Received: 4 July 2021 | Revised: 31 May 2022 | Accepted: 10 June 2022

DOI: 10.1111/ejop.12812

ORIGINAL ARTICLE

European Journal of Philosophy WILEY

# The importance of self-knowledge for free action

Joseph Gurrola

Department of Philosophy, University of Maryland, College Park, Maryland, USA

**Correspondence**
Joseph Gurrola, Department of Philosophy, University of Maryland, 4300 Chapel Drive, College Park, MD 20742, USA.
Email: gurrola@umd.edu

## Abstract

Much has been made about the ways that implicit biases and other apparently unreflective attitudes can affect our actions and judgments in ways that negatively affect our ability to do right. What has been discussed less is that these attitudes negatively affect our freedom. In this paper, I argue that implicit biases pose a problem for free will. My analysis focuses on the compatibilist notion of free will according to which acting freely consists in acting in accordance with our reflectively endorsed beliefs and desires. Though bias presents a problem for free action, I argue that there are steps agents can take to regain their freedom. One such strategy is for agents to cultivate better self-knowledge of the ways that their freedom depends on the relationship between their conscious and unconscious attitudes, and the way these work together to inform action and judgment. This knowledge can act as an important catalyst for agents to seek out and implement short- and long-term strategies for reducing the influence of bias, and I offer four proposals along these lines. The upshot is that though bias is a powerful influence on our actions, we need not resign ourselves to its negative effects for freedom.

## 1 | MOTIVATING THE PROBLEM

Philosophical work on implicit biases often centers on questions related to the pernicious influences that these biases can have on our moral behavior (Gendler, 2014; Holroyd & Puddifoot, 2021; Kelly & Roedder, 2008; Levy, 2017a). Less has been said about how implicit biases and their unreflective nature may make our actions less

free since they may cause us to act in ways we would not endorse for reasons we would not endorse.[1] One potential exception to this is the literature on implicit bias and moral responsibility (Brownstein, 2016b; Faucher, 2016; Glasgow, 2016; Levy, 2014, 2017b; Washington and Kelly, 2016; Zheng, 2016). However, while some views of free will require that free actions be those for which an agent is also morally responsible, these questions can be treated separately (Doris, 2015, 11) and I intend to treat them separately here.[2]

On the face of it, the observation that there are attitudes that operate below the level of conscious awareness that inform our actions seems to provide strong grounds for thinking that at least some of our actions that are informed by these unreflective attitudes are not free (Brandenburg, 2016). Here I will focus on a specific subclass of these actions that make the problem of implicit bias for free action especially salient: those for which we have reflectively endorsed a reason for action. In these kinds of cases, we have deliberated over what to do, and perceive ourselves as acting in accordance with our reflective commitments. These cases should be especially concerning since they are cases that, intuitively, best exemplify free action according to views that ground free action in the exercise of rational agency.[3]

Even if one believes that acting freely requires something more than acting in accordance with one's reflective values or concerns—for example, having had the ability to do otherwise—it is reasonable to assume that a person's acting freely involves acting in accordance with what they have decided is valuable (Frankfurt, 1971; Shoemaker, 2003). Importantly, this condition on free action is broader than the concern we have for behaving morally or behaving according to our moral values, since we often hold reflective values that are unrelated to our concern for being good persons. For example, an artist might have values related to her being a good artist, or an athlete may have values related to her desired accomplishments. Since our unreflective attitudes have the potential to affect actions related to our diverse set of reflectively endorsed values and concerns, they pose a formidable problem for free will. Here, I present evidence that suggests that actions that appear to be motivated by our reflectively endorsed values can be motivated by implicit attitudes that we would not endorse. I argue that this shows that reflective endorsement of reasons for action and the perception that our reflective reasons motivate our actions is often not sufficient for free action.

## 2 | FREEDOM OF THE WILL AS ACTION IN ACCORDANCE WITH REFLECTIVE ENDORSEMENTS

On one prominent compatibilist approach to free will, acting freely consists in acting in accordance with some subset of one's reflectively endorsed desires, beliefs, or cares (Frankfurt, 1969, 1971; Shoemaker, 2003). On these views, to be free is to have one's actions be motivated by the desires or goals that one reflectively endorses. These accounts are higher order accounts of free will since they hold that actions are free if they are caused by first-order desires that one desires to have (Frankfurt, 1969). For example, according to the Frankfurtian account of free will, an action is free if it is caused by the desires that one desires be efficacious in informing action.

One challenge posed by these higher order accounts of free action is that one can have multiple higher order volitions. For example, one might desire both that one's actions be informed by one's desire to be efficient and one's desire to patiently sort through evidence before deciding to act. In many cases, one will need to choose which of several first-order desires to act in accordance with if the first-order desires are inconsistent. One way to solve this problem is to claim that in these kinds of cases, free action requires that one make a decision that identifies oneself with one desire over another, thereby alienating oneself from the other relevant desires (Frankfurt, 1987). In these kinds of cases, one makes a choice not only about which higher order desire should guide one's actions in the moment, but also which higher order desire should take precedence in future decisions of a similar kind (Frankfurt, 1987, 175). Thus, though an agent may continue to hold the alienated desire, they have acted in such a way that ensures internal harmony "within the person," such that though the desires may be opposed in some objective sense, their presence within the agent does not result in a form of ambivalence about the desires for the agent

(ibid. 172). Frankfurt holds that so long as one acts in accordance with the desires one has reflectively chosen for oneself and endorsed as one's motivations, one is free, despite the presence of alienated attitudes.

A different solution to the problem of competing desires or cares in deliberation is to cultivate better self-knowledge. Shoemaker (2003) holds that if one spends time reflecting on one's cares, one will come to learn how much one values one's individual cares and which of one's cares are most important. This kind of self-knowledge should then make prima facie difficult decisions, like choosing between benefitting one's needy parents and one's children, less difficult, and will make one choose in a more wholehearted manner, perhaps without needing to reason about one's decision at all (Shoemaker, 2003, 110). While Shoemaker and I both advocate for a form of self-knowledge as a means to improving free action, we focus on different kinds of self-knowledge. Shoemaker focuses on knowledge of one's reflective desires, whereas I focus on knowledge of how one's capacities for reflective decision making comprise reflective and unreflective components. I mention this partly because it makes clear how our theses concerning self-knowledge and freedom differ, but also because one can accept either approach without accepting the other. For example, one might deny that there is always some fact of the matter about what we care about most, and thus be skeptical about the value of cultivating knowledge of one's reflective cares.[4] By contrast, one might accept this view but be skeptical that knowledge of the kind I endorse is useful given the recalcitrance of implicit attitudes to long term change (Lai et al., 2016).

According to the theory of free will I have summarized here, free action consists in acting in accordance with one's reflectively endorsed desires, specifically those desires about how one wishes to act. This theory of free will captures the intuition that one is unfree if one's actions are the result of desires or causes that one does not endorse—desires that are in a sense alien to oneself (Levy, 2014). In the next section, I summarize forms of implicit bias that I argue pose a serious problem for free action. The main problem I will discuss is that implicit biases can act as defeaters for attributions of free will for actions that seem from the perspective of the reflective agent to be free, but which are unfree given the unendorsed influence of implicit bias. I focus primarily on cases where one's implicit biases and reflective commitments fail to align. However, in Section 4.2, I also discuss what Holroyd (2016) terms "alignment" cases, ones where an agent's implicit attitudes align with at least some of their reflective commitments, though they may conflict with the reflectively endorsed goals and values that they desire be action-guiding.

## 3 | IMPLICIT ATTITUDES AND THEIR UNENDORSED INFLUENCE ON REFLECTIVE ACTIONS

In this section, I discuss two examples of implicit bias—stereotyping and in-group bias—that make it less likely that we will act in accordance with our endorsed desires in a given situation.[5] To be clear, these implicit biases can affect our seemingly free actions without our noticing; when implicit biases operate below the level of reflective awareness, we may falsely believe that the causes of our actions are our reflective cares, and thus believe that our actions are free. The point of this discussion is to show that in many cases where our reflectively endorsed actions are informed by our implicit biases, our actions are not free, since we do not reflectively endorse and in many cases would actively disavow the biases that best explain our actions. The upshot of this discussion, taken up in Section 4, is that a greater awareness of and intervention into the effects of these biases may afford us the opportunity to make our actions free.

### 3.1 | Implicit racism and sexism

Implicit bias is typically exemplified by the propensity of agents to exhibit racist or sexist behavior toward individuals of certain groups despite agents' explicitly avowed nondiscriminatory attitudes (Gendler, 2011; Levy, 2014). A paradigmatic case that illustrates the effects of implicit bias on behavior is one where an individual's actions are informed

by nonreflective attitudes with content that they actively disavow—for example, a person who exhibits a pattern of treating people who are White as if they are more intelligent, though they reflectively endorse statements like "there are no intrinsic differences in intelligence among races."

Here, it is important to note that the type of implicit bias at play in cases of implicit racism and sexism can be divided into two different kinds of attitudes or influences: implicit *stereotypes* about people belonging to a certain group, and implicit *affective attitudes* (Carruthers, 2018; Carruthers, 2015; Gilbert, Swencionis, & Amodio, 2012; Amodio and Devine 2006). Following Leslie (2017), I will assume that implicit stereotypes take the form of generics comprising semantic contents such as *Hispanics are fiery*, and *Asians are good at math* (Carruthers, 2018; Del Pinal & Spaulding, 2018; Leslie, 2017; Wodak and Leslie, 2017). Implicit affective attitudes, on the other hand, are evaluative attitudes that, depending on the context and an agent's activated attitudes, cause the agent to react to individuals in a positive or negative way based on the individuals' perceived social category (Brownstein, 2016a; Carruthers, 2018).

A competing view is that implicit biases have a unitary nature, underlain by mixed cognitive and affective associations (Gawronski & Bodenhausen, 2006; Gendler, 2008; Madva & Brownstein, 2018). On the associationist picture, there are not two kinds of bias or influences—cognitive and affective—but rather one type of attitude that comprises affective and cognitive associations (Gendler, 2008; Madva & Brownstein, 2018). While I assume a view according to which these influences can be distinguished as different types of attitudes, I do not pretend to offer any arguments in support of this here, though see Carruthers (2018) for a recent defense of the view grounded in what we know about cognitive and affective processing and their apparent reliance on different networks in the brain. I employ the distinction because I find it useful for discussing the experimental evidence on bias, which often isolates either affective or cognitive features of bias when explaining certain behaviors and judgments. I hasten to emphasize, however, that my analysis of the implications of implicit bias for freedom should stand on either picture of the nature of implicit attitudes. An analysis that assumed an associationist picture of implicit attitudes would merely require that I analyze the effects of stereotype and affective bias as effects of different features of the same kind of attitude, as opposed to the effects of two different kinds of attitudes.[6] In the next section, I discuss experimental evidence that is said to test for *associations* between kinds and concepts. I use the term 'association' mainly for ease of explication, as it follows the convention of the researchers conducting the experiments reviewed below, not to signal my commitment to the associationist view. People who take different views on the underlying nature of implicit attitudes can nevertheless agree on what the data show about the influence of these attitudes.

Though much of the literature surrounding implicit bias focuses on negative stereotypes and evaluative biases, implicit stereotypes and affective attitudes need not be negative in nature; for example, one may hold the implicit belief that *men are natural leaders*, which, depending on the context, might dispose them to assign tasks requiring greater leadership capacities to men—a positive outcome for that group (Carruthers, 2018; Del Pinal & Spaulding, 2018). Similarly, one may hold a positive implicit affective bias toward individuals from one's racial in-group (Spaulding, 2018). But regardless of whether the implicit stereotypes and affective biases are positive or negative, their influence on one's actions can constitute a defeater for free action if it is not an influence one does or would endorse.

It is generally accepted that the influence of implicit bias on our action and reasoning is closed off to reflective processes (Carruthers, 2018; Frankish, 2016; Gendler, 2014; Holroyd, 2016). Along these lines, Tamar Gendler argues that implicit bias often manifests without a "distinctive phenomenology" (Gendler, 2014, 197). According to this view, implicit stereotypes and affective attitudes do not generate internal conflict between an agent's reflectively endorsed beliefs or commitments and those attitudes operating in the background of her decision-making.[7] Consequently, implicitly biased agents who exhibit biased behavior are often unaware of the ways that their biases inform their actions. For example, agents on admissions committees who hold implicit stereotypes about the relative intelligence of individuals of different races or genders may be unaware of their disposition to view candidates of a certain race or gender as more qualified than those of a different race or gender (Gendler, 2014, 194; Schwitzgebel, 2011).

Sometimes, evidence of one's implicit affective attitudes and stereotypes may reach the level of conscious awareness. For example, in deciding to cross the street upon seeing a group of people belonging to a certain racial category approaching, an agent may acknowledge that her action is being driven by a feeling of fear for which she can provide no rational justification. Or consider an agent presented with evidence of inconsistency in the way she judges the reprehensibility of identical actions (say, protests for political causes) based on the political affiliations of the groups involved. In both cases, though the agent cannot be sure she possesses the implicit attitude she suspects she has, she can factor in the possible influence of an implicit bias and adjust her reflective reasoning accordingly.

In the following two sections, I will review the role that bias plays in informing our actions in a manner closed off to reflective reasoning.

## 3.2 | Implicit stereotyping as social cognition

Implicit stereotyping results from humans' reliance on reflexive categorization practices for efficiently navigating the world and is sustained by the perpetuation of stereotypes and structures that reinforce, and at times reward, stereotyping (Gendler, 2011, 43; Huebner, 2016; Leslie, 2017; Spaulding, 2021). Beginning at an early age, humans form categories into which they group objects and individuals according to shared properties (Gendler, 2011; Rhodes & Baron, 2019). The deployment of categories in perception allows individuals to make quick discriminations in an environment that presents them with "overwhelming detail" (Gendler, 2011, 39). Though these categorization processes are more likely to be deployed when agents are under significant cognitive load (Frankish, 2016; Levy, 2014; Spaulding, 2018), reflexive categorization is also "fundamental to how we make sense of the world." (Gendler, 2011, 39; Rhodes & Baron, 2019; Spaulding, 2018) Additionally, that we exist in societies in which race and gender are particularly salient social categories and in which racial and gender stereotypes are frequently circulated and reinforced makes it the case that the reflexive categorizations we make often reflect our knowledge of racial and gender stereotypes (Gendler, 2011; Huebner, 2016; Spaulding, 2018; Leslie, 2017). Agents do not have to think a stereotype is accurate or reflectively endorse the stereotype for them to be influenced by it. Rather, mere exposure to a stereotype is often sufficient for it to play a role in an agent's reflexive categorizations (Gendler, 2011; Leslie, 2017; Spaulding, 2017; Spaulding, 2018). Once an agent has learned a stereotype about a group, that stereotype then informs how the agent perceives individuals of that group, causing the agent to attend to those pieces of evidence that reinforce the stereotype and to discount those pieces of evidence that challenge it (Gendler, 2011; Gendler, 2014; Spaulding, 2018).

Implicit racial and gender stereotyping has been documented across various experimental paradigms. On one version of the Implicit Association Test (IAT), which claims to test for subjects' associations between male and female genders and certain professions, researchers found that subjects more quickly identified female elementary school teachers than male elementary school teachers, and more quickly identified male accountants than female accountants (White & White, 2006). This effect was exhibited even though more women than men hold accounting positions, and comparison of explicit associations over time showed that subjects in this study were less likely to explicitly stereotype accounting as a male profession than were subjects in similar studies conducted decades earlier (White & White, 2006). A different experimental paradigm that aimed at determining the influence of stereotypes related to competence in the sciences found that science faculty were more likely to hire male candidates than equally matched female candidates with identical CVs where the names had been changed (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). In the experiment, a large group of science faculty at various universities were provided with identical CVs with only the gender of the names changed and asked to evaluate the hireability and competence of the candidate as well as starting salary and potential for future mentoring opportunities. The study found that women candidates were rated as significantly less hireable and seen as less competent than equally qualified male candidates; they also received a significantly lower starting salary and fewer prospective mentoring opportunities. The authors of the study emphasize that the biased responses were due primarily to prevailing stereotypes about the competence of women in STEM fields, rather than positive or negative affective biases, as the

evaluators rated the female candidates as more likeable than the male candidates, and the gender of the evaluators did not make a difference in competence ratings (Moss-Racusin et al., 2012, 16,477).

Experiments on the effects of priming, which "measure the effects of subtle cues in the environment on [agents'] emotional and cognitive responses" also illustrate the problematic effects of implicit bias (Spaulding, 2018, 28). In one subliminal priming task that tested the effects of priming police officers with "words related to the social category Black," officers primed with these words were more likely to associate a "hypothetical youth…whose race is unspecified" with markers of "delinquency" and categorize them as "more likely to reoffend" (Spaulding, 2018, 28; Graham & Lowery, 2004). Gendler (2014) describes another priming experiment in which subjects were asked to "shoot" only those targets who were armed, after being asked about their knowledge of prevalent stereotypes of African Americans (Gendler, 2014, 207). In the experiment, subjects were "more likely to 'shoot' an armed target if the target [was] of African descent than if the target [was] White, and more likely to refrain from shooting an unarmed target if the target [was] White than if the target [was] of African descent." (Gendler, 2014, 207) Researchers noted that the biased behavior of participants in these experiments was best explained by participants' knowledge of stereotypes about Black men in American culture, such as the stereotypes that they are violent or aggressive (Correll, Park, Judd, & Wittenbrink, 2002). This conclusion was supported by the fact that the same level of racial bias was exhibited by White and African American participants in the study. And since African American participants reported knowledge of stereotypes about Black men in American society, but were unlikely to hold "strong prejudices against their own ethnic group," researchers took this as evidence that the presence of stereotypes (as opposed to the presence of an affective bias) best explained the biased results (Correll et al., 2002, 1,325).

These examples illustrate the ways that our implicit stereotypes, once activated by features of our environment, can lead us to act in ways that conflict with one of our most crucial desires: the desire to treat people fairly and equally, regardless of race, gender, or other features that may not be integral to who they are as individuals. If we are disposed to think of individuals from certain groups in terms of problematic stereotypes and treat them differently because of our implicit stereotypes,[8] then we have good reason to think that our actions toward them conflict with our reflective commitments regarding the fair treatment of others. This presents a challenge for freedom on any theory of free will that holds that free actions are those whose motivating reasons are reflectively endorsed. In cases where implicit stereotypes are implicated as causes of action, agents cannot reflectively endorse their influence due to their unreflective nature and may even actively disavow their contents.[9]

In cases where biases manifest in biased behaviors but not in an awareness of the bias's influence in deliberation, the conflict between our commitments and our implicit attitudes will not result in a "distinctive phenomenology—a sort of mental discord or conflict" (Gendler, 2014, 191). For example, an implicit stereotype may alter an agent's space of reasons by disposing them toward certain interpretations of the evidence over others without their being any the wiser. In these kinds of cases, an agent may believe they are acting in accordance with their reflective desires, and their evidence—including the evidence provided to them by the relative fluidity of their final decision—will suggest that this is the case (Proust, 2013). Their bias will have shaped their view of the evidence so that they are more responsive to evidence that fits the stereotype than they are toward evidence that conflicts with it, and this means that they are unlikely to notice any conflict between their avowed commitments and their actions (Gendler, 2011; Spaulding, 2018).

For example, if I am on a job search committee, I may view some candidates as more likable than others, but also as less competent overall if their race or gender is associated with stereotypes of incompetence, as in the case mentioned above (Gendler, 2014; Moss-Racusin et al., 2012). My bias in this instance will be opaque to me because I do not experience internal conflict about my bias, even if I experience conflict about which candidate to choose. My bias has made it the case that I will attend to evidence that reinforces stereotypes, rather than evidence that challenges stereotypes. But because I lack knowledge of the effects of bias on my reasons for choosing, I lack the motivation to reflect on the veridicality of the evidence and to question whether my reasons are well supported. My implicit bias has shaped my epistemic space—the space in which I reason or deliberate—in such a way that the reasons available to me for choosing one candidate over another seem authoritative and clear. I believe I am selecting a candidate not because I am biased, but because according to my evidence they are more qualified (Gendler, 2011).

What these kinds of cases show is that I can act in accordance with my reflectively endorsed desires while also being influenced by implicit stereotypes whose influence I would disavow.

On the account of free will in question, an action is free if it follows from reflectively endorsed desires, but given what I have shown above about the role implicit stereotypes play in shaping our evidence, it follows that an action can appear to meet the criteria of a free action for an agent while being unfree due to the biasing nature of implicit attitudes. It follows that in many cases, free action is indiscernible from the perspective of the agent from unfree action informed by bias. One might know one's cares well and be an especially careful reasoner and still fail to act freely, given the unreflective influence of implicit attitudes on reflective action.

### 3.3 | In-group and out-group bias as affective bias

Thus far, I have discussed implicit bias as a feature of implicit stereotyping. In this section, I discuss a form of implicit affective bias—in-group and out-group bias.

In-group bias, or in-group favoritism, is typically characterized as a bias resulting in agents exhibiting preferential treatment toward individuals they consider to be like them in certain respects and exhibiting less preferential treatment toward individuals they deem to be unlike them in certain respects (Ames, 2004; Rhodes & Baron, 2019; Spaulding, 2017).[10] This preferential treatment is often grounded in an affective bias, where one is disposed to make comparatively more positive evaluations of one's in-group members than one's out-group members (Baron & Dunham, 2015; Rudman, Greenwald, Mellott, & Schwartz, 1999).[11,12] While the effects of in-group bias may sometimes rise to the level of conscious awareness—as when an agent acknowledges in response to a philosophical thought experiment like the trolley problem that they care more about the well-being of their family member than the well-being of a colleague or a complete stranger—they also operate in the background of our decision-making processes. When in-group bias operates this way, it shapes the way that we attend to and interpret evidence in ways that cause us to view in-group members more favorably and out-group members less-favorably without our knowledge of the influence of this bias (Carruthers, 2018; Spaulding, 2017).

In-group bias is also a product of the more general phenomenon of reflexive social categorization, discussed above. Reflexive sorting practices are "automatic" in the sense that we do not control when they are deployed (Spaulding, 2017, 4,015). We take in vast amounts of information while navigating social interactions, and the deployment of social sorting practices allows individuals to navigate more efficiently by attending to situationally salient properties of individuals, such as race, gender, or occupation (Rhodes & Baron, 2019; Spaulding, 2018). But, just as the categorization processes that inform implicit bias avail themselves of problematic stereotypes, so too do the social sorting practices that result in in-group favoritism. When an agent categorizes an individual as part of her out-group, she is more likely to perceive that individual's actions and intentions in terms of stereotypes that attach to the individual's socially salient features than she is to attend to those features that challenge the relevant stereotypes (Ames, 2004; Spaulding, 2017). By contrast, when agents perceive individuals as part of their in-group, they attend more carefully to their specific actions and impute individual intentions and motivations to those individuals, rather than explaining and predicting their behavior solely in terms of stereotypes (Ames, 2004). Additionally, when agents sort individuals into their in-group, they are likely to judge and interpret those individuals' actions more charitably than they are to interpret the actions of individuals who they sort as part of their out-group (Spaulding, 2018). These judgments also lead to a disparity in the ways we empathize with individuals whom we reflexively judge to be "like us" or "unlike us" (Spaulding, 2018, 32; Bloom, 2016).

Like implicit stereotyping, the effects of in-group bias have important consequences for our ability to act in accordance with our reflective commitments related to the equal treatment of others. If I am unaware of the degree to which my reflexive social-sorting practices shape the ways that I attend to the evidence in my social interactions, or if I am unaware of the way that my environment will make certain implicit beliefs and affective biases more salient, then I will be unaware of how my deliberation is influenced by biases whose effects on the way I perceive

others I would find problematic (Brownstein, 2016a; Carruthers, 2018). I can act with an awareness of my reflective desires and commitments—and so believe I act freely—and, even so, display an unconscious bias whose influence I would, upon reflection, disavow as conflicting with my reflective commitments, and whose presence acts as a defeater for my claim to free action. In the next section, I will show how cultivating better self-knowledge of our unreflective attitudes can help facilitate free action.

## 4 | SELF-KNOWLEDGE AS A TOOL FOR FREE ACTION

In my discussion of the relationship between implicit bias and free action, I have argued that an action can meet all the subjective conditions for being free—that is, an action can seem to one as if it is motivated by one's reflective attitudes—while being compromised by implicit biases that one would disavow. One upshot of this is that actions which appear to be free from the perspective of the agent can also be actions that illustrate biased behaviors that the agent would disavow. Perhaps more consequentially, one can act in ways that form a chain or pattern of biased behavior that one would also disavow. In this section, I present a case in which an agent's pattern of action illustrates this point. I then argue that knowledge of our biases is a helpful tool for reflecting on what has gone wrong for freedom in such cases, and for working to ensure that future actions are not susceptible to the same defeaters for free action.

*The case*: Kate, an admissions committee member at a local university, takes pride in her job and always ensures that she carefully reflects on each admission decision she makes. Kate pays particular attention to the previous academic success of applicants, as well as to comments in students' letters of recommendation that indicate that the student has sufficient academic mettle to succeed at her university. She is sometimes deeply conflicted about which students to admit, but this is only because she cares deeply about ensuring that she admits those students who most deserve to be admitted. Kate is also a staunch supporter of causes for racial justice on social media and spends a significant amount of her time volunteering at a local women's mission in which many of the residents are women of color.

One day after the most recent admission cycle, Kate receives a report on patterns of bias in the admissions decisions at her university. The report identifies each admissions committee member by a confidential six-digit number. Next to the identifying numbers of committee members, the report indicates the level of bias exhibited by each member's admissions decisions on a scale of "benign" to "severe", as well as the most salient bias exhibited by their admissions decisions. Attached to the report is Kate's confidential six-digit identifier, 629701. She locates her number on the report and is unnerved to discover the following sequence of terms: "629701; Severe; Racial". Kate receives a second report shortly thereafter that provides further insight into her biased decisions. When selecting between equally matched applications of White and Black candidates, Kate admitted White applicants at a rate of 1.8 to 1 over Black candidates. Kate is almost twice as likely to admit White candidates over equally matched Black candidates.

What is the best way to analyze Kate's case? Kate's actions appear, at least partially, to be the product of her reflective commitments since they follow from deliberation and she has reflective awareness of the ways that her decisions are being guided by these commitments. But, importantly, Kate's actions also reflect a bias that she strongly disavows as conflicting with one of her most important reflective commitments. While we could decide to identify Kate with each of her decisions that formed the pattern of bias as well as with the reflective commitments that informed each of her decisions, neither of these identifications provide insight into the pattern of bias that has emerged. That is to say, it seems that we can analyze each of Kate's apparently free actions in terms of her reflective commitments at that moment and still fail to uncover any explanation for the pattern of implicit racial bias. Moreover, Kate has what seems to be a good, reflective awareness of her reasons for choosing one candidate over another, and our knowledge of Kate's political advocacy and dedication to fair admissions practices lends credence to the idea that Kate is operating without any explicitly biased attitudes which might affect her admissions decisions (Levy, 2017b).

It might be argued that Kate's biased actions were guided by emotional reactions that reflect evaluative biases toward certain individuals that spoke in favor of some students being admitted over others. That is, one might argue that insofar as Kate's actions are the result of emotional reactions linked to her racial biases, then we can reasonably understand and explain her implicitly biased actions in terms of reflective endorsements of certain evaluative judgments. This might leave open an explanation of Kate's actions in terms of *explicit* bias, since it assumes that the biasing attitudes were available to conscious reflection.

But this seems like an odd story to tell, especially considering the earlier description of implicit bias as something that lacks a distinctive phenomenology. Unlike explicitly biased attitudes that one recognizes as one's own, but which one disavows, implicitly biased attitudes fail to generate any such reflection. Often, implicitly biased attitudes and judgements seem completely justified on the surface, since the agent is oblivious to their connection to biased reasoning and categorization processes. The agent understands these attitudes in terms of the reflective reasons she has for holding these attitudes, not in terms of any background biases. If the influence of Kate's implicitly biased attitudes lacks a distinctive phenomenology, and their presence generates no internal conflict with her explicitly avowed anti-racist attitudes, then the effects of implicit bias on her admissions decisions seem poorly explained by talk of reflective commitments.

Moreover, the mere presence of an affective evaluation like a gut feeling one way or the other is not sufficient for attributing an explicit bias to an agent. This is because an agent may accept the evaluation without knowing (and therefore being unable to endorse) the basis of the evaluation. To explain Kate's actions in terms of an explicit bias, one would have to show not only that Kate was aware of and endorsed her emotional reaction but also that Kate was aware that the reaction was itself a manifestation of bias and that she endorsed the *bias*, not merely its emotional output.

Acknowledging the difficulty of linking the influence of Kate's implicit bias with her reflective commitments illustrates how a lack of self-knowledge about one's unreflective attitudes limits one's capacity to act in accordance with one's reflective commitments. Obtaining better self-knowledge of what she reflectively believes or desires will not help Kate mitigate the negative effects of her bias in her admissions decisions (though see Frankish, 2016 for an account of how one can use one's reflective attitudes to override the influence of one's unreflective attitudes). Kate knows she reflectively endorses the importance of fair admissions practices, and that she is deeply committed to causes of racial justice. These commitments inform her actions in the workplace and in her private life. But further reflection on these attitudes will be unlikely to provide her with the tools to mitigate her bias unless she also obtains knowledge about the ways that her judgments of others are informed by categorization processes that exploit problematic stereotypes and cause her to view certain individuals less charitably.

Kate's actions are not free since they are informed by stereotypes and affective attitudes that lead her to make biased decisions. While none of Kate's actions on their own may illustrate a conflict between her judgments that certain individuals should be admitted over others and her commitment to racial justice, a wider perspective on her actions illustrates the ways that her actions fail to align with one of her most strongly held commitments. This is where knowledge about ourselves as reflective agents subject to the effects of bias becomes especially important for free action. If by reflecting on the way she is structured as a reflective agent influenced by certain biases Kate can develop strategies to modulate the influence of her implicit bias, and if this results in a reduction in bias in her admissions decisions, then Kate's admissions decisions will be more free since on the one hand they follow from her reflective attitudes, and on the other hand they will be less susceptible to biases that are opposed to these attitudes.

Many of us are like Kate: we care a great deal about treating people fairly, as well as about causes of social justice like racial and gender equality. Given the effects of implicit bias and social sorting practices, we have good reason to think that many of our everyday actions that follow from our reflective judgments are actions that result in our treating people unfairly. This means that many of the actions that, at least from our first-person perspective as reflective agents, qualify as free are also actions that may not be free since they are informed by attitudes that conflict with our reflective commitments and desires. Knowledge of the ways that bias causes us to treat people unfairly affords us the opportunity to take steps so that we are less susceptible to its problematic influence. In Section 4.2, I will show how we can take steps to ensure that our actions reflect our commitment to treating people fairly,

despite the color of their skin, their gender, or whether they are different from us in other superficial ways. Through this process of mitigating the negative effects of bias on our actions, we can ensure that more of our actions are free.

### 4.1 | On harmony cases

So far, I have discussed the consequences for freedom in cases where the conflict between one's implicit biases and one's reflective commitments are relatively clear cut. But one might wonder what my account has to say about more complex cases where the conflict is not as clear, or perhaps where no apparent conflict exists at all. Can implicit attitudes still generate problems for freedom in these cases?

Holroyd (2016) distinguishes "alignment" or "harmony" cases—those in which the implicit attitudes that inform an agent's actions align with their reflective attitudes—from "conflict" cases in which the implicit attitudes that inform action fail to align with the agent's reflectively endorsed attitudes. Consider a variety of an alignment case—what Holroyd (2016) calls a "protocol-adhering" case—where an avowed racist's implicit racist attitudes inform action in unwanted ways, such that they lead to actions that conflict with an important goal the achievement of which requires non-prejudiced action.[13] Holroyd offers the example of a person, B, who is an avowed racist except in the professional setting where he cares deeply about hiring the most qualified candidates. B forms the commitment to act in a non-racist manner toward candidates of color so that he can achieve his goal of hiring the best candidate. Despite his commitment, and because of the influence of his implicit racist attitudes, he proceeds to behave in a prejudiced manner toward the best qualified candidate who happens to be a person of color, and this has a negative effect on the interview, leading B to discount the candidate's qualifications and preventing B from achieving his goal. What should we say about how B's implicit attitudes influence the freedom of his actions?

To treat alignment cases requires that we distinguish between the content of an attitude and the influence it has. An agent may be influenced by implicit attitudes whose contents align with attitudes whose influence the agent reflectively endorses. But an agent may also be influenced by implicit attitudes that align with the content of her reflectively endorsed attitudes, but which fail to align with her reflective desires about which attitudes should inform her actions. In the case of B, I submit that a problem for B's freedom is posed by a failure of alignment between the influence of his implicit attitudes and his desires about how attitudes with those contents should influence his actions in a professional setting; this is true despite the fact that the *contents* of his implicit and explicit attitudes appear to align.

It seems reasonable to think that, like B, agents can form reflective commitments to act in accordance with certain values over others and take steps to ensure that this is the case. When they believe they are acting in accordance with the values they have previously committed to acting on but instead act on other implicit attitudes without being any the wiser, their freedom is compromised. This is true even if the efficacious implicit attitudes are ones whose contents and influence they would endorse in contexts outside of the one in question.

Though implicit attitudes pose a problem for freedom in many cases, there are some cases where the influence of these attitudes does not negatively affect freedom. These are straightforward alignment cases, where an agent endorses the contents of the efficacious unreflective attitudes as well as their influence. Since both the contents and the influence of the implicit attitudes in question align with the agent's reflective commitments, there is no corresponding problem for free action. For example, an explicit racist whose implicit racist attitudes inform some of his racist actions in a manner he endorses does not qualify as unfree on the view advanced here, despite the influence of implicit attitudes. This conclusion follows from the above analysis of the problems that implicit attitudes pose for free action, which focused primarily on the conflict between the influence of these attitudes and our reflective commitments. Because there is no conflict between the agent's reflective attitudes and his efficacious unreflective attitudes in these cases, there is no problem for free action.[14]

This response to straightforward harmony cases dovetails nicely with intuitions about adjacent questions of how we should morally evaluate actions influenced by implicit attitudes. Though I will not argue for this view here, some may hold the intuition that the racist in the straightforward alignment case bears more responsibility for his racist actions than the racist who disavows the influence of his racist implicit attitudes, since the racist in the straightforward alignment case endorses his racist actions and desires to behave in a racist manner, while the protocol-adhering racist does not (Holroyd, 2016).[15] The present analysis suggests that there is an important difference in whether the respective agents' actions are free: the action of the agent in the straightforward alignment case is free, while the action of the agent in the protocol-adhering case is unfree since the racist implicit attitudes that inform his actions fail to align with his reflective desires about which attitudes should inform his actions. This difference in freedom may help explain the intuition that the agent in the straightforward alignment case is more responsible than the agent who desires to act in a non-racist manner but fails because of the influence of racist implicit attitudes.

The importance of considering alignment cases is that they illustrate how acting freely in light of the influence of implicit attitudes is even more complicated than it might have seemed. While conflict cases like the Kate case illustrate the most obvious problem implicit attitudes present for freedom—that the contents of our implicit attitudes conflict with our reflective ones—Holroyd's discussion of alignment cases suggests that sometimes even implicit attitudes whose contents align with our reflective attitudes can lead us to act in ways that we would deem unfree if they influence our actions in ways we would not endorse. Reflection on alignment cases also suggests that the mere presence or influence of implicit attitudes is not grounds for claiming a lack of freedom. The extent of the alignment matters. If an agent endorses the implicit attitudes and their influence on her actions—that is, if she acts based on the influence of attitudes whose influence she endorses or would endorse—then there are few reasons related to the presence of implicit attitudes to think she is unfree.

## 4.2 | The value of self-knowledge for free action

I have claimed that self-knowledge is helpful for free action since it provides us with knowledge we need to take steps to modulate our actions so that they better align with our reflective commitments, but I have said very little about how exactly cultivating self-knowledge might help us to do this.[16] It seems odd to think that mere knowledge of the organization of our reasoning processes as subject to the effects of bias suffices to ensure freer action. Additionally, research on the recalcitrance of implicit attitudes in the face of interventions to change them seems to speak against the value of self-knowledge (Lai et al., 2016). If agents cannot do anything about the existence of their implicit biases, how does knowing about them help agents act more freely?

The value of self-knowledge is grounded in the fact that agents will not attempt to take steps to mitigate the effects of bias unless they have knowledge of their biases. And even if the steps that agents can take to change their biases are limited, they are not completely powerless in organizing themselves or their environments in ways that can mitigate the effects of their biases on their actions and decisions (Holroyd & Kelly, 2016; Madva, 2017). In short, what power an agent has over the influence of their biases is unlikely to be exercised unless the agent knows that implicit bias presents a problem for free action in the first place, and this knowledge is just the kind of knowledge that I have argued we ought to cultivate if we want to act more freely.

But mere knowledge of the problem is unlikely to mitigate its effects. For knowledge to be helpful, agents must be motivated to change and have some knowledge of strategies that can reduce the effects of implicit bias on their reflective actions. The former condition seems easily met, since it is reasonable to think we are usually motivated to change circumstances we see as impinging on our most important values and cares. The second condition is admittedly more difficult to satisfy, as agents may have to acquaint themselves with the social and empirical literature on interventions to curb the effects of bias in order to get a reasonably good idea about how they might modulate the influence of bias and which interventions are unlikely to work.[17] For example, an individual will have to learn that

attempting to suppress the influence of stereotypes in outgroup interactions is a bad strategy that can lead to disastrous "rebound" effects (Macrae, Bodenhausen, Milne, & Wheeler, 1996). Still, given the motivational force to which concerns about freedom and morality are apt to give rise, we should think that individuals with the resources to do so might undertake the project of learning how they can reduce the influence of bias.[18]

The remaining question is whether there are interventions that can reduce the influence of bias so that one can act more freely, making good on the promise that self-knowledge can lead to freer action. While some philosophers are skeptical of the steps agents can take to reduce biased action and decision making (Brandenburg, 2016), leading them to conclude that freedom with respect to one's implicit attitudes is largely limited by social circumstances, others are more sanguine about the prospects for exercising control over biases in decision and action (Holroyd & Kelly, 2016). The approach I advocate for here involves the exercise of what Holroyd and Kelly (2016), following Clark (2007), term ecological control. Ecological control involves deploying strategies that rely on a "robust, reliable source of relevant order in the body…brain, and/or in one's local environment," rather than relying primarily on one's in-the-moment reflective capacities for guiding action to achieve one's ends (Clark, 2007, 4). Below, I discuss three examples of ecological control aimed at mitigating the effects of bias by shaping one's environment to reduce the likelihood that bias will be deployed in the first place.[19]

Research suggests that biases are more likely to be deployed and are more difficult to correct for when agents are under significant cognitive load (e.g., when they are stressed, hungry, or fatigued) or are under time pressure (Burgess, Beach, & Saha, 2017; Gilbert, Pelham, & Krull, 1988; Macrae, Hewstone, & Griffiths, 1993). This suggests that one way for agents to reduce biased action is to reduce the amount of stress or cognitive load they experience in contexts where biases are apt to be activated and deployed. For example, in professional settings where they are tasked with evaluating candidates for hire or admission, agents can ensure that this process is completed during times of the day where they are relatively relaxed and well rested, such as after their afternoon lunch break, rather than just before. By contrast, it would be unwise to schedule this process for the end of the workday or after a department meeting, when one is likely to be most tired or in need of a break. This kind of intervention is relatively local and low cost and is likely to have benefits in other settings as well; for example, it can also reduce factors that might lead to biased grading or unfair evaluation of employees. So long as an agent has the flexibility to schedule their workday in this manner, they stand to benefit from the reduction of cognitive load in their evaluative judgments of candidates.[20]

The second intervention aims at reducing the availability of information that is apt to activate bias. To reduce the likelihood that they will form evaluations based on bias, agents can anonymize the material they review about candidates where this is feasible (Holroyd & Kelly, 2016). This will be especially beneficial for agents who need to evaluate candidates for hire or admission in a manner that eschews considerations about superficial characteristics such as group membership (Goldin & Rouse, 2000). The process of anonymizing can also be implemented by instructors, editors, or those in positions to evaluate the performance of candidates in a professional setting. In these contexts, individuals can set up processes to anonymize submissions of work and the files they use to track overall performance. Since stereotypes and affective biases are often triggered by information indicating group membership, we should expect that reducing the amount of identifying information available to evaluators in many contexts should result in fewer biased judgments.

Anonymizing candidate criteria stands to benefit the freedom of agents in more ways than one. If representation of traditionally underrepresented groups is increased as a result of anonymized evaluation, this will increase the number of counter-stereotypic exemplars in a field. We should expect that greater exposure to counter-stereotypic exemplars will reduce implicit bias in those acquainted with the exemplars and that as the number of counter-stereotypic exemplars in a field grows, this effect will be more widespread (Dasgupta & Greenwald, 2001; Saul, 2013). Second, the presence of counter-stereotypic exemplars helps to reduce stereotype threat, which should in turn decrease the rate at which underrepresented individuals leave a field based on impressions that they do not belong (Saul, 2013; Steele, 2011). Reducing stereotype threat in underrepresented individuals therefore stands to further increase the representation of counter-stereotypic exemplars, and the increase in exposure to counter-

stereotypic exemplars in a field may in turn lead to further reductions in biased action and judgment directed toward members of the underrepresented group.[21]

A final example of exercising ecological control also trades on the value of exposure to counter-stereotypic exemplars (Holroyd & Kelly, 2016). As mentioned above, research suggests that priming individuals with information about counter-stereotypic exemplars from a group can moderate their implicit biases toward individuals of that group (Dasgupta & Greenwald, 2001; Dasgupta & Rivera, 2008). In situations where individuals know that they have a bias against a certain group, they can prime themselves with photos of counter-stereotypical exemplars within contexts where they form judgments about members of that group. For example, the person on the admissions committee who knows that they hold an anti-Black bias can set the screensaver on their office computer to display images of prominent Black intellectuals and leaders—individuals whose images should prime positive information about Black individuals that can counter the person's anti-Black bias. The same method could be employed by individuals who have good reason to think that they hold implicit biases against other groups. Reviewing the files of applicants in a context that primes one to hold more positive evaluations of a certain group should reduce the degree to which bias is manifested in one's evaluations of members of that group.

Finally, agents can reduce instances of biased judgment by cultivating mindfulness through the process of mindfulness meditation. Recent research suggests that agents who practice mindfulness meditation can reduce their reliance on biased attitudes in reflection, resulting in fewer biased actions and judgments (Burgess et al., 2017; Gendler, 2014; Lueke & Gibson, 2015; Oyler, Price-Blackshear, Pratscher, & Bettencourt, 2021). Agents who practice mindfulness meditation develop a greater capacity to suppress the influence of reflexively activated attitudes, such as stereotypes. They also become better at identifying and accepting the presence of reflexive attitudes in a non-judgmental manner that allows them to treat them as objects for consideration rather than as authoritative reasons for judgment and action (Lueke & Gibson, 2015; Zhang et al., 2019). Agents who practice mindfulness meditation also exhibit a reduction in stress and an increase in their ability to regulate their mood, which is helpful for reducing reliance on implicit attitudes given that they are most often deployed under conditions of cognitive load (Burgess et al., 2017; Zhang et al., 2019). By practicing mindfulness meditation, agents can reduce the instances of reflexive activation of implicit attitudes, and when they are activated, agents will have better tools for recognizing and moderating their influence. In effect, cultivating mindfulness helps agents slow down their thinking, even in situations where reflexive judgments and evaluations are apt to be activated and would generally have a strong biasing effect on action and judgment.

In contrast to the three interventions I described above, which are meant to be relatively short-term and limited to specific contexts, mindfulness meditation stands to have wider effects for an agent's ability to reduce the influence of bias. This is explained by the fact that practicing mindfulness meditation cultivates a capacity that can be employed in various contexts. However, as with any other skill, developing mindfulness takes time and practice, and consequently requires a longer commitment from agents for its cultivation. But the benefits of increased mindfulness as a means of reducing biased action in the longer term seem to make this commitment worthwhile, especially considering that there are other personal benefits to cultivating mindfulness, such as an overall reduction in stress (Burgess et al., 2017).

In this section, I argued that we should think that self-knowledge is important for free action. The interventions I proposed are examples of how agents can use their self-knowledge to mitigate the effects of bias on their actions and decisions. Most of the interventions I have proposed are relatively local in their implementation; this is a benefit, since agents should have an easier time implementing these structures to reduce bias. Of course, the tradeoff is that local interventions may be less effective at reducing bias on a larger scale. But when we consider the positive long-term effects of some interventions like anonymizing admissions decisions, we can see that though an intervention may be limited to one context, its effects can still be widespread.

Importantly, my arguments leave room for social approaches to ameliorating the effects of implicit bias (Brandenburg, 2016; Huebner, 2016). Indeed, it may be the case that when it comes to reducing bias, approaches that prioritize structural and societal changes, both in terms of the information that is circulated and the physical

structure of our societies, are best (Huebner, 2016). To the extent this is true, agents should also work toward longer term societal and structural changes that are apt to reduce bias on a larger scale. (For example, in addition to implementing anonymous grading in their own courses, a person might advocate for across-the-board anonymous student grading at their university, or anonymized admissions procedures where this is feasible.) In addition to gaining benefits for their own freedom, working toward these kinds of projects can also benefit others' freedom, as well as their ability to do right.

But acknowledging that societal and structural change is the best approach for eliminating bias altogether is consistent with acknowledging the importance of interventions that moderate the effects of one's own biases. Just because bias is apt to exist so long as there are structures that maintain it does not mean that agents cannot or should not take steps to curb their own biased behavior, even if the effects of their interventions are often limited to certain contexts. As I argued above, we have strong reasons to intervene to reduce the influence of implicit bias on our actions. And there is little reason to think of social and individual approaches to reducing bias as opposed to each other in some zero-sum manner, where benefits to individual agency consume limited resources that could otherwise be used for societal change. Reducing bias in our actions can have downstream effects that contribute to a less biased environment, especially where we reduce bias in institutional settings like universities and the workplace.

## 5 | CONCLUSION

In this paper, I have shown how implicit bias poses a challenge to freedom under one prominent type of compatibilist free will. Though I have focused on implicit attitudes that lead to morally harmful consequences, my analysis should concern anyone who cares about their capacity to act in accordance with the attitudes and concerns they reflectively endorse. Mitigating the effects of bias on reflective action stands to have a significant impact on one's life since one can become more free than one otherwise would have been. Not addressing these attitudes means resigning oneself to a life where one would never be sure whether one's reflective actions were free or not, and where one would have good reasons to think that many of one's actions are not free.[22]

### ACKNOWLEDGMENT

Open access funding enabled and organized by Projekt DEAL.

### ORCID

*Joseph Gurrola* https://orcid.org/0000-0002-9680-9159

### ENDNOTES

[1] An exception is Brandenburg (2016) who also argues that implicit bias poses a problem for freedom. Though my and Brandenburg's conclusions about how implicit bias can affect freedom are similar, I focus specifically on the most troublesome cases, those in which a reason for action has been reflectively endorsed. We also differ in our views on the implications for freedom (e.g., whether freedom is limited) and in the solutions that we propose. My discussion in section 4 will briefly touch on this disagreement.

[2] One reason we should think that these questions can be treated separately is that we can imagine cases in which people act freely but in which, in virtue of their failing to meet some condition required for blameworthiness, we would think they lack the necessary conditions for being morally responsible with respect to some act or omission (Heyndels & De Mesel, 2018; Shoemaker, 2017). For example, one might argue that if I fail to notice an apparently injured person on my way to give a talk for which I am late, I may not be morally responsible for failing to help them in virtue of my not noticing them (Vargas, 2013). This can be true even if I navigate to the talk freely, understood as my acting in a manner that is an expression of my values.

[3] Since the view I discuss holds that implicit bias presents a challenge for freedom on accounts that ground freedom in the exercise of one's rational capacities, the problems for freedom may also extend to free action on accounts that ground freedom in the ability to act based on reasons, or on the presence of a reasons-responsive mechanism (Fischer &

Ravizza, 1998). This is because agents who act on the influence of implicit biases are not actually responsive to reasons related to reflective values, such as concerns for egalitarianism or good reasoning. Moreover, the effects of implicit attitudes are arguably widespread, undermining the view that in most cases our actions are not motivated by these irrational influences.

[4] Thanks to an anonymous referee for suggesting this concern.

[5] I use the term "bias" throughout to refer both to implicit stereotyping and implicit affective biases since both are likely to lead to behavior that illustrates a bias toward or against some group or other. I distinguish between these two types of implicit attitude below.

[6] Finally, one might wonder how the distinction between affective and cognitive attitudes fits within the ongoing debate about whether implicit biases are propositional in nature, and, whether they are like beliefs (Levy, 2014; Mandelbaum, 2016; Spaulding, 2021). According to the view that distinguishes between affective and cognitive biases, only cognitive attitudes are propositional in nature, while affective attitudes are not propositional. However, claiming that stereotypes are primarily cognitive in nature does not commit one to the claim that stereotypes are bona fide beliefs (Carruthers, 2018). Although Leslie, (2017) treats generics, the attitudes underlying stereotypes, as "essentialized beliefs", acceptance of this view is not required for grasping that stereotypes are a form of essentialized thinking (Spaulding, 2018; Spaulding, 2021). Moreover, my summary of the empirical research on bias and my analysis of its effects on reflection should be plausible on either the belief or non-belief view of stereotypes and biases more generally. This is important to recognize, as the status of biases as beliefs is as yet unsettled. For an insightful review of the state of the art of the debate see Spaulding (2021).

[7] By "internal conflict", I mean a conflict of commitments or beliefs of which the agent is conscious. One way to think of this is as a conflict that arises in an agent's space of reasons—that space in which an agent passively observes the competition for influence among commitments that could potentially inform her action (Shoemaker, 2003).

[8] It is worth noting that even where a stereotype is statistically correct, we may nevertheless want to disavow its influence. This might be the case if we hold the commitment to treat others as individuals, regardless of their membership in a particular group and any stereotypes that might be relevant. For example, if in expecting to meet my colleague's spouse whom I know is an elementary school teacher I assume the spouse will be a woman and am therefore surprised when I find out he is a man, I might chide myself for basing my expectations on a stereotype; I'd do this regardless of the truth of the stereotype that most elementary school teachers are women. Thanks to Peter Carruthers for suggesting this example.

[9] Complicating this concern is that most people recognize the content of the stereotypes that end up informing their actions. Indeed, an agent may reflectively disavow the content of a stereotype in making their decision—for example, a hiring manager deciding between a male and gender-non-conforming candidate may reflectively disavow the false stereotype that *men are natural leaders*—and nevertheless act in a manner shaped by this very stereotype (Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

[10] These categorizations need not correspond to salient categories in their respective societies (e.g., race, gender, age) and can be artificial, as shown by research on the minimal group effect (Bigler, Jones, & Lobliner, 1997; Dunham, Baron, & Carey, 2011; Spaulding, 2018).

[11] For a comprehensive discussion of the literature on social categorization and its biasing effects, see Spaulding, 2018, chapter 3.

[12] In-grouping and out-grouping also results in expectations for in-group and out-group members. For example, one expects people sorted into one's in-group and out-group to abide by the norms of their respective groups, and to engage in helping behavior toward members of their respective groups in cases of inter-group conflict (Pun, Birch, & Baron, 2021; Rhodes & Baron, 2019; Roberts, Gelman, & Arnold, 2017).

[13] Holroyd (2016) calls this a "protocol-adhering case" since the agent has the goal of adhering to a protocol about how to behave, such as the commitment to behave in a non-racist manner in a certain context, that conflicts with their otherwise reflectively endorsed implicit and explicit attitudes.

[14] Though straightforward alignment cases avoid the problem that implicit attitudes pose for freedom, this should not assuage our concern over their negative effects for freedom. There are many contexts, most notably social ones, where the actions of those committed to the fair treatment of others are apt to be unfree due to the influence of attitudes, such as false stereotypes or ingroup biases, that they do not endorse. And recall that even when the contents of our implicit attitudes align with our reflective ones, we may still fail to act freely if we do not endorse the influence of these attitudes.

[15] This is not to suggest that the protocol-adhering racist is free of responsibility due to the influence of bias. He may be morally responsible on the grounds that he cultivates explicitly racist attitudes that are apt to influence his behavior even when he does not want them to (Holroyd, 2016).

[16] Thanks to two anonymous referees for voicing this concern.

[17] Thanks to Cody Gomez for this suggestion.

[18] Admittedly, not all agents will be able to engage in this process in equal measure. Differences in leisure time and financial stability will all influence the amount of time agents will be able to devote to this project, and stressors may negatively affect an individual's motivation to learn and the amount they glean from the material they consume. Still, it is reasonable to think that some individuals will take steps to learn how to be more free and improve their ability to do right if these are things they care about, and that they will be motivated to get it right, even if this takes time.

[19] Two of the interventions I discuss here—anonymization of materials and exposure to counter-stereotypic exemplars in contexts where one makes evaluations—are suggested in Holroyd and Kelly (2016). I expand on the rationale for these interventions here.

[20] Agents can also take steps to implement this kind of scheduling on a structural level. For example, a manager might schedule the workday of their office so that their employees complete reviews of dossiers immediately following their lunch break rather than before. They might also incentivize this kind of process by providing a reward (e.g., light lunch) to anyone willing to schedule their time to achieve the same ends.

[21] Anonymization may be especially good as a process for reducing biased judgments since while its effects are apt to be widespread, the intervention itself can be relatively localized (for example, in the decisions of individual instructors or committees).

[22] I am grateful to Peter Carruthers, Eric Saidel, Vanessa Singh, Midnight Singh, and Tyler Loveless, as well as to two anonymous revewiers for their comments on earlier versions of this article. I am also grateful to Hallie Liberto, Dan Moller, and the University of Maryland Philosophy Work in Progress Workshop for helpful discussion about the ideas and arguments in this paper.

## REFERENCES

Ames, D. R. (2004). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology*, *87*(3), 340–353.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. *Journal of personality and social psychology*, *91*(4), 652.

Baron, A. S., & Dunham, Y. (2015). Representing 'us'' and 'them': Building blocks of intergroup cognition. *Journal of Cognition and Development*, *16*(5), 780–801.

Bigler, R. S., Jones, L. C., & Lobliner, D. B. (1997). Social categorization and the formation of intergroup attitudes in children. *Child Development*, *68*(3), 530–543.

Bloom, P. (2016). *Against empathy: The case for rational compassion* (First ed.). New York, NY: Ecco, an imprint of HarperCollins Publishers.

Brandenburg, D. (2016). Implicit attitudes and the social capacity for free will. *Philosophical Psychology*, *29*(8), 1215–1228.

Brownstein, M. (2016a). Context and the ethics of implicit bias. In *Implicit bias and philosophy* (Vol. 2). Oxford: Oxford University Press.

Brownstein, M. (2016b). Attributionism and moral responsibility for implicit bias. *Rev.Phil.Psych.*, *7*, 765–786. https://doi.org/10.1007/s13164-015-0287-7

Burgess, D. J., Beach, M. C., & Saha, S. (2017). Mindfulness practice: A promising approach to reducing the effects of clinician implicit bias on patients. *Patient Education and Counseling*, *100*(2), 372–376.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford, UK: Oxford University Press UK.

Carruthers, P. (2018). Implicit versus explicit attitudes: Differing manifestations of the same representational structures? *Review of Philosophy and Psychology*, *9*(1), 51–72.

Clark, A. (2007). Soft selves and ecological control. In Ross, D., Spurrett, D., Kincaid, H., & Stephens, G. L. (Eds.), *Distributed Cognition and the Will: Individual Volition and Social Context* (pp. 101–122). Cambridge, MA: MIT Press.

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

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

Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long–term contact and short–term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, *26*(1), 112–123.

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

Doris, J. M. (2015). *Talking to our selves: Reflection, ignorance, and agency* (First ed.). Oxford, United Kingdom: Oxford University Press.

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development*, *82*(3), 793–811. https://doi.org/10.1111/j.1467-8624.2011.01577.x

Faucher, L. (2016). Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics. In *Revisionism and moral responsibility for implicit attitudes*. Oxford, UK: Oxford University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge, UK: Cambridge University Press.

Frankfurt, H. (1987). Identification and wholeheartedness. In *The Importance of What We Care About*. Cambridge, UK: Cambridge University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*(23), 829.

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

Frankish, K. (2016). Playing double. In *Implicit bias and philosophy* (Vol. 1). Oxford: Oxford University Press.

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

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

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*(1), 33–63.

Gendler, T. S. (2014). I—The third horse: On unendorsed association and human behaviour. *Aristotelian Society*, *88*(1), 185–218.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia*, *50*(14), 3600–3611.

Glasgow, J. (2016). Alienation and responsibility. In *Implicit bias and philosophy* (Vol. 2). Oxford: Oxford University Press.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review*, *90*(4), 715–741.

Graham, S., & Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and Human Behavior*, *28*, 483–504.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Heyndels, S., & De Mesel, B. (2018). On Shoemaker's response-dependent theory of responsibility. *Dialectica*, *72*, 445–451. https://doi.org/10.1111/1746-8361.12243

Holroyd, J. D., & Kelly, D. (2016). Implicit bias, character and control. In: Masala, A., & Webber, J. (Eds.) *From Personality to Virtue Essays on the Philosophy of Character*. Oxford, UK: Oxford University Press.

Holroyd, J. (2016). VIII—What Do We Want from a Model of Implicit Cognition? In *Proceedings of the Aristotelian Society* (Vol. 116, pp. 153–179). Oxford, UK: Oxford University Press.

Holroyd, J., & Puddifoot, K. (2021). Implicit bias and epistemic oppression in confronting racism. *Journal of the American Philosophical Association*, 1–20. https://doi.org/10.1017/apa.2021.12

Huebner, B. (2016). Implicit bias, reinforcement learning, and Scaffolded moral cognition. In *Implicit bias and philosophy* (Vol. 1). Oxford: Oxford University Press.

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, *3*(3), 522–540.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016.

Leslie, S. J. (2017). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, *114*(8), 393–421.

Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. *Noûs*, *48*(1), 21–40.

Levy, N. (2017a). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research*, *94*(1), 3–26.

Levy, N. (2017b). Am I a racist? Implicit bias and the ascription of racism. *The Philosophical Quarterly*, *67*(268), 534–551.

Lueke, A., & Gibson, B. (2015). Mindfulness meditation reduces implicit age and race bias: The role of reduced automaticity of responding. *Social Psychological and Personality Science*, *6*(3), 284–291.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Wheeler, V. (1996). On resisting the temptation for simplification: Counterintentional effects of stereotype suppression on social memory. *Social Cognition*, *14*(1), 1–20.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, *23*(1), 77–87.

Madva, A. (2017). Biased against debiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. *Ergo: An Open Access Journal of Philosophy*, *4*, 145–179.

Madva, A., & Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs*, *52*(3), 611–644.

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.

Oyler, D. L., Price-Blackshear, M. A., Pratscher, S. D., & Bettencourt, B. A. (2021). Mindfulness and intergroup bias: A systematic review. *Group Processes & Intergroup Relations*, *25*, 1368430220978694–1368430220971138.

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford, UK: Oxford University Press.

Pun, A., Birch, S. A. J., & Baron, A. S. (2021). The power of allies: Infants' expectations of social obligations during intergroup conflict. *Cognition*, *211*, 104630.

Rhodes, M., & Baron, A. (2019). The development of social categorization. *Annual review of developmental psychology*, *1*, 359–386.

Roberts, S. O., Gelman, S. A., & Arnold, K. H. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science*, *41*, 576–600.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the implicit association test. *Social Cognition*, *17*(4), 437–465.

Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In: Hutchison, K., & Jenkins, F. (Eds.), *Women in Philosophy: What Needs to Change*? Oxford, UK: Oxford University Press.

Schwitzgebel, E. (2011). Self-ignorance. In *Consciousness and the self: New essays* (pp. 184–197). Cambridge, UK: Cambridge University Press.

Shoemaker, D. (2003). Caring, identification, and agency *. *Ethics*, *114*(1), 88–118.

Shoemaker, D. (2017). Response-dependent responsibility; or, a funny thing happened on the way to blame. *Philosophical Review*, *126*(4), 481–527.

Spaulding, S. (2017). Do you see what I see? How social differences influence mindreading. *Synthese*, *195*(9), 4009–4030.

Spaulding, S. (2018). *How we understand others: Philosophy and social cognition*. Abingdon, Oxford: Routledge, an imprint of the Taylor & Francis Group.

Spaulding, S. (2021). Beliefs and biases. *Synthese*, *199*(3), 7575–7594.

Steele, C. M. (2011). Whistling Vivaldi: How stereotypes affect us and what we can do. New York, NY: *WW Norton & Company*.

Vargas, M. (2013). Situationism and moral responsibility: Free will in fragments. In: Clark, A., Kiverstein, J., & Clark, A. (Eds), *Decomposing the Will*. Oxford, UK: Oxford University Press.

Washington, N., & Kelly, D. (2016). Who's Responsible for This?: Moral Responsibility, Externalism, and Knowledge about Implicit Bias. In *Implicit Bias and Philosophy*, *Volume 2: Moral Responsibility*, *Structural Injustice*, *and Ethics*. Oxford University Press.

White, M. J., & White, G. B. (2006). Implicit and explicit occupational gender stereotypes. *Sex Roles*, *55*(3), 259–266.

Wodak, D., & Leslie, S. J. (2017). The mark of the plural: Generic generalizations and race. In *The Routledge companion to philosophy of race* (pp. 277–289). Routledge.

Zhang, Q., Wang, Z., Wang, X., Liu, L., Zhang, J., & Zhou, R. (2019). The effects of different stages of mindfulness meditation training on emotion regulation. *Frontiers in Human Neuroscience*, *13*, 208.

Zheng, R. (2016). Attributability, accountability, and implicit bias. In *Implicit Bias and Philosophy*, *Volume 2: Moral Responsibility*, *Structural Injustice*, *and Ethics*. Oxford, UK: Oxford University Press.

**How to cite this article:** Gurrola, J. (2022). The importance of self-knowledge for free action. *European Journal of Philosophy*, 1–18. https://doi.org/10.1111/ejop.12812