Received: 15 June 2021 | Accepted: 21 August 2021

DOI: 10.1111/josp.12442

ORIGINAL ARTICLE

JOURNAL OF SOCIAL PHILOSOPHY

Responsibility for implicitly biased behavior: A habit-based approach

Josefa Toribio[1,2]

[1]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

[2]Department of Philosophy, University of Barcelona, Barcelona, Spain

Correspondence
Josefa Toribio, Department of Philosophy, University of Barcelona, Montalegre 6-8, 08001 Barcelona, Spain.
Email: jtoribio@ub.edu

Funding information
Agència de Gestió d'Ajuts Universitaris i de Recerca, Grant/Award Number: 2017-SGR-63; Ministerio de Ciencia e Innovación, Grant/Award Number: PGC2018-095909-B-I00

1 | INTRODUCTION

Most of us would sincerely and justifiably express views that portray ourselves as unprejudiced agents, consciously committed to egalitarianism in all its forms. Yet, we are often surprised to discover that we behave in ways that betray such egalitarian, self-reported beliefs. We may thus very well find that it takes us longer to e.g., categorize women as surgeons or Mediterranean people as hard-working, even if we disavow sexism and racism. The mismatch between our explicit egalitarian beliefs and the stereotypically biased, often discriminatory, prejudicial behavior we often display makes it plausible to think that there must be mental states, other than our explicit beliefs, which are at least partially responsible for such a behavior. The expressions "implicit biases," "implicit attitudes," "implicit prejudices" or simply "prejudices," here taken as synonyms, are used to refer to such posits. Implicit biases are automatically activated by certain categorical cues in relevant contexts and permeate our perception, actions, and decisions. They are seldom the object of awareness,[1] which makes them especially resilient to change—resilient, but not unchangeable.

Some authors characterize implicit biases as traits, i.e., broad-track dispositions that manifest themselves not just in behavior, but also in e.g., attention patterns and emotions. Broad-track dispositions are thus like character traits (Machery, 2016). A construal of implicit biases in terms of dispositions is also at the heart of e.g., Céline Leboeuf's (2020) account, who regards them as perceptual habits (see below). More often, however, implicit biases are taken to be occurrent mental states with a content that reflects stereotypical social features and is evaluatively charged.

The central cases in the literature are those in which the content of such mental states is at odds with our explicit (self-reported) attitudes. It is, in fact, the low correlation between implicit and self-reported attitudes that has driven most of the research on implicit biases in social psychology and the reason I take the implicitly biased egalitarian as the central case in this paper.[2]

This lack of consensus about the ontology of implicit biases extends to the question of their representational structure when they are considered to be occurrent mental states. On the one hand, according to what is sometimes called "the associationist view," implicit biases are *sui generis* states, which, unlike explicit attitudes, have an associative structure (see e.g., Brownstein & Madva, 2012; Gendler, 2008a, 2008b; Madva, 2016; Toribio, 2018). On the other hand, on the so-call "propositionalist view," implicit biases are propositional states, i.e., either beliefs (De Houwer, 2014; Egan, 2011; Hughes & Barnes-Holmes, 2011; Mandelbaum, 2013, 2016; Mitchell & De Houwer, 2009) or states that fall short of being beliefs, but are, nevertheless, propositionally structured (Levy, 2015). There are also hybrid models such as Sullivan-Bissett's (2019) unconscious imagination model and models that fall outside the propositional/associationist dichotomy, such as Nanay's (2021) mental imagery account.

Regardless of their ontology, implicit biases are taken to have an affective component, with some associationist accounts such as Gendler's (2008a, 2008b) or Madva and Brownstein (2018) taking the affective element to be an essential part of their content. And, again, no matter how they are conceptualized, there is no question that implicit biases trigger an (often) discriminatory behavior toward members of the applicable social category—a discriminatory behavior that is directed toward these members just in virtue of their membership to that category. The focus of this paper is the nature of this discriminatory behavior. In particular, I aim to defend the view that the discriminatory behavior triggered by implicit biases is best understood as a type of habitual action—as a harmful, yet deeply entrenched, passively acquired, and socially relevant type of habit.[3]

As I have mentioned, the appeal to habits in the explanation of both implicit and explicit biases is not new. Helen Ngo (2016), following Merleau-Ponty's *Phenomenology of Perception*, develops an account of racism as "deeply embedded in our bodily habits of movement, gesture, perception, and orientation" (Ngo, 2016, p. 847). Also following the steps of Merleau-Ponty (1945/2002), Céline Leboeuf (2020) argues for a view of implicit biases as perceptual, embodied, social habits. Nathifa Green (2020) relies on the notion of habit to give an account of stereotype threat, i.e., the feeling of being at risk of underperforming in tasks which are traditionally associated with stereotypes of the social group to which a person belongs—be it gender, race, sexual orientation, nationality, or profession. Green argues that the anxiety and alienation triggered by stereotype threat is a form habit disruption. Stereotype threat consists in the feeling of being knocked out of a person's skills and habits normal "flow."[4]

I acknowledge the importance and depth of all these analyses, most of which are located within the phenomenological tradition or within the so-called embodied and embedded approach to cognition. I offer an argument for the thesis that the prejudicial behavior is best conceptualized as a type of passively acquired habit that completements, rather than compete, with similar views in the phenomenological tradition. This central claim is, to be clear, not a thesis about implicit biases themselves. It is a thesis about the behavior triggered by such implicit biases—understood as either mental or physical action. My argument here is an inference to the best explanation, premised on the view of habitual action that psychology and neuroscience have to offer. I elaborate on some of the marks of habitual action and suggest a more precise characterization of the central notions involved in such an account.

If the thesis that I defend here is plausible, and we accept that an important part of our life as egalitarian agents is nevertheless a life dotted with prejudicial habits—automatic actions of which we seem to be unaware—can we hold ourselves responsible for such prejudicial actions? Understanding prejudicial behavior as a passively acquired type of habit offers a diagnosis of the array of intuitions that inform extant approaches to responsibility for this type of actions. It also helps develop an ability-based account of the responsibility we have for prejudicial action that justifies a blaming response to the prejudiced agent while acknowledging that she is not blameworthy for her actions. This ability-based account of responsibility is neutral regarding any of the characterizations of implicit biases portrayed above.

The paper is organized as follows: In Section 2, I outline the main features traditionally associated with the notion of habit and highlight some of the neurobiological and neuropsychological processes that underwrite habitual action so as to pin down the type of habit that I consider to be relevant for characterizing prejudicial behavior. In Section 3, based on the previous considerations, I argue for the claim that prejudicial behavior is a passively acquired type of habitual action. In Section 4, I address the issue of the responsibility we have for our habits. As habits are ultimately susceptible of being controlled, the agent is responsible, and ought to be blamed, I argue, for failing to develop the ability to spot the kind of situations that require the exercise of her intellectual, moral, social, and prudential obligations. Being thus responsible, however, is consistent with the agent's not being blameworthy. For the automaticity of the blamed agent's implicitly biased behavior makes it unintentional relative to intellectual, moral, social, and prudential values that she already cares about.[5]

2 | HABITS: FROM CONCEPTS TO NEUROBIOLOGY AND BACK

Let me begin with Steve Matthews' (2017, p. 395) characterization of habit, which aims to capture what we typically mean by "habit" while avoiding theoretical commitments that belong to specific views on the topic:

> Habits are phenomena arising out of social contexts; they are acquired through repetition; they occur as a result of enduring structures within an agent; they are activated in the presence of environmental triggers; and this activation occurs automatically while remaining susceptible to control.

Matthews' characterization includes paradigmatic cases not just of bodily action but also of an important part of our mental lives, such as our "ways of thinking, of perceiving, of feeling, and of willing" (Matthews, 2017, p. 398). His is a plausible and neutral way of providing the first pass at a widely accepted folk-psychological view of habit.

In a more scientific vein, William James (1890) famously offers a view of habits in the fourth chapter of his *Principles of Psychology* as phenomena that belong to the sub-personal domain, separated from the intentional and rational realm of cognitive processing, and sharply distinguished from goal-directed behavior—an influential view within neuroscience and neuropsychology. Automaticity trumps flexibility in habits, according to James, while goal-directed action is controlled, flexible and sensitive to the contingencies in achieving the goal. Although there are by now more nuanced accounts when it comes to habits in humans, Anthony Dickinson (1985) was probably the first to account for this distinction in non-human animals. Through

the manipulation of the relationship between rats' trained behavior and its outcome, Dickinson established that habitual behavior, unlike goal-directed action, has two distinctive marks: outcome devaluation and action-outcome contingency degradation. Rats which have been trained to press a lever to get a reward are said to have developed a habit if they keep pressing the lever, at the same rate, even after the reward has been deployed of all its value (outcome devaluation) and even after the reward appears only randomly, without being correlated with the lever press (action-outcome contingency degradation).[6]

The mechanisms responsible for habitual behavior do not seem to change much across mammals, including humans, where they are not superseded by higher cortical areas. Instead, the most relevant mechanism is the so-called sensorimotor loop, which connects the somatosensory and motor cortex with two of the basal ganglia nuclei: the dorsolateral striatum and the posterior putamen—also heavily involved in other emotional, memory, pattern recognition and movement-related functions. When the basal ganglia circuits involving the putamen and the dorsolateral striatum increase their activation, i.e., when behavior becomes automatic and habitual, the medial prefrontal cortex, which governs explicit decision-making, inhibits its activity. Ann Graybiel, whose pioneering research on habits is as influential as Dickinson's, has provided ample evidence about the relevance of the dorsolateral striatum in habit formation (see e.g., Graybiel, 1998, 2008). She has shown that neural activation patterns in the dorsolateral striatum change as a result of an action becoming habitual. While the neurons in this region fire continuously throughout a new action or while learning a new task, once the action become habitual, these neural firings cluster together at the beginning and the end of the acquired habit, with no neural activation in between. This phenomenon, known as "chunking," makes the habit extremely difficult to break and gives it the robustness that makes habitual action insensitive to changes in response outcomes.

Repeatedly responding to a contextual cue is thus key for the development of habitual action. This is just a case of the more wide-ranging phenomenon known as Hebbian learning, i.e., the increase in synaptic strength that occurs between cells that repeatedly fire together. Through this sort of routine, the brain thus develops strategies that are computationally efficient by discharging cortical areas and hence, awareness, from the supervision of their performance. Neuroscience matches up here with a certain tradition in philosophy that has its origin in Felix Ravaisson's *Of Habit* (2008), who first formulated what he called "the double law of habit" as follows: "[t]he continuity or the repetition of passion weakens it; the continuity or repetition of action exalts and strengthens it" (Ravaisson, 2008, p. 49). More recently, Clare Carlisle (2014) has developed an account of habit around this idea. In a nutshell, the thought is that when an action becomes habitual, our sense of ourselves, the sensory feedback that comes with the action, becomes weaker while the repeated action becomes easier, swifter, and effortless. Yet, this sensory awareness, never fades away completely. Rather, as our actions become habitual, their automaticity impinges a second nature on us. That's why habit can be a "general and permanent way of being" (Ravaisson, 2008, p. 25). The work of influential figures such as Bordieu (1980/1990), Gibson (1979), or Merleau-Ponty (1945/2002) can also be situated within this tradition, where habits are portrayed as "dynamically configured stable patterns, strengthened and individualized by their enactment ... forming an integral part of individual embodied intentionality" (Barandiaran & Di Paolo, 2014, p. 6). With an emphasis on the deep relationship between habitual behavior and properties of the environment, the account of implicit biases as habits developed by Helen Ngo's (2016), Céline Leboeuf's (2020), and Nathifa Green's (2020), mentioned in the Introduction, should also be located within this view. But, as we have seen, even if briefly, the neurobiology of habits already places great emphasis on the relevance of the social environment for establishing

the way in which habits are acquired and maintained. When thinking about implicit bias or prejudicial behavior in terms of habits, the relevant environmental cues will obviously be socially relevant properties such as the color of our skin or our gender. Such environmental cues trigger the kind of automatic responses that bypass conscious awareness and become long-lasting and fixed in the way typically illustrated by habits. The neurobiology of habit also teaches us that the fast and automatic ways in which our implicit biases move us to act often grant us an epistemic advantage, especially under time constraints. The presence of such epistemic benefits is partially explained by the efficiency of the computational strategies involved in the creation of habits triggered by genuine, albeit unjust, regularities in our social environment.[7]

Now, two crucial qualifications are in order to complete my characterization of habits. First, it is important to distinguish between habits developed unreflexively, without even noticing, such as the habit of pushing up your glasses or biting your nails, and habits that we consciously develop and set ourselves to practice as part of a conscious decision, such as the habit of meditate for 10 min before going to sleep or brushing your teeth after eating chocolate. The latter play an important role in the acquisition of certain skills, and there is a sense in which most skills, like the skill of the expert piano player, are built upon the layers of habitual actions involved in regular practice. However, the explicit purposiveness of this kind of habits, based on explicit learning objectives, rules them out as theoretically interesting for our purposes. The habits of the former kind, the habits that we fall into unreflexively, are the type of habits that interest me here. My claim is that prejudicial behavior is this *type* of habitual action.

Second, the one feature that most commonly characterizes habits is their automaticity. Moors and De Houwer's (2006) offer a rich conceptual analysis of the notion of automaticity as an umbrella term constituted by a variable set of features, which are neither necessary nor sufficient. Automatic processes or actions often share, on their view, a subset of the following features: they are unintentional, goal-independent, uncontrolled/uncontrollable, autonomous, i.e., initiated and run unintentionally, without any intervening processing goal being able to stop or avoid them, purely stimulus driven, unconscious, efficient, i.e., requiring minimal attention thus triggering a subjective experience of effortlessness, and fast. Furthermore, the presence of these features in automatic processes and actions is *gradual* and *relative* to some standard, which acts as "a standard of comparison" (Moors & De Houwer, 2006, p. 321).

Just a word about the uncontrolled/uncontrollable feature of automatic actions. As we saw above, our pre-theoretical understanding of habits includes the idea that, although they are triggered automatically, they are still controllable. Bill Pollard (2006, 2010) argues that, even if habits are automatic, there is always the possibility of intervention, and it is this possibility of intervention that puts the subject in control and guarantees responsibility for their habits, since when the subject intervenes to alter the automatic course of action of a habit, they do it for a reason. Is there a contradiction here? Not really. The features of uncontrolled and uncontrollable as applied to the notion of automaticity capture both the lack of our intending to act in the way we do and the lack of our intending to be guided by any other sub-goals during the execution of the action such that they may change, interrupt, or prevent its completion. Habits, inasmuch as they are automatic actions are, in this sense, uncontrolled and uncontrollable, but they remain appropriate objects of agentive control when an overriding goal is present. As Wayne Wu (2016) suggests, it is a mistake to think of processes or actions are either automatic or controlled *simpliciter*. Actions as automatic or controlled *relative to* one of its features, i.e., relative to at least one of the possible correct characterizations of the action. For instance, my sitting 5 feet away from an African American man in a waiting room is automatic relative to my sitting 5 feet away, and relative to the

standard distance that I would sit, were I to do it in a room with only white people. Yet, my action is controlled since it is the result of my intention of sitting somewhere in the room.

I can now fully formulate the claim I will defend in the next Section. Prejudicial behavior is a passively acquired type of habitual action. This type of habitual action is automatic in the following sense: it is unintentional relative to at least one of its features, and it is, to some degree and relative to some standard, goal-independent, uncontrolled, autonomous, purely stimulus driven, unconscious, efficient, and fast.

3 | PREJUDICIAL BEHAVIOR AS A (TYPE OF PASSIVELY ACQUIRED) HABIT

Keith Payne's (2001) now classic weapon identification task is a good illustration of the prevalent implicit racism in North American society. This experiment shows that participants identify weapons much faster when primed with pictures of black faces as opposed to pictures of white faces. Participants are also more prone to misidentify tools as weapons when primed with pictures of black faces, again as opposed to pictures of white faces. If we go back to our explicative definition of habit and its neuroscientific counterpart offered in the previous Section, we can appreciate how well this case of implicit racism fits the bill of habitual action. First, just photographs of black faces are enough of an environmental cue for triggering the type of automatic response—misidentification of tools as guns—which bypasses conscious awareness and is supported by past exposure to stereotypes about African Americans. It is the kind of behavior that arises out of structural properties of our social environment. Second, the observed behavior is acquired through repeatedly acting in the same biased and prejudiced ways we have observed in our fellow citizens, thus contributing to such prevailing unjust regularities in our social environment. Repetition promotes an automatism that grants us epistemic advantages—even if it also often leads to mistakes. Third, the fast and inflexible nature of the responses in this experimental setup illustrates the existence of an enduring structure within the agent, a pathway between perception and action built out of the repetition of similar passively learnt prejudiced patterns of behavior. These new pathways that make the behavior so automatic, fast, and effortless are developed to avoid deploying extra cognitive energy in familiar (unjust) social environments. The sheer perception of a black face is enough to activate such pathways. Fourth, this activation is automatic, internalized without awareness but not without complete control. The behavior triggered by implicit racism illustrated by the above case is not a pure reflex, even if its habitual nature makes it rigid and difficult to control. Finally, an important reason for thinking about prejudicial behavior as a passively acquired habit is that habits, in general, and this type of habits in particular, change (mainly, if not only) through changes in the environment in which the habits are acquired. Although prejudicial behavior can occasionally be modulated by logical and evidential considerations (see e.g., Mandelbaum, 2016), these changes are often lab-bound and short-lived. Prejudicial behavior "in the wild" is much more resilient to change as a result of instruction or counter-stereotypical evidence—long-term changes tend to come, when they do, through environmental changes.

The workings of social structures with their inherent injustice express themselves in practices, habits, and conventions that have a direct impact on individuals from a very early age. We fall into the habit of prejudicial behavior, as we fall into any of our other passively acquired habits. Through constant interaction with our social environment, which includes individuals, groups, and institutions, we acquire distinct and enduring ways of acting that often allow us to

navigate more efficiently through our life—the navigation is more efficient because such ways of acting match the default models already present in the unjust society we live in. This process of habit formation shapes our social identity as agents and informs all aspects of our life, from the way we eat to the way treat others. Passively acquired habits are particularly definitional of this social identity, precisely because, by falling into them, we construe and develop a typical way of relating to our environment that it is not under the control of the beliefs and attitudes we endorse and it is, for that reason, very difficult to overcome at any given time. Characterizing prejudicial behavior as a type of passively acquired habitual action also fits nicely structural views of implicit biases such as Haslanger's (2006, 2015). According to Haslanger, the source of the problem with implicit biases is structural rather than individual, and the analysis and the strategies to understand and fight against the prejudicial actions they trigger cannot afford to ignore structural properties of our social environment. Even so, as Haslanger herself suggests, just focusing on structural social factors would not provide the whole picture either. For changes in patterns of behavior of the kind exemplified by prejudicial actions are mainly promoted by changes in experiential context, so that no change is likely in the absence of social change.

4 | HABITS AND RESPONSIBILITY

The theoretical landscape regarding the kind of responsibility (if any) that we have for prejudicial behavior is complex. My aim in this Section is not to settle the question of which of these views is the most plausible. Rather, I would like to show that treating prejudicial behavior as a passively acquired type of habit helps accounting for the intuitions that fuel these different views as well as showing where some of them go astray.[8] Of course, not all our habits thus acquired have a moral significance. Twirling one's hair, going to work through a particular route, or changing into sweatpants upon arriving home are prima facie neutral vis-à-vis responsibility. But the good-natured, cooperative roommate's habit of always leaving the lights on when exiting a room or never placing her dirty clothes in the laundry basket is not neutral. It is responsibility-laden, because in so acting, she places a burden on the other household members. Her actions can become the target of blaming attitudes. Keeping these mundane, simpler cases in mind, will facilitate making the point I defend below regarding prejudicial action: the roommate's habits show that she lacks a certain type of ability: the ability to recognize a situation as one requiring her to step back and exercise the kind of critical appraisal which would lead her to change her automatically generated course of action. Lacking this recognitional ability, however, co-exists with the good-natured roommate and, in general, with the egalitarian (but inattentive) agent's already caring about the moral, prudential, and social considerations that justify the blaming response. Blaming is justified as a way of bringing the blamed agent back into a realm of social and moral values that she already endorses. At the same time, the blamed agent is not blameworthy, precisely because she avows the relevant egalitarian principles beforehand.

Part of the difficulty of assessing whether agents are responsible for their prejudicial behavior and, if so, in which way, stems from two conflicting intuitions. On the one hand, it would seem too demanding to blame committed egalitarians for their prejudicial behavior when such actions are the result of implicit biases, whose influence escapes their radar of awareness and their direct control. Remember, our roommate's habit is the habit of a good-natured, cooperative roommate. This intuition fuels, for instance, Neil Levy's (2012) and Jennifer Saul's (2013) accounts, inherited from what it is known as volitional accounts of responsibility, according to which we are not responsible for that over which we have no control (Fischer & Ravizza, 1998). Since we do not have

any direct control over the influence of our implicit biases on our behavior, it follows from this view that we are not responsible for it.[9] If the white egalitarian's action of e.g., sitting far from the only African American man at the dentist office is conceptualized as a passively acquired habit, then it is an automatic action, it is something of which she is not aware and cannot directly control. Hence, it does not seem appropriate to blame her. Or, at least, it does not seem appropriate to blame her in the same way in which we would blame an overt racist behaving in the same way. At the same time, deeming this type of prejudicial behavior as a habit also allows to establish that, although unconscious and uncontrolled, this type of actions remain susceptible to supervision and censorship, as any other habit. A lot would depend on whether the agent is able to spot their own discriminatory behavior, perhaps alerted by others, and if so, like Saul claims, "they may ... be blamed if they fail to act properly on the knowledge that they are likely to be biased" (Saul, 2013, p. 55).[10]

This last point leads directly to the second, conflicting intuition mentioned earlier. The intuition that even the egalitarian who behaves prejudicially should be held responsible for her behavior because she contributes, albeit unconsciously, to discrimination and harm and hence she promotes the perpetuation of the social inequalities she explicitly disavows. Similarly, the cooperative, good-natured roommate appears responsible for her actions, as she contributes, albeit unconsciously, to more expensive electricity bills and fails to assume her chores in the household where she lives. But entering the domain of responsibility is not an easy task. The theoretical landscape is complex. To navigate through it, I find it useful the way Holroyd et al. (2017) divide questions about responsibility in three different groups. The first set of questions aims at determining whether the action to be considered reflects properties of the agents that could justifiably be attributed to them and hence put them at fault.[11] The second set of questions addresses whether agents should be blamed for actions which are thus attributed to them. These two sets of questions reflect Watson's (1996) classic distinction between responsibility as attributability and responsibility as accountability, where only the latter carries the weight of sanction and reactive attitudes, while the former is just a form of aretaic evaluation. These first two sets of questions are typically considered backward-looking regarding responsibility. Finally, the third set of questions has to do with the obligations agents incur when they behave in ways that put them at fault. This third set, by contrast, focuses the discussion on the forward-looking issue of the obligations we are expected to fulfil to counteract any harm done by our actions—regardless of whether they truly reflect something about our moral character and whether we should be considered blameworthy for the wrongs done.

Michael Brownstein's (2016) analysis illustrates an affirmative answer to the first set of questions. According to his version of attributionism, we are responsible for actions, omission, desires, and attitudes which we do not explicitly acknowledge as defining the person that we take ourselves to be, but which reflect, nevertheless, how we relate to others and, in so doing reflect our true self.[12] Brownstein considers prejudicial behavior a reflection of the agent's deep cares, understood not in the usual psychological sense, i.e., not as the evaluative judgments the agent identifies herself with, but rather in what he calls "the ontological sense," i.e., as the set of unreflective, often unacknowledged attitudes that manifest themselves in her dispositions to act—regardless of whether she consciously identifies herself with the attitude. Prejudicial behavior is thus, according to Brownstein, open to aretaic evaluations because it is attributable to the individual's *deep self*, which he characterizes as "a functional concept representing an agent's stable and identity-grounding attitudes" (Brownstein, 2016, p. 769). Yet, Brownstein remains neutral when considering whether attitudes such as blame or punishment are appropriate for egalitarian agents who exhibit prejudicial behavior.[13]

A characterization of prejudicial behavior as a passively acquired type of habit fits the intuition behind this understanding of responsibility as attributability inasmuch as the agent's cares, understood in this ontological sense, bypass the agent's psychologically motivating reasons. The feature of the automaticity of habits that attributionism seems particularly well suited to capture is its unintentionality and, again its unawareness. In viewing prejudicial behavior as a habit, we highlight the fact that agents who exhibit this type of behavior are not aware of obvious facts about themselves, such as the harm they cause to others. Yet, if we think about the simpler cases of habits, is it really appropriate not to blame our good-natured roommate for her actions? Jules Holroyd (2012) seems to push for a negative answer to this question—which enters the second of the categories mentioned above—when, after arguing that there are scenarios that justify our holding individuals responsible for their prejudicial behavior, she adds: "we cannot conclude that we ought not to regard individuals as liable for blame for *being influenced* by implicit bias" (Holroyd, 2012, p. 297). One of the points that Holroyd makes in order to vindicate attributions of blameworthiness to egalitarian agents who nevertheless exhibit prejudicial behavior is that, even though the implicit biases that trigger such a behavior may not be under the agent's direct rational control, they are under other kind of (indirect) control, i.e., interventions of different kinds that aim to stop or regulate the automaticity of the activation of implicit biases as a result of environmental triggers. What Holroyd and Kelly's (2016) call "ecological control" thus fits the idea that, although habitual action is automatic and hence uncontrollable at a certain level of execution, it remains susceptible to agential control of a different kind, as when the chain smoker ends up taking control of her habit by sheer power of the will, but also by rearranging her environment to avoid triggering cues long associated with smoking. Again, these considerations highlight the importance of the agent's development of a certain type of ability that allows her to spot situations that invite rethinking her automatic patterns of behavior—an ability that, we assume, can and hence ought to be developed.

Finally, Robin Zheng (2016) exemplifies one possible answer to the forward-looking question of the duties we acquire with regard to the harm which we may unconsciously inflect on others.[14] On Zheng's account, even if our prejudicial behaviour does not reflect badly on us and is thus not something for which we should be blamed, we are still responsible for it inasmuch as we incur corrective obligations to deal with the harm our wrongdoings inflict on the targets of our discrimination. We are indeed responsible for our prejudicial behavior, but this does not entail an assessment of our character; nor should the appropriate response involve blame. Instead, prejudicial actions call for "non-appraising responses," i.e., responses that involve "compensation, apology, and redress" (Zheng, 2016, p. 74).[15] Being responsible for our prejudicial behavior entails, on Zheng's account, appropriately being the target of others' expectations and demands regarding such actions, but in a non-appraisal-based form of moral criticism, i.e., without thereby inviting blame and punishment. If Zheng is right, we should not blame our good-natured but rather careless roommate. Instead, we should ask her to apologize and to remedy the situation by e.g., paying more for the electricity bill or assuming extra chores in the household. But is this enough? Furthermore, does the demand for an apology, compensation and redress not fulfil standard blaming criteria? After all, to demand an apology, compensation and redress involves expressing a moral disapproval as well as asking for a change in behavior; it involves censuring and devaluating the person that receives the demand, and it often triggers negative feelings in them—all characteristic marks of blame.

It is time to take stock. When assessing the responsibility that we have for prejudicial behavior, there is one issue that appears in almost any account of the theoretical landscape—one that fits nicely the characterization of prejudicial behavior as a type of passively acquired habit.

Passively acquired habits of the kind exemplified by prejudicial actions give us a powerful second nature—as we saw e.g., Ravaisson suggests. Our ways of moving through the social world and our forms of perceiving and attending become part of our identity as agents, shaping our behavior through family- and institution-based unchallenged practices. But, habits, even passively acquired habits, can be controlled, reflected upon, and changed. The automaticity of our habits is consistent with the development of the sensitivity and the abilities required to change them. Both aspects must thus inform any account of the responsibility we have over them. To say that an agent is responsible for her prejudicial behavior is to say, first, that the agent is capable of developing the required ability that allows her to step back and critically appraise her actions. The automaticity of our habitual behavior makes it, however, very difficult that the situations requiring this critical reflection simply "pop-out" to the agent, since *ex hypothesis* the explicit egalitarian is effectively fulfilling all her intellectual and moral obligations to the best of her knowledge and hence it does not appear to be blameworthy. Yet, we do hold agents responsible for their habits and we rightly do so because agents are capable of what we can call *forward-looking tuneability by reasons*. The idea here is very close to Dennett's (1984) take on the issue of control in relation to what he calls the "could have done otherwise" principle in discussions of free will. The "could have done otherwise" principle can only be sensibly interpreted, Dennett claims, as the possibility of a properly functioning agent modifying her actions in the future as the result of being prompted to corrections by the provision of training or feedback. Learning is thus of the essence. Someone "could have done otherwise" only if she is able to learn from the particular outcome of her actions; only if she is "cognitively tuneable" so as to act differently when facing similar situations in the future (Dennett, 1984, pp. 139–144). The requisite *kind* of tuneability is, to be clear, tuneability, by the exchange of reasons and arguments.

Regarding an agent's responsibility for her prejudicial behavior as a type of habit thus allows us to introduce a view of responsibility that resembles Michael McKenna's (2012) conversational stance or Daphne Brandenburg's (2018) nurturing stance, but which, importantly, still makes the prejudiced agent an appropriate target for blame. Considering prejudicial behavior as a habit, and hence as an action that is ultimately susceptible of being controlled, justifies the practice of blaming since the prejudiced agent retains the kind of agential control that makes her capable of a forward-looking tuneability by reasons—reasons that she already acknowledges as fitting reasons. This notion of blame is close to the notion of proleptic blame developed by Regina Rini in her (2020) *The Ethics of Microaggression*. Rini's central idea of proleptic blame draws on Bernard William's work, who introduces the concept to characterize scenarios in which it is not so clear that the person behaving prejudicially really cares about the relevant moral considerations, so the blaming is done in the hope of getting her to care about them. Proleptic blame, as Rini describes it, leads the blamed agent to value such relevant moral considerations, but it does so somewhat indirectly—identifying something the blamed person already cares about, such as gaining or maintaining the esteem of those, the blamer(s), whom she already respects. My account, although similar in having a forward-looking slant, places the blamed agent already within the realm of the relevant moral considerations, i.e., the blamed agent already possesses the right kind of motivation to acknowledge the relevant values, but the automaticity of her habits makes her careless and inattentive. In keeping with Rini's approach, however, the blaming response need not be and, I suggest, should not be expressed through anger or contempt. Instead, it is best cashed out in terms of an invitation for the prejudiced agent to reflect on her actions, to encourage her to develop an ability that has been compromised—the ability to spot the kind of situations that requires an exercise of her (intellectual, moral, social, or prudential) obligations. The appropriateness of the blaming response depends on the agent's being thus tuneable. The

key point of my account is that blaming the egalitarian agent who nevertheless behaves prejudicially is consistent with the agent not being blameworthy. Blaming is an effective tool for bringing the blamed agent back into a realm of values that she already cares about. Yet, she is, for thus caring about the right values, not blameworthy. Considering prejudicial behavior as a habit, and hence as automatic, helps to appreciate the consistency of this combo, because that the action is automatically entails that it is unconscious, goal-independent, and unintentional relative to its intellectual, moral, social, or prudential dimension.[16]

Let me finish with two general corollaries of the account of responsibility for prejudicial action developed here. First, since a lot depends on agents being in a position to know about the existence and prevalence of implicit biases—and not just on their knowledge about e.g., the prevalence of racism or sexism—and also, given the central role given to agents being capable of developing an ability to spot the situations that require the exercise of their moral, intellectual, social, and prudential obligations, it may look like only cognitively sophisticated and socially sensitive agents could be held responsible when they behave prejudicially. There is some truth in this conclusion. Both responsibility and the appropriateness of blame depends on the agent's being forward-looking tuneable by reasons in the way described above. This seems to suggest that just highly reflective, socially sensitive, and intellectually sophisticated subjects should be the target of both responsibility and blame. However, we should not commit unreservedly to an unduly under-intellectualized, somewhat condescending view of the average citizen. As Hahn et al.'s previously mentioned (2014) study shows, extremely cognitively unsophisticated subjects seem surprisingly able to predict how their implicit biases will influence their behavior in a wide range of contexts and across different experimental conditions, even when they are told very little about the test or about what implicit biases are. The message to take home here, I suggest, is that responsibility comes in degrees, for thoughtfulness, socially sensitivity, and intellectual sophistication also come in degrees, so those of us who are better informed about the unintentional harm we can inflict on our fellow citizens must take greater responsibility for our actions.

Second, the main target of this paper is prejudicial behavior caused by implicit bias that is directed toward others, and the relevant notion of responsibility developed here has its home in the realm of interpersonal relationships. This, of course, does not exclude scenarios in which agents behave prejudicially toward members of their own group. Implicit misogyny is, for instance, not uncommon among self-declared non-sexist women, and my account applies to such cases in the same way in which it applies to any other kind of prejudicial behavior. It could be argued, however, that there are also occasions in which what is going on is not the stigmatization of others through our actions—whether they belong or not to our social group—but a form of self-stigmatization, as when implicit biases manifest themselves as stereotype threat. If so, it would be harder for the agent to intervene upon herself in order to reflect on her own actions in the way suggested by the forward-looking tuneability transaction I suggest.[17]

We must tread carefully here though. My account requires the prejudiced agent, the habit-driven agent, to be able to break the automaticity of her old reactive patterns and reflect on the relevant actions—the actions that cause harm to others. But the agent who experiences stereotype threat feels her expert performance disrupted as a harm, and she feels that way as a result of other people's prejudices, not hers. As Nathifa Green (2020) reminds us, "[o]n the receiving end of that harm, stereotype threat is a consequence of bias, not a cause" (Green, 2020, p. 147). Even if stereotype threat could be considered a form of self-stigmatization, we must not get confused about whose responsibility is at stake here. It is the responsibility of those whose stereotypes cause the threat through their prejudicial actions, and my account would apply seamlessly here.

Shifting the responsibility holder onto the agent experiencing stereotype threat would be not only a misapplication of my account, but also morally wrong, as Green also reminds us.

Research for this chapter was supported by MICINN/AEI/FEDER, EU, under grant agreement PGC2018-095909-B-I00, and AGAUR, under grant agreement 2017-SGR-63.

CONFLICT OF INTEREST

I have no conflict of interest to declare.

ORCID

Josefa Toribio https://orcid.org/0000-0002-5180-6755

ENDNOTES

[1] Recent meta-analyses in social psychology suggest, however, that we may be more aware of the *content* of implicit biases than it has been previously assumed. For instance, Hahn et al. (2014) have shown that subjects are remarkably good at predicting their results on tests that measure implicit biases (see also Gawronski et al., 2006; Hall & Payne, 2010). Also, when subjects are experimentally forced to take e.g., their gut reactions toward gay people as indicators of their attitudes, the gap between implicit and explicit attitude measures is narrower (Ranganath et al., 2008). These results need not mean that they are aware of the content of their bias. It may just be that there are confused about what they should consider an implicit bias in the first place, as Hall and Payne (2010) suggest. Finally, lack awareness of the content of implicit biases is different from the lack of awareness we have of their *source*, and also different from our failing to be aware of their *impact* on other mental states, psychological processes, or behavior (Gawronski et al., 2006, p. 486). Unawareness of the impact of implicit biases remains unchallenged.

[2] We cannot rule out, however, what Holroyd (2016) calls 'harmony cases', in which explicit and implicit attitudes are aligned. Holroyd's case of the explicit racist, who is strongly motivated to adhere to egalitarian principles and explicitly forms the intention of acting in a non-discriminatory way in a hiring process, yet fails to do, is one of those cases (Holroyd, 2016, p. 160). For reasons that will become clear at the end of the paper, even if, occasionally, the explicit racist's prejudicial behavior could justifiably be attributed to her implicit bias, the blaming response does not exclude blameworthiness, as it does, I argue, in standard cases of committed egalitarians' prejudicial actions.

[3] From now on I will use 'prejudicial behaviour' or 'prejudicial action' to refer to the behaviour that results from implicit biases, not just any kind of prejudicial behaviour—not the behaviour of the explicitly committed racist, sexist or homophobe, when it is triggered by their explicit biases. Of course, any action is the result of a combination of factors, so to talk about behaviour triggered by implicit biases is an oversimplification. We should also keep this in mind. Michael Brownstein (2016) uses the acronym 'BEIB' ("behavioural expression of implicit bias"), and Luc Faucher (2016) uses the acronym 'BEIA' ("behavioural expression of implicit attitudes") to capture the same idea.

[4] In psychology, Devine and collaborators have occasionally characterized social group stereotypes as habits (e.g., Carnes et al., 2015; Devine, 1989) but their treatment is almost anecdotal.

[5] My account thus resembles Regini Rini's (2020) work on the type of blame deserved by those of us who commit microaggressions, which are just instances of prejudicial behaviour as characterized in this paper, i.e., "small act[s] of insult or indignity, relating to a person's membership in a socially oppressed group, which seems minor on its own but plays a part in significant systemic harm" (Rini, 2020, p. 13). In Section 4, I shall discuss in which way my account shares some features, but it is ultimately different from Rini's account. I thank one of the referees for this Journal for having drawn my attention to Rini's work.

[6] When it comes to human behaviour, however, we should be wary of establishing a sharp distinction between habits and goal-directed action though. Some habits are goal-directed, as it is, for instance, tying one's shoelaces or driving home through the same route every day. On some philosophical views, such as Aristotle's first philosophical treatment of the concept, habits are in fact acquired dispositions that help the agent to improve their performance in the pursuit of a goal.

[7] Gendler (2008b, 2011) points out that this poses a difficult normative dilemma for us. The more we know about our social environment, the more likely it is that we act in biased and prejudiced ways. So, we must either ignore genuine socially relevant regularities, thus incurring into epistemic costs, or we have to deploy extra cognitive energy in suppressing and controlling the readily available information that we get from our social environment. Katherine Puddifoot (2017) argues that there is no real dilemma here and that automatic responses that ignore statistically relevant social inequalities often bring epistemic benefits that offset the obvious epistemic costs they carry. My point here, however, is factual rather than normative, i.e., such *egalitarian* responses, as Puddifoot calls them, are computationally costly, and seldom as fast and automatic as it would be desirable.

[8] Henceforth, whenever talking about habits, it will short for 'passively acquired habits.'

[9] Of course, volitional arguments differ greatly among themselves. Levy's position, for instance, stems from an argument about the nature of implicit biases themselves. He argues that we should give up trying to reason about responsibility for implicit biases and the behaviour they trigger based on pre-theoretical intuitions since implicit biases are what he calls 'patchy endorsements', i.e., neither beliefs nor associations, and our folk-psychological reasoning gets off-track when dealing with such a theoretical notion (Levy, 2015). Saul (2013) also qualifies her volitional view with an important proviso: individuals who learn about the likelihood of their behaviour being influenced by implicit biased without implementing measures to eliminate such an influence should be blamed.

[10] Saul's remark thus highlights the observation I made above about the agent's ability to recognise a situation as calling for appraisal. I'll return to this issue below.

[11] Of course, the same holds when considering responsibility for actions that reflect well on the agents and hence when the issue is whether agents should be credited for such actions. I focus here only on the negative side of responsibility for obvious reasons.

[12] What Susan Wolf (1990) calls 'Real Self' Views of responsibility.

[13] Talbert (forthcoming) warns us about thinking that attributionism is only concerned with attributability as a lesser kind of responsibility, one that does not invite responses such as blame or reactive attitudes such as resentment. For contemporary attributionists, of the kind represented by e.g., Hieronymi (2014), Scanlon (1998), Smith (2012, 2015) or Talbert (2016) himself, however, "*attributability is enough for accountability*" (Talbert, forthcoming, p. 7). Being responsible thus entails, on what we can call the strong reading of attributionism, being open to reactive attitudes such as blame or resentment. My own proposal takes this appreciation on board (see below).

[14] Zheng labels this forward-looking sense of responsibility "accountability," but the label here is confusing as it does not match Watson's original distinction.

[15] Again, I compress Zheng's view here enormously. She does admit that to get off the hook from the point of view of responsibility (in the attributionist sense), two conditions have to be met: "(1) The agent would not upon reflection endorse the influence of the difference-making implicit bias on her action. (2) The agent has done what she can reasonably be expected to do with respect to avoiding and responding to the implicit bias" (Zheng, 2016, p. 72). I just take these conditions to be part and parcel of what we take an egalitarian but implicit biased agent to be.

[16] Although blaming is consistent with nonblameworthiness in the way described, it does not necessarily exclude blameworthiness. So-called 'harmony cases' described above illustrate this possibility. For the e.g., explicit racist that, on occasion, ends up behaving prejudicially, due to her implicit rather than their explicit bias, is wilfully committed to antiegalitarian values, even if she sporadically holds egalitarian intentions. The development of the right kind of ability through a forward-looking process of tuneability by reasons will be thwarted in such cases, if, as it seems likely, the blamed "harmonious" racist fails to experience the justificatory force of the reasons involved in such forward-looking blaming process.

[17] I thank one of the referees for this Journal for pressing me on this point.

REFERENCES

Barandiaran, Xabier E., and Ezequiel A. Di Paolo. 2014. "A Genealogical Map of the Concept of Habit." *Frontiers in Human Neuroscience 8*, Article 522. https://doi.org/10.3389/fnhum.2014.00522.

Bourdieu, Pierre. 1980/1990. *The Logic of Practice*. Palo Alto, CA: Stanford University Press.

Brandenburg, Daphne. 2018. "The Nurturing Stance: Making Sense of Responsibility without Blame." *Pacific Philosophical Quarterly 99*(S1): 5–22. https://doi.org/10.1111/papq.12210.

Brownstein, Michael. 2016. "Attributionism and Moral Responsibility for Implicit Bias." *Review of Philosophy and Psychology 7*: 765–86. https://doi.org/10.1007/s13164-015-0287-7.

Brownstein, Michael, and Alex Madva. 2012. "Ethical Automaticity." *Philosophy of the Social Sciences 42*(1): 67–97. https://doi.org/10.1177/0048393111426402.

Carlisle, Clare. 2014. *On Habit. Thinking in Action*. London and New York: Routledge.

Carnes, Molly, Patricia G. Devine, Linda Baier Manwell, Angela Byars-Winston, Eve Fine, Cecilia E. Ford, Patrick Forscher, Carol Isaac, Anna Kaatz, Wairimu Magua, Mari Palta, and Jennifer Sheridan. 2015. "Effect of an Intervention to Break the Gender Bias Habit for Faculty at one Institution: A Cluster Randomized, Controlled trial." *Academic Medicine: Journal of the Association of American Medical Colleges 90*: 221–30.

De Houwer, Jan. 2014. "A Propositional Model of Implicit Evaluation." *Social and Personality Psychology Compass 8*(7): 342–53. https://doi.org/10.1111/spc3.12111.

Dennett, Daniel. 1984. *Elbow Room. The Varieties of Free Will Worth Wanting*. MIT Press.

Devine, Patricia G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology 56*: 5–18. https://doi.org/10.1037/0022-3514.56.1.5.

Dickinson, Anthony. 1985. "Actions and Habits: The Development of Behavioural Autonomy." *Philosophical Transactions of the Royal Society London Biological Sciences 308*: 67–78.

Egan, Andy. 2011. "Comments on Gendler's 'The Epistemic Costs of Implicit Bias'." *Philosophical Studies 156*: 65–79. https://doi.org/10.1007/s11098-011-9803-5.

Faucher, Luc. 2016. "Revisionism and Moral Responsibility for Implicit Attitudes." In *Implicit Bias and Philosophy. Volume II: Moral Responsibility, Structural Injustice, and Ethics*, edited by Michael Brownstein and Jennifer Saul, 115–45. Oxford, New York: Oxford University Press.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Gawronski, Bertram, Wilhelm Hofmann, and Christopher J. Wilbur. 2006. "Are 'Implicit' Attitudes Unconscious?." *Consciousness and Cognition 15*: 485–99.

Gendler, Tamar Szabó. 2008a. "Alief and Belief." *The Journal of Philosophy 105*(10): 634–63. https://doi.org/10.5840/jphil20081051025.

Gendler, Tamar Szabó. 2008b. "Alief in Action (and Reaction)." *Mind and Language 23*(5): 552–85. https://doi.org/10.1111/j.1468-0017.2008.00352.x.

Gendler, Tamar Szabó. 2011. "On the Epistemic Costs of Implicit Bias." *Philosophical Studies 156*: 33–63. https://doi.org/10.1007/s11098-011-9801-7.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Graybiel, Ann M. 1998. "The Basal Ganglia and Chunking of Action Repertoires." *Neurobiology of Learning and Memory 70*: 119–36. https://doi.org/10.1006/nlme.1998.3843.

Graybiel, Ann M. 2008. "Habits, Rituals and the Evaluative Brain." *Annual Review of Neuroscience 31*: 359–87. https://doi.org/10.1146/annurev.neuro.29.051605.112851.

Green, Nathifa. 2020. "Stereotype Threat, Identity, and the Disruption of Habit." In *Implicit Bias: knowledge, Justice, and the Social Mind*, edited by Erin Beeghly and Alex Madva, 134–52. London and New York: Routledge.

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.

Hall, Deborah L., and B. Keith Payne. 2010. "Unconscious Influences of Attitudes and Challenges to Self-Control." In *Self Control in Society, Mind, and Brain*, edited by Ran Hassin, Kevin N. Ochsner, and Yaakov Trope, 221–42. Oxford: Oxford University Press.

Haslanger, Sally. 2006. "What Good are our Intuitions? Philosophical Analysis and Social Kinds." *Proceedings of the Aristotelian Society 80*(1): 89–118.

Haslanger, Sally. 2015. "Social Structure, Narrative and Explanation." *Canadian Journal of Philosophy 45*(1): 1–15.

Hieronymi, Pamela. 2014. "Reflection and Responsibility." *Philosophy and Public Affairs 42*(1): 3–41. https://doi.org/10.1111/papa.12024.

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

Holroyd, Jules. 2016. "What Do We Want from a Model of Implicit Cognition?" *Proceedings of the Aristotelian Society 116*(2): 153–79.

Holroyd, Jules, and Dan Kelly. 2016. "Implicit Bias, Character and Control." In *From Personality to Virtue: Essays on the Philosophy of Character*, edited by Alberto Masala and Jonathan Webber, 106–34. New York, USA: Oxford University Press.

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

Hughes, Sean, Dermot Barnes-Holmes, and Jan De Houwer. 2011. "The Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives." *The Psychological Record 61*(3): 465–98. https://doi.org/10.1007/BF03395772.

James, William. 1890. *Principles of Psychology*. New York: Henry Holt.

Leboeuf, Céline. 2020. "The Embodied Biased Mind." In *Implicit Bias: knowledge, Justice, and the Social Mind*, edited by Erin Beeghly and Alex Madva, 41–56. London and New York: Routledge.

Levy, Neil. 2012. "Consciousness, Implicit Attitudes, and Moral Responsibility." *Noûs 48*(1): 21–40. https://doi.org/10.1111/j.1468-0068.2011.00853.x.

Levy, Neil. 2015. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs 49*(4): 800–23. https://doi.org/10.1111/nous.12074.

Machery, Eduard. 2016. "De-Freuding Implicit Attitudes." In *Implicit Bias and Philosophy. Volume I: Metaphysics and Epistemology*, edited by Michael Brownstein and Jennifer Saul, 104–29. Oxford, New York: Oxford University Press.

Madva, Alex. 2016. "Why Implicit Attitudes are (Probably) not Beliefs." *Synthese 193*: 2659–84. https://doi.org/10.1007/s11229-015-0874-2.

Madva, Alex, and Michael Brownstein. 2018. "Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind." *Noûs 52*(3): 611–44.

Mandelbaum, Eric 2013 "Against Alief." *Philosophical Studies 165*: 197–211.

Mandelbaum, Eric. 2016. "Attitude, Association, and Inference: On the Propositional Structure of Implicit Bias." *Noûs 50*(3): 629–58.

Matthews, Steve. 2017. "The significance of habit." *Journal of Moral Philosophy 14*: 394–415. https://doi.org/10.1163/17455243-46810073.

McKenna, Michael. 2012. *Conversation and Responsibility*. Oxford, New York: Oxford University Press.

Merleau-Ponty, Maurice. 1945/2002. *Phenomenology of Perception*. Translated by Donald A. Landes. Routledge.

Mitchell, Chris J., Jan De Houwer, and Peter F. Lovibond. 2009. "The Propositional Nature of Human Associative Learning." *Behavioral and Brain Sciences 32*(2): 183–98. https://doi.org/10.1017/S0140525X09000855.

Moors, Agnes, and Jan De Houwer. 2006. "Automaticity: A Theoretical and Conceptual Analysis." *Psychological Bulletin 132*(2): 297–326. https://doi.org/10.1037/0033-2909.132.2.297.

Nanay, Bence. 2021. "Implicit Bias as Mental Imagery." *Journal of the American Philosophical Association*. https://doi.org/10.1017/apa.2020.29.

Ngo, Helen. 2016. "Racist Habits: A Phenomenological Analysis of Racism and the Habitual Body." *Philosophy and Social Criticism 42*(9): 847–72. https://doi.org/10.1177/0191453715623320.

Payne, B. Keith. 2001. "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality and Social Psychology 81*(2): 181–92.

Pollard, Bill. 2006. "Explaining Actions with Habits." *American Philosophical Quarterly 43*(1): 57–69.

Pollard, Bill. 2010. "Habitual Actions." In *A Companion to the Philosophy of Action*, edited by Timothy O'Connor and Constantine Sandis, 74–82. Chichester: Wiley-Blackwell.

Puddifoot, Katherine. 2017. "Dissolving the Epistemic/Ethical Dilemma over Implicit Bias." *Philosophical Explorations 20*(S1): S73–S93. https://doi.org/10.1080/13869795.2017.1287295.

Ranganath, Kate A., Colin Tucker Smith, and Brian A. Nosek. 2008. "Distinguishing Automatic and Controlled Components of Attitudes from Direct and Indirect Measurement Methods." *Journal of Experimental Social Psychology 44*: 386–96. https://doi.org/10.1016/j.jesp.2006.12.008.

Ravaisson, Felix. 2008. *Of Habit. (1837) De l'Habitude*. Translated by Clare Carlisle and Mark Sinclair. London and New York: Continuum Books.

Rini, Regina. 2020. *The Ethics of Microaggression*. London and New York: Routledge.

Saul, Jennifer. 2013. "Implicit Bias, Stereotype Threat and Women in Philosophy." In *Women in Philosophy: What Needs to Change?* edited by Katrina Hutchison and Fiona Jenkins, 39–60. Oxford: Oxford University Press.

Scanlon, Thomas Michael. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Smith, Angela M. 2012. "Attributability, Answerability and Accountability. In Defence of a Unified Account." *Ethics 122*: 575–89.

Smith, Angela M. 2015. "Responsibility as Answerability." *Inquiry 58*: 99–126. https://doi.org/10.1080/0020174X.2015.986851.

Sullivan-Bissett, Ema. 2019. "Biased by our Imaginings." *Mind and Language 34*: 627–47. https://doi.org/10.1111/mila.12225.

Talbert, Matthew. Forthcoming. "Attributionism." In *The Oxford Handbook of Moral Responsibility*, edited by Dana K. Nelkin and Derek Pereboom. Oxford, New York: Oxford University Press.

Talbert, Matthew. 2016. *Moral Responsibility: An Introduction*. Cambridge, UK: Polity Press.

Toribio, Josefa. 2018. "Implicit Bias: From Social Structure to Representational Format." *Theoria. An International Journal for Theory, History and Foundations of Science 33*(1): 41–60. https://doi.org/10.1387/theoria.17751.

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics 24*(2): 227–48. https://doi.org/10.5840/philtopics199624222.

Wolf, Susan. 1990. *Freedom Within Reason*. New York: Oxford University Press.

Wu, Wayne. 2016. "Experts and Deviants: The Story of Agentive Control." *Philosophy and Phenomenological Research 93*(1): 101–26. https://doi.org/10.1111/phpr.12170.

Zheng, Robin. 2016. "Attributability, Accountability, and Implicit bias." In *Implicit Bias and Philosophy. Volume II: Moral Responsibility, Structural Injustice, and Ethics*, edited by Michael Brownstein and Jennifer Saul, 63–89. Oxford, New York: Oxford University Press.

AUTHOR BIOGRAPHY

Josefa Toribio is an ICREA Research Professor at the University of Barcelona. She previously held positions at the University of Sussex, Washington University in St. Louis, the University of Indiana, Bloomington, and the University of Edinburgh. Her current research focuses on the analysis of central topics in the philosophy of mind and the philosophy of cognitive science, with a special emphasis on the philosophy of perception and implicit attitudes.

How to cite this article: Toribio, J. (2021). Responsibility for implicitly biased behavior: A habit-based approach. *Journal of Social Philosophy*, 00, 1–16. https://doi.org/10.1111/josp.12442