Oxford: Oxford University Press), 248–75.

Vargas, M. R. (2017) Implicit Bias, Responsibility, and Moral. In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*, vol. 4 (Oxford: Oxford University Press), 219–47.

Washington, N., and D. Kelly. (2016) 'Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias'. In M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy, vol. 2: Moral Responsibility, Structural Injustice, and Ethics* (Oxford: Oxford University Press), 11–36.

Wicker, A. W. (1969). 'Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects'. *Journal of Social Issues*, 25, 41–78.

Xiao, Y. J., G. Coppin, and J. J. V. Bavel. (2016) 'Perceiving the World Through Group-Colored Glasses: A Perceptual Model of Intergroup Relations'. *Psychological Inquiry,* 27, 255–74.

https://doi.org/10.1017/apa.2021.45 Published online by Cambridge University Press