PSYCHOLOGICAL INQUIRY
2022, VOL. 33, NO. 3, 219–231
https://doi.org/10.1080/1047840X.2022.2115729

Routledge
Taylor & Francis Group

REPLY

Reflections on the Difference Between Implicit Bias and Bias on Implicit Measures

Bertram Gawronski[a], Alison Ledgerwood[b], and Paul W. Eastwick[b]

[a]University of Texas at Austin, Austin, Texas; [b]University of California, Davis, California

We are pleased about the considerable interest in our target article and that there is overwhelming agreement with our central thesis that, if the term *implicit* is understood as *unconscious* in reference to bias, implicit bias (IB) should not be equated with bias on implicit measures (BIM) (Cesario, this issue; Corneille & Béna, this issue; Cyrus-Lai et al., this issue; De Houwer & Boddez, this issue; Dovidio & Kunst, this issue; Melnikoff & Kurdi, this issue; Norman & Chen, this issue; Olson & Gill, this issue; Schmader et al., this issue; but see Krajbich, this issue; Ratliff & Smith, this issue). We are also grateful for the insightful commentaries, which continue to advance the field's thinking on this topic. The comments inspired us to think further about the relation between IB and BIM as well as the implications of a clear distinction between the two. In the current reply, we build on these comments, respond to some critical questions, and clarify some arguments that were insufficiently clear in our target article. Before doing so, we would like to express our appreciation for the extreme thoughtfulness of the commentaries, every single one of which deserves their own detailed response. For the purpose of this reply, we will focus on recurring themes and individual points that we deem most important for moving forward.

We start our reply with basic questions about the concept of *bias*, including the difference between behavioral effects and explanatory mental constructs, the role of social context, goals, and values in evaluating instances of bias, and issues pertaining to the role of social category cues in biased behavior. Expanding on the analysis of the bias construct, the next sections address questions related to the implicitness of bias, including the presumed unconsciousness of BIM, methodological difficulties of studying unconscious effects, and the implications of a broader interpretation of *implicit* as *automatic*. The next sections again build on the discussions in the preceding sections, addressing questions about the presumed significance of IB research for understanding societal disparities and the value of BIM research if IB is treated as distinct from BIM. The final section presents our general conclusions from the conversation about our target article and several suggestions on how to move forward.

Reflections on Bias

Bias as a Behavioral Phenomenon

Our analysis of IB is based on a behavioral definition of bias as the effect of social category cues (e.g., cues used to construct racial and gender categories) on behavioral responses. This definition is based on the notion that bias should be conceived of as a behavioral phenomenon that needs to be explained rather than a "thing" that people have that would explain their biased behavior (see De Houwer, 2019; Payne & Correll, 2020). As we noted in our target article, explanations of the latter type can easily become circular when (1) biased behavior is explained by the proposition that people have bias and (2) the bias people are presumed to have is inferred from the biased behavior that needs to be explained (see Cervone et al., 2001; Fleeson & Jayawickreme, 2021; Gawronski & Bodenhausen, 2015). A behavioral definition of bias avoids such explanatory circularity by clearly distinguishing between bias as a behavioral phenomenon that needs to be explained and the mental processes and representations proposed to explain biased behavior.

Although some commentators explicitly supported our behavioral conceptualization of bias (Corneille & Béna, this issue; De Houwer & Boddez, this issue; Ratliff & Smith, this issue), others expressed concerns that a purely behavioral definition could miss important aspects of bias. Dovidio and Kunst (this issue) discussed the importance of attitudes, ambivalence, and intrapersonal responses for understanding bias; Olson and Gill (this issue) highlighted the role of motivation and opportunity to control automatically activated attitudes in the expression of bias; and Schmader et al. (this issue) pointed to the significance of beliefs, attitudes, stereotypes, motivations, and regulatory processes. We fully agree that all of these mental constructs are important for understanding bias, as well as the development of effective interventions to reduce bias (see Schmader et al., in press). However, the obvious value of the proposed mental constructs in *explaining* bias does not imply that they should be used to *define* bias. In fact, doing so would undermine their explanatory role, because it would create a purely semantic link between biased behavior as the to-be-explained phenomenon and the mental constructs proposed to explain biased behavior, which leads to circular explanations and logical fallacies in the understanding of the to-be-explained

CONTACT Bertram Gawronski gawronski@utexas.edu University of Texas at Austin, Austin, TX, USA.

phenomenon (De Houwer et al., 2013; Gawronski & Bodenhausen, 2015).

Our quest not to refer to mental constructs in a definition of bias as a behavioral phenomenon echoes earlier concerns by attitude researchers to clearly "distinguish between the inner tendency that is attitude and the evaluative responses that express attitudes" (Eagly & Chaiken, 2007, p. 582). Equating mental attitudes with their behavioral expressions would be unproblematic if there was a one-to-one relation between the two such that differences in mental attitudes generally involve corresponding differences in evaluative responses, and vice versa (see De Houwer et al., 2013). However, behavioral influences of attitudes are often disrupted by motivational processes, and these processes can shape evaluative responses over and above mental attitudes (see Olson & Gill, this issue; Schmader et al., in press). Applied to the current question, these issues prohibit direct equations of biased behavior with biased attitudes, because two individuals may have the same biased attitude but differ in the degree to which they show bias in their behavior (e.g., when one of them is motivated to suppress the expression of their biased attitudes and the other is not; see Schmader et al., in press). Conversely, two individuals may differ in terms of their biased attitudes but nevertheless show the same degree of biased behavior (e.g., when someone who is motivated to conceal their biased attitudes behaves in the same way as someone without biased attitudes; see Schmader et al., in press). These concerns apply not only to self-reports and blatant expressions of biased behavior; they are also highly relevant for responses on implicit measures, which are known to be shaped by multiple distinct processes, only some of which are related to underlying attitudes (see Calanchini et al., 2014; Conrey et al., 2005). Thus, similar to the concern that a direct equation of mental attitudes and behavioral evaluations undermines our understanding of when and how attitudes guide behavior (Eagly & Chaiken, 2007), including mental constructs (e.g., attitudes) in a definition of bias can undermine our understanding of the complex processes underlying biased behavior. A purely behavioral definition of bias, such as the one proposed in our target article, avoids these problems by clearly distinguishing between bias as a behavioral phenomenon that needs to be explained (*explanandum*) and the mental processes and representations proposed to explain biased behavior (*explanans*).

The significance of distinguishing between biased behavior and underlying mental processes can be illustrated with findings cited by Dovidio and Kunst (this issue), suggesting that members of disadvantaged groups who show anti-ingroup BIM are at greater risk for mental health problems. Dovidio and Kunst (this issue) argue that these relations are the product of intrapersonal processes in people's minds, which might be missed when bias is defined at a purely behavioral level. Although we agree that intrapersonal processes are essential for understanding the link between anti-ingroup BIM and mental health, we would argue that (1) a purely behavioral definition of bias facilitates a more nuanced understanding of the processes underlying this link and (2) a reference to mental constructs in the definition of bias is detrimental rather than helpful in this endeavor. From the perspective of a purely behavioral definition, anti-ingroup BIM represents negative evaluative responses to one's ingroup on an implicit measure. Such responses should not be treated as a direct indicator of anti-ingroup attitudes, because they are jointly shaped by (1) negative thoughts about one's ingroup and (2) the effectiveness of inhibitory processes in suppressing the behavioral expression of these thoughts (see Conrey et al., 2005). Moreover, recent research suggests that, while individual differences in inhibitory control on implicit measures are relatively stable over time, the activation of unwanted thoughts is highly variable (Elder et al., 2022). Thus, to the extent that mental health problems more likely arise from stable than unstable factors, ineffective inhibition of negative thoughts about one's ingroup (and the systemic factors that support or undermine inhibitory control) might play a more significant role for the observed link between anti-ingroup BIM and mental health problems than the unwanted thoughts per se.[1] This important nuance is missed when bias is defined in mental terms, for example when anti-ingroup BIM is equated with anti-ingroup attitudes. A purely behavioral definition of bias avoids these issues, allowing for a more nuanced analysis of the link between anti-ingroup BIM and mental health problems.

Evaluating Instances of Bias

Another concern about our definition of bias is that it is too broad in the sense that it subsumes effects that we may not want to call bias. Dovidio and Kunst (this issue) argued that effects of social category cues on behavioral responses should be called bias only when they are unjust or unfair; Schmader et al. (this issue) suggested that the consequences for the target are essential for classifying behavior as biased; and Norman and Chen (this issue) pointed to cases where the absence of differential treatment rather than its presence may be deemed bias (e.g., failing to tailor one's directions to accommodate a person's ability to use the stairs or an elevator). These concerns seem especially important in response to claims of "reverse bias" against members of dominant groups (e.g., Cesario, this issue; Cyrus-Lai et al., this issue).

We fully agree that social context is fundamentally important for evaluating instances of biased behavior and appreciate the commentaries that pushed for a deeper consideration of this point. Discussions of bias cannot and should not be divorced from the historical conditions and societal hierarchies within which those biases operate (Salter et al., 2018; Sidanius et al., 2004). At the same time, we think it is useful to distinguish between (1) effects of social

[1]This conclusion should not be misinterpreted to suggest that members of disadvantaged groups just have to work harder to suppress unwanted negative thoughts about their ingroups. It simply means that factors determining the effectiveness of inhibitory processes have to be considered for understanding the link between anti-ingroup BIM and mental health problems, and these factors can be outside of a person's control (e.g., impaired inhibitory control due to thoughts about financial problems; see Mani et al., 2013).

category cues on behavioral responses and (2) the (un)desirability of such effects (see Corneille & Béna, this issue). This distinction is important, because whether an effect of social category cues on behavioral responses produces a desirable or undesirable outcome depends on the specifics of history and context as well as one's goals and values.

For example, to build on Norman and Chen's (this issue) insightful scenario, many would agree that it is desirable to take category cues into account when deciding how to give directions to someone who is walking versus in a wheelchair. In contrast, many would agree that it is undesirable to take those same category cues into account when deciding how much of a raise to give someone based on their stellar work record. Likewise, judgments about desirability will depend on one's values and goals. For example, consider a woman who calls the police on families barbecuing in the park, but she does that only when the family members have dark brown skin but not when they have light beige skin. Because such differential treatment reproduces existing social hierarchies, it is likely to be perceived as acceptable by someone who wants to maintain or enhance these hierarchies, but as morally wrong by someone who wants to reduce them. Goals and values of this kind are relevant not only for moral evaluations of bias by non-academics, but also for evaluations by social scientists.

If researchers decide to define something as bias only when it produces an undesirable outcome, which effects of social category cues count as *bias* will depend on context, goals, and values, as well as the interplay between them. What counts as bias for one researcher may be completely different from what counts as bias for another researcher. Therefore, we think it is useful to distinguish between the definition of bias as a behavioral phenomenon and the question of whether it is (un)desirable, while underscoring the importance of explicitly discussing both. To be clear, this means acknowledging that a researcher's personal values and assumptions are not and cannot be left at the laboratory door (see Ledgerwood et al., in press; Reddy & Amer, 2022). For example, like many of our commentators, we believe that instances of bias deserve moral condemnation when they uphold asymmetric power structures and histories of oppression, whereas instances of bias that reduce historical inequalities may be morally desirable (e.g., affirmative action programs). By explicitly acknowledging the possibility that effects of social category cues on behavioral responses can be morally desirable, we can also acknowledge the moral need for what some commentators called "reverse bias" to compensate for a history of oppression and unfair treatment. Yet, any such judgments are extrinsic to our definition of bias as the effect of social category cues on behavioral responses. We treat them as moral judgments about bias rather than judgments referring to intrinsic features of bias. Likewise, our definition of bias allows researchers to ask important questions about whether the antecedents and consequences of bias are different depending on, for example, whether a given instance of bias upholds versus challenges societal inequalities. For example, certain goals or ideologies might lead to reduced biases overall, whereas other goals or ideologies might push people toward hierarchy-challenging biases and away from hierarchy-enhancing biases, or vice versa (see Hudson et al., 2019; Jones, 1998).

Such a definition of bias is also consistent with the use of the term *bias* in the broader literature on judgment and decision-making, where biases are treated as judgmental tendencies that can lead to inaccurate and maladaptive judgments in some contexts and to accurate and adaptive judgments in other contexts (Kruglanski & Ajzen, 1983). At the same time, we recognize that it may conflict with a lay understanding of bias as bad—something that should always be reduced—and we recognize the importance and challenge of using language that will clearly communicate ideas not only to scientists but also the public. Still, we think accepting a lay definition of bias as bad will likely create confusion—for example, leading people to assume that colorblindness and treating everyone exactly the same is necessarily morally desirable, a problem underscored by Norman and Chen's (this issue) extremely insightful analysis (see also Fryberg & Stephens, 2010; Jones, 1998; Yi et al., in press).

The Role of Social Category Cues

To avoid the issues addressed in the preceding section, Corneille and Béna (this issue) suggested a radical departure from extant terminology: instead of using the morally laden term *bias*, researchers should describe their findings as effects of social categorization. We fully agree that avoiding the term *bias* could be helpful to avoid potential misunderstandings. However, the proposed emphasis on social categorization conflicts with our goal to clearly distinguish between behavioral effects and explanatory mental constructs. Whereas effects of social category cues on behavioral responses are purely behavioral phenomena, social categorization is a mental process that may explain effects of social category cues, but this process should not be equated with the to-be-explained phenomenon (De Houwer et al., 2013, Gawronski & Bodenhausen, 2015). A clear distinction between effects of social category cues and social categorization seems especially important in light of findings suggesting that social category cues can sometimes influence responses independent of how a target is categorized (e.g., Blair et al., 2002, 2004; Livingston & Brewer, 2002). Such effects are captured by a behavioral conceptualization like the one we proposed in our target article, but they cannot be captured by a mental conceptualization in terms of social categorization. Likewise, a focus on social category cues rather than social categorization aligns well with calls for psychologists to move away from relying on social categories as explanatory constructs and toward examining how people use specific features to assign status in a dynamic and context-dependent way (Cikara et al., in press; Helms et al., 2005).

A closely related concern by Corneille and Béna (this issue) is that social categories are defined at the perceiver level and that, therefore, our definition of bias is not purely behavioral. We would argue that, although this concern applies to an alternative conceptualization in terms of social

categorization, it does not apply to our original conceptualization. There is a clear difference between social category cues at the stimulus level (e.g., skin color) and perceived category membership at the mental level (e.g., categorization of a person with lighter vs. darker skin color as White vs. Black). Our definition of bias refers specifically to social category cues at the stimulus level. As such, it is purely behavioral in the sense that it refers exclusively to aspects of stimuli and behavioral responses without invoking explanatory mental constructs (see De Houwer et al., 2013, Gawronski & Bodenhausen, 2015).

An important question raised by Norman and Chen (this issue) is whether our definition of bias captures cases involving category ambiguity. We appreciate their suggestion to explicitly discuss category ambiguity, which we think connects well with our definition of bias. In our view, category ambiguity often arises from the presence of mixed category cues, with some cues suggesting one category and others suggesting a different category. Such cases still involve effects of category cues, although the overall set of category cues is inconsistent rather than consistent. Such a conceptualization also implies the possibility that inconsistency itself may influence responses, potentially producing unique effects that cannot be understood as the additive product of individual cues. For example, a person's behavior toward a gender-ambiguous target may be distinct from the mere average of that person's responses toward an individual with unambiguous male features and an individual with unambiguous female features (Stern, 2022). In technical terms, these considerations suggest that, when studying effects of social category cues on behavioral responses, researchers should investigate not only main effects of individual category cues but also their interactions.

In addition to category ambiguity arising from inconsistent configurations of category cues, another possibility involves cases where category cues are weakly pronounced or absent. Norman and Chen (this issue) correctly note that such cases do not align well with the emphasis on effects of social category cues in our definition of bias. However, upon further reflecting on their thought-provoking argument, we think our definition can cover such cases, albeit in a more indirect way that may not seem obvious from the emphasis on social category cues. To identify effects of absent category cues on behavioral responses, one would need to show that absence of category cues elicits behavioral responses that are different from the ones when category cues are present. Moreover, to confirm that observed differences in responses are indeed driven by the absence of category cues in the "cues-absent" condition rather than the specific category cues in the "cues-present" condition, one would have to demonstrate that the observed differences generalize to a broad range of specific category cues in the "cues-present" condition. Thus, hypotheses about the effects of absent category cues necessarily involve comparisons to counterfactual cases involving present category cues, the latter of which is central to our definition of bias. Thus, although effects of absent category cues are not directly covered by our definition of bias, their significance is captured indirectly by the need to compare cases with and without category cues. Because some people show aversive reactions to category ambiguity associated with either inconsistent or absent category cues (Stern, 2022), we deem it important to acknowledge the potentially unique properties of category ambiguity and their relation to our definition of bias.

Reflections on the Implicitness of Bias

Is Bias on Implicit Measures Unconscious?

We are pleased that the authors of 9 out of the 11 commentaries agree with our conclusion that IB should not be equated with BIM if the term *implicit* in IB is understood as *unconscious* (Cesario, this issue; Corneille & Béna, this issue; Cyrus-Lai et al., this issue; De Houwer & Boddez, this issue; Dovidio & Kunst, this issue; Melnikoff & Kurdi, this issue; Norman & Chen, this issue; Olson & Gill, this issue; Schmader et al., this issue). However, because most of what we said in our target article would be obsolete if BIM were unconscious, the validity of our conclusion should not be determined solely on consensus. Rather, it seems essential to seriously engage with any counterarguments that may question our conclusion (Krajbich, this issue; Ratliff & Smith, this issue), even if these opposing views are not shared by the majority of our commentators.

One argument, put forward by Krajbich (this issue), is that the available evidence suggesting awareness of BIM is ambiguous, because the prediction tasks employed to measure awareness of BIM (e.g., Hahn et al., 2014) may inadvertently raise participants' awareness of their own biases. A related concern raised by Ratliff and Smith (this issue) is that, while the available evidence clearly speaks against complete unawareness of BIM, it does not rule out the possibility that people are unaware of their BIM when they do not pay attention to their biases.

We agree with the basic idea underlying these arguments. However, we would argue that it stretches the meaning of *unconscious* to a level that undermines a thorough understanding of unconscious processes. Although cognitive scientists have been unable to come up with a consensually accepted nominal definition of *unconscious* (Norman, 2010), a widely accepted operational criterion for determining (un)awareness of mental representations is whether people are able to verbally report them (Timmermans & Cleeremans, 2015). If we interpret the term *unconscious* in a manner to subsume any mental representation that, although verbally reportable, is not activated every second of the day 24/7, the distinction between *conscious* and *unconscious* would become semantically equivalent to the distinction between *activated* and *dormant* representations (Gawronski et al., 2006). In that case, we would have to call a person's liking for their best friend unconscious whenever the person is not actively thinking about it. We do not think such an expansive interpretation of *unconscious* is helpful for understanding the operation of unconscious representations (i.e., mental representations that people are unable to verbally report but nevertheless influence their behavior).

Another counterargument pertains to our thesis that surprise reactions in response to IAT feedback may be driven by a mismatch between the naïve metric used by participants to describe the extremity of their biases and the metric used by researchers to convert numeric IAT scores into verbal feedback (e.g., *strong preference for White people compared to Black people*). To the extent that the two metrics do not align, participants may be surprised about their IAT feedback, not because they are unaware of their bias, but because their personal description does not match the description in the feedback they receive (see Gawronski, 2019). Ratliff and Smith (this issue) were not convinced by this argument, citing the following five reasons:

> First, participants in these studies self-reported their preferences on the exact scale on which they received feedback; thus, the format was not entirely novel. Second, participants in these studies are defensive even when they receive feedback indicating only a slight implicit preference. Third, we have manipulated the format in which we give feedback and are unable to attenuate the basic defensiveness effect. Fourth, a re-analysis of the data from Howell et al. (2015) shows that the discrepancy between IAT feedback and self-report predicts defensiveness even among participants who report having previously taken an IAT (and are thus familiar with the format by which participants receive feedback). Finally, although we recognize that our anecdotal experience will not be recognized by everyone as a legitimate source of evidence, we note that together we have spoken to tens of thousands of people at more than 60 organizations about the fact that behavior can be influenced by social group cues in ways that are often unrecognized in the moment. And many people—people who are not in psychology research study pools or well-versed in behavioral science—are truly, genuinely surprised. (pp. 215–216)

In response to Ratliff and Smith's rebuttal, we would like to point out that their first, third, and fourth points misconstrue our original argument, which is about the metrics used to link performance levels to verbal labels, not the wording itself. To the extent that the metric used by participants does not align with the metric used by the experimenters, there would be a mismatch between participants' self-assessment and the experimenter's feedback, which is sufficient to cause a surprise reaction.

Regarding Ratliff and Smith's (this issue) second point (see also Goedderz & Hahn, 2022), it is worth noting that, according to our misaligned-metrics interpretation, more extreme feedback should lead to greater surprise only if there is a multiplicative relation between participants' naive metric and the metric used by researchers. However, feedback extremity should have no effect on surprise reactions if there is an additive relation between the two metrics. To illustrate this point, imagine two participants, one of whom perceives themselves to have a small bias of 1 based on their naïve self-assessment while the other perceives themselves to have a large bias of 3.[2] Now, assume a multiplicative "distortion" of this self-assessment by a factor of 2 in the researcher's feedback, which would suggest bias feedback of 2 for the first participant and bias feedback of 6 for the second participant. In this case, the second participant should be much more surprised, because the discrepancy between their self-assessment and the feedback is larger (i.e., 3) compared to the first participant (i.e., 1). However, that is not the case for an additive "distortion" where the discrepancy is exactly the same for the two participants. For example, if one assumes an additive "distortion" of 2, the bias feedback would be 3 for the first participant and 5 for the second participant, implying that the discrepancy between participants' self-assessments and experimenter feedback is exactly the same for the two participants (i.e., 2). This scenario illustrates that, if there is an additive relation between participants' naive metric and the metric used by researchers to label different levels of IAT performance, misaligned metrics should not necessarily lead to greater surprise as a function of feedback extremity. Hence, counter to Ratliff and Smith's (this issue) argument (see also Goedderz & Hahn, 2022), the fact that even feedback suggesting a slight degree of BIM can cause defensive (or surprise) reactions does not provide evidence for the idea that BIM is unconscious.

Regarding the fifth point in Ratliff and Smith's (this issue) rebuttal, we wonder if the anecdotal surprise reactions have anything to do at all with unawareness of bias, but instead reflect surprise about how one's conscious thoughts and feelings can influence performance in the IAT. Over the past years, the first author has used a classroom exercise, in which students collectively complete a flower-insects IAT with timed stimulus presentations on a classroom screen. Participants' task is to clap their legs with their left or right hand, with the required responses matching the ones in the so-called "compatible" and "incompatible" blocks of the standard IAT. Students are generally surprised about how difficult it is to quickly and accurately respond in the "incompatible" block of the task, even without receiving verbal feedback about their individual performance. Does this mean that the students are unaware of their preference for flowers over insects? We do not think so. It seems much more likely that they are surprised about how their conscious preference makes it so difficult to respond in the task. Although this observation is—like Ratliff and Smith's observation—merely anecdotal, it makes us even more skeptical about whether surprise reactions about IAT performance tell us anything about unawareness.

Another counterpoint put forward by Ratliff and Smith (this issue) is that, although participants may be aware of the effects of social category cues on some trials of an implicit measure, they may be not aware of such effects on all trials. Similarly, it seems possible that, although some participants may be aware of the effects of social category cues on their responses on an implicit measure, this may not be the case for all participants. We appreciate this point and agree that it is most likely true, but we would argue that it does not permit an equation of BIM and IB, if we define IB as an unconscious effect of social category cues on behavioral responses. To illustrate our concern, imagine a study in which all participants were aware of the effects of social category cues on their responses for 50% of the trials of an implicit measure and unaware for the other 50%.

[2]The numbers in this example are meant to reflect hypothetical levels of bias, not numeric IAT scores.

Correspondingly, imagine a study in which 50% of the participants were aware of the effects of social category cues on all of their responses on an implicit measure and 50% were unaware for all of their responses. Would it make sense to call the implicit measure in these studies a measure of unconscious effects of social category cues? We do not think such a classification makes sense, because the same logic could be used to call it a measure of conscious effects of social category cues. It would certainly be justified to call the measure in the two studies a measure of bias without further qualification. However, it would be arbitrary to call it a measure of unconscious bias, just as it would arbitrary be to call it a measure of conscious bias.

The Difficulty of Studying Unconscious Effects

Several commentaries noted the difficulty of studying unconscious effects of social category cues (Corneille & Béna, this issue; Cyrus-Lai et al., this issue; Krajbich, this issue; Ratliff & Smith, this issue; Schmader et al., this issue). We fully agree with this assessment. Although carefully controlled lab experiments are a valuable tool to determine the extent to which behavioral responses are influenced by social category cues, determining the unconscious nature of such effects is an extremely challenging task (see Timmermans & Cleeremans, 2015).

As some commentators pointed out, the difficulty of studying unconscious effects is partly rooted in the fact that every effect involves multiple different aspects that people may be aware or unaware of (Ratliff & Smith, this issue; Schmader et al., this issue). Ratliff and Smith (this issue) specifically noted that people may be (un)aware of (1) the response-eliciting stimulus, (2) their response to the stimulus, or (3) the causal link between the stimulus and their response (see also Gawronski & Bodenhausen, 2012). Applied to our definition of bias, these aspects correspond to (1) social category cues, (2) one's behavior, and (3) the causal link between the two. Although we agree that it can be interesting to study effects of stimuli that are presented outside of awareness (e.g., effects of subliminally presented stimuli) or effects on behaviors that people may not be aware of (e.g., effects on eye blinking rates), the qualifier *implicit* in our definition of IB was meant to refer specifically to the third aspect. We deliberately formulated our definition of IB as ***unconscious effects*** *of social category cues on behavioral responses*; we did not define IB as *effects of* ***unconscious social category cues*** *on behavioral responses* or *effects of social category cues on* ***unconscious behavioral responses***. The reason for our emphasis on effects was that, in most real-world settings, people are aware of social category cues (i.e., subliminal exposure to social category cues seems extremely unusual) and people are most often aware of what they are doing (e.g., they are aware that they are hiring a job candidate or that they are calling the police), but they may not be aware of that their actions are influenced by social category cues. For example, people may be perfectly aware that a job candidate has a prototypically female name and that they are making a hiring decision, but they may be unaware that their hiring decision is influenced by gender cues. Similarly, people may be perfectly aware that a person waiting inside a Starbucks has dark brown skin and that they are calling the police on that person, but they may be unaware that their decision to call the police is influenced by the person's skin color. These examples belong to a broader category of unconscious effects where people are aware of specific stimulus properties as well as their behavioral responses, but they may be unaware of how their behavior is influenced by those stimulus properties (see Ledgerwood et al., 2018).

However, even with a high level of conceptual precision about the intended referent of the qualifier *implicit*, empirically establishing unawareness of a causal effect is an extremely difficult endeavor (see Timmermans & Cleeremans, 2015). We fully agree with Corneille and Béna (this issue) that claims about unconscious effects of social category cues generally require thorough awareness checks. If no evidence for unawareness can be provided, researchers should abstain from making claims about unawareness, or at least clarify the speculative nature of their claims. We also agree with Cyrus-Lai et al. (this issue) that research on unconscious effects of social category cues should move beyond approaches in which unawareness is inferred from null effects. What is needed are approaches that establish unawareness from statistically significant effects rather than non-significant effects (although Bayesian statistics might be helpful for interpretations of null effects). Cyrus-Lai et al. (this issue) present some valuable suggestions in this regard, including experimental manipulations to increase the salience of potential effects of social category cues and tests of interaction effects between a manipulation of social category cues and measures of awareness.

Some bias researchers may not be interested in embracing the challenges of studying unconscious effects of social category cues. That is perfectly legitimate. However, in such cases, it would seem appropriate to limit conclusions to bias and refrain from making claims about unconsciousness. Indeed, an argument could be made that the dominant concern with IB has distracted researchers from studying blatant forms of bias, which still represent a major factor underlying the perpetuation of social disparities (see Corneille & Béna, this issue). Regardless of whether one agrees or disagrees with this view, not everyone may be interested in whether effects of social category cues are conscious or unconscious—some researchers may just be interested in bias without further qualification. Yet, if researchers are interested in studying IB, they should provide empirical evidence for their claims about unawareness; if they are not interested in accepting this methodological challenge, it would seem appropriate to refrain from making claims about unawareness.

What About IB as Automatic Bias?

Several commentators suggested that, instead of using the term *implicit* as synonymous with *unconscious*, it might be better to use it in a manner that is synonymous with the

broad umbrella term *automatic* (De Houwer & Boddez, this issue; Olson & Gill, this issue; Ratliff & Smith, this issue), focusing specifically on the unintentionality feature of automaticity (De Houwer & Boddez, this issue; Dovidio & Kunst, this issue; Krajbich, this issue; Olson & Gill, this issue; Ratliff & Smith, this issue). Indeed, a case could be made that an emphasis on unawareness could be detrimental, in that describing IB as unconscious could inadvertently lead to a rejection of responsibility for one's actions (Melnikoff & Kurdi, this issue; Ratliff & Smith, this issue; see also Daumeyer et al., 2019; Redford & Ratliff, 2016) and raising awareness in IB interventions could have other unintended effects (Corneille & Béna, this issue). To provide a context for our reply to these points, we deem it helpful to first explain why our target article focused on unawareness as the central characteristic of IB, before we move on to discuss the difference between unconscious and unintentional bias and its implication for the difference between IB and BIM. To foreshadow our conclusion: we agree with Corneille and Béna (this issue) that it might be time to jettison the term *implicit* as a qualifier for bias, and instead ask researchers to use the more specific terms *unconscious* (when they mean unconscious) and *unintentional* (when they mean unintentional). As we explain in this section, there are reasons to believe that both unconscious biases and unintentional biases are important for understanding social disparities. However, their specific roles are fundamentally different, echoing our argument in the target article that unconscious bias should not be equated with unintentional bias.

Two Schools of Thought

From the very beginning, research using implicit measures was shaped by two distinct schools of thought (see Gawronski, De Houwer, & Sherman, 2020; Payne & Gawronski, 2010). One school of thought is associated with the development of the evaluative priming task (EPT) to measure the automatic activation of attitudes (Fazio et al., 1986), which provided the basis for using the EPT as an unobtrusive measure of attitudes (Fazio et al., 1995). Central to the development of the EPT was the idea that attitudes, conceptualized as object-evaluation associations of varying strength, are activated unintentionally upon encountering a target object if the association between the object and its summary evaluation is sufficiently strong (see Olson & Gill, this issue). The second school of thought is associated with the development of the IAT (Greenwald et al., 1998), which was inspired by research on implicit memory suggesting that people can have memory traces they are unable to verbally report but nevertheless influence behavior. This idea is prominently reflected in Greenwald and Banaji's (1995) definition of *implicit cognition* as "introspectively unidentified (or inaccurately identified) trace of past experience that mediates [responses]" (p. 5).

A notable difference between the two schools of thought is that they emphasize different features of automaticity in their characterizations of implicit measures. Whereas the first school of thought emphasizes *unintentionality* as the central feature that distinguishes implicit from explicit measures, the second school of thought emphasizes *unawareness* of the underlying memory traces. The concept of IB was an intellectual product of the second school of thought, whose proponents suggested that people can behave in a biased manner without being aware that their behavior is biased (e.g., Banaji & Greenwald, 2013; Greenwald & Krieger, 2006). Notably, advocates of the first school of thought have repeatedly expressed concerns against using the term *implicit* as qualifier of measured constructs (e.g., bias), suggesting that it should instead be used to describe features of measures (e.g., Fazio & Olson, 2003). Responses on implicit measures were assumed to reflect the unintentional activation of attitudes in memory, not unawareness of the measured attitude (see Olson & Gill, this issue). For the sake of brevity, we will refer to the first school of thought as *unintentionality school* and the second school of thought as *unconsciousness school*.

Back to Implicit Bias

Although proponents of the unconsciousness school have recently backtracked from their early claims about unawareness of the constructs captured by implicit measures (e.g., Greenwald & Banaji, 2017),[3] the original conceptualization of IB as unconscious and its equation with BIM is still widespread in both the scientific literature and the broader discourse of this work. Our target article was inspired by two concerns about this state of affairs. First, in line with the concerns expressed by proponents of the unintentionality school, we aimed to clarify that there is no basis for the idea that BIM is unconscious. Second, reviving some aspects of the ideas advanced by proponents of the unconsciousness school, we aimed to make a case that this does not implicate a rejection of IB as the unconscious effect of social category cues on behavioral responses. Our broader point underlying these concerns is that the common equation of BIM and IB was detrimental to progress in understanding IB, because it led researchers to use BIM as an indicator of IB instead of directly studying IB.

What does this mean for the proposal to use the term *implicit* in a manner that is synonymous with the term *automatic* (De Houwer & Boddez, this issue; Olson & Gill, this issue; Ratliff & Smith, this issue)? As we explained in our target article, we do not think such a reinterpretation is helpful in advancing the science of IB, because the term *automatic* subsumes multiple distinct features (i.e., unintentionality, unawareness, efficiency, uncontrollability). Because these features do not overlap (see Bargh, 1994), the broad umbrella term *automatic* creates conceptual ambiguity if researchers do not specify to which of these features they are referring (see Corneille & Béna, this issue; Melnikoff & Kurdi, this issue). As noted by Corneille and Béna (this

[3]Different from their early claims about unconsciousness, proponents of the unconsciousness school now state that the term *implicit* should be used in the sense of *indirectly measured* (e.g., Greenwald & Banaji, 2017). We refer to the discussion in our target article for conceptual problems with this conceptualization.

issue), scientific progress is achieved by greater conceptual precision, not greater conceptual ambiguity. Several commentators acknowledged this issue, suggesting that work in this area should focus specifically on unintentionality (Dovidio & Kunst, this issue; Krajbich, this issue; Melnikoff & Kurdi, this issue; Ratliff & Smith, this issue). If IB were reinterpreted as unintentional effect of social category cues on behavioral responses, the equation of IB and BIM would be justified, because there is little doubt that implicit measures capture unintentional responses. However, as we explained in our target article, such a reinterpretation of *implicit* perpetuates the current sphere of inattention for unconscious effects of social category cues. Because unintentional is not the same as unconscious, shifting the focus from unconscious bias to unintentional bias continues to miss a potentially important factor in the perpetuation of social disparities.

Unconscious and Unintentional Bias

The significance of the difference between unconscious and unintentional bias can be illustrated with a central question in research on racial bias in police officers' decision to shoot, reflected in a tendency to more frequently shoot at unarmed Black targets compared to unarmed White targets (for a review, see Payne & Correll, 2020). One potential interpretation of this difference is that it reflects an unintentional effect of social category cues on response selection, involving an impulsive tendency to pull the trigger in response to Black but not White targets, which could be suppressed given sufficient time and mental resources. An alternative interpretation is that it reflects an unconscious effect of social category cues on basic perceptual processes, involving a tendency to mistakenly perceive harmless objects as weapons when they are held by a Black person but not when they are held by a White person. An important difference between the two accounts pertains to the correction of erroneous responses when participants have an opportunity to reflect on an initial speeded response without being able to see the target person and the relevant object (Payne et al. 2005). According to the unintentionality account, participants should correct their initial errors when they are given an opportunity to reflect on their initial responses even when they are unable to see the target person and the relevant object during the reflection period. In contrast, the unconsciousness account suggests that initial errors should remain uncorrected when participants are given an opportunity to reflect on their initial response but are unable to see the target person and the relevant object.

Payne et al. (2005) tested these competing predictions using a variant of the weapon identification task (WIT, Payne, 2001). The WIT is based on the notion of sequential priming, in that participants are briefly presented with a Black or White face prime, followed by a brief presentation of a gun or a harmless object as the target. The target object is replaced by a masking stimulus and participants are asked to indicate whether the target object was a gun or a harmless object. A common finding in the WIT is that participants misidentify harmless objects more frequently as guns when they were primed with a Black face than when they were primed with a White face (for a review, see Payne & Correll, 2020). Integrating an opportunity for reflection and error correction in the WIT, Payne et al. (2005) found that participants almost always corrected their initial errors, suggesting that racial bias in weapon identification is driven by unintentional effects of social category cues, not unconscious effects. These results suggest that unintentionality may indeed be more important for understanding social disparities, at least for racial disparities in police officers' use of lethal force.

But that is not the whole story. In a study that combined Payne et al.'s (2005) correction paradigm with Correll et al.'s (2002) first-person shooter task, Correll et al. (2015) investigated whether Payne et al.'s (2005) finding replicates for simulated shooting decisions (rather than classifications of target objects) and more complex visual stimuli involving full-body presentations of Black and White individuals holding either a gun or a harmless object in the context of real-world backgrounds. The results were remarkably different. Although participants corrected initial errors on a small number of trials, a strong racial bias continued to emerge under correction conditions. Analyses using Drift Diffusion Modeling (see Ratcliff et al., 2016) further showed a significant effect of race on the start point parameter reflecting "initial assumptions," but not the drift rate parameter reflecting "evidence accumulation" (see also Krajbich, this issue). Together, these findings suggest that, although social category cues can influence decisions to shoot in an unintentional manner, unconscious effects on basic perceptual processes play a major role in tasks that more closely resemble real-world settings (see Payne & Correll, 2020). Thus, exclusively focusing on unintentional effects and ignoring the possibility of unconscious effects involves a risk of missing important factors contributing to social disparities. More seriously, if BIM is treated as a model for unintentional bias in real-world settings (see De Houwer & Boddez, this issue), research using implicit measures may suggest misleading (and potentially inaccurate) conclusions due to the low resemblance of their task structure with real-world decision contexts (e.g., the false conclusion from Payne et al.'s, 2005, study that unconscious effects on basic perceptual processes do not matter for racial bias in decisions to shoot).

Based on these differences and the heavy focus on unintentionality in the commentaries to our target article, it might be helpful to follow Corneille and Béna's (this issue) suggestion to jettison the term *implicit* as a qualifier of *bias*, and instead use the term *unconscious* when one means unconscious and the term *unintentional* when one means unintentional (see also Corneille & Hütter, 2020). Using these more specific terms, the main argument of our target article translates into the proposition that, although implicit measures may be well suited to capture unintentional bias, they are not suitable to measure unconscious bias, the latter of which may contribute social disparities in a manner that is fundamentally different from unintentional bias.

Reflections on Implications

Understanding Social Disparities

In our target article, we discussed two potential mechanisms underlying unconscious biases: (1) biased interpretation of ambiguous information and (2) biased weighting of mixed information. Different from the conceptually distal links between real-world behavior and unintentional bias on implicit measures, the contexts in which the proposed underpinnings of unconscious bias tend to operate have clear counterparts in real-world settings (e.g., hiring and promotion decisions, jury selection, criminal sentencing, policing; see Gawronski, Ledgerwood, & Eastwick, 2020). However, one commentator expressed skepticism about the idea that findings from experimental lab research—which subsumes most of the research on biased interpretation and biased weighting—could be used to understand social disparities in real-world settings (Cesario, this issue). We would argue that, although this skepticism was expressed under the disguise of scientific rigor, its tacit underlying principles seem rather unreasonable once they are spelled out (see also Ledgerwood et al., 2022; Mora et al., 2022; Okonofua, 2022; Payne & Banaji, 2022). If findings from experimental lab work could not be applied to real-world contexts that do not permit experimental manipulation, we would not be able to use findings on the laws of gravitation in experimental physics to understand the movement of planets in the orbit (see Payne & Banaji, 2022). We do not think this is a reasonable stance to evaluate applications of basic science. Yet, if the skepticism is exclusively directed against research on social biases, questions could be asked about the underlying motivations for selectively applying ostensible principles of scientific rigor to one specific area of research but not to others (see Ditto & Lopez, 1992; Lord et al., 1979).

Related to this issue, some commentators expressed concerns that the societal significance of research in this area has been overstated, given the weak empirical basis for the strong claims that have been made by some researchers (Corneille & & Béna, this issue; see also Cesario, this issue). We agree with this concern, but with an important qualification. Based on our assessment of more than a quarter century of research using implicit measures (see Gawronski, De Houwer, & Sherman, 2020), we concur that the contribution of implicit measures to understanding social disparities seems disappointingly small, especially if one considers the enormous amount of research that has be done in this area. We attribute this state of affairs to the dominant, yet empirically questionable, narrative according to which responses on implicit measures provide uncontaminated indicators of trait-like unconscious representations that coexist with functionally independent conscious representations. Although this narrative has been challenged by multiple scholars from the very beginning (for a review, see Gawronski et al., 2022), their concerns had little impact on the dominance of this narrative in bias research using implicit measures. Thus, despite more 25 years of research using implicit measures, the contribution of unintentional biases to social disparities is still unclear.

A similar conclusion can be reached for unconscious biases, albeit for very different reasons. Because the dominant focus on BIM has created the mistaken impression that we were already studying unconscious effects of social category cues, we still know very little about unconscious bias—different from the massive number of studies using implicit measures to investigate unintentional bias. Thus, given the undisputable experimental documentation of social biases in the real world (e.g., Bertrand & Mullainathan, 2004; Bordieri et al., 1997; Moss-Racusin et al. 2012), the role of ignorance in maintaining social hierarchies (Mueller, 2020; Salter et al., 2018), and the intuitively plausible significance of unconscious bias for the perpetuation of social disparities, we would encourage a shift in the current research agenda from the currently dominant focus on unintentional biases captured by implicit measures to the still poorly understood phenomenon of unconscious bias.

What Is the Value of Implicit Measures?

Our rejection of BIM as an indicator of unconscious biases raises the question of whether implicit measures still have any value for research on social biases. Some commentators seemed rather skeptical about that, noting that the research program on BIM has lost considerable momentum over the last years—partly due to unresolved debates about the predictive validity of BIM and meta-analytic evidence questioning the presumed causal role of BIM in discriminatory behavior (Cyrus-Lai et al., this issue). Other commentators seem more optimistic, noting a potential role of implicit measures as a model of real-world bias under suboptimal processing conditions (De Houwer & Boddez, this issue). Yet, such a role requires that the processes and processing conditions that shape responses on implicit measures correspond to the processes and processing conditions of to-be-modeled real-world behavior (Gawronski, De Houwer, & Sherman, 2020; Gawronski & De Houwer, 2014). If either their underlying processes or their processing conditions do not align, using implicit measures as a model for bias in real-world settings can suggest misleading conclusions, as we illustrated with the example of unconscious versus unintentional racial bias in decisions to shoot. Moreover, although we agree that BIM may serve as a model for understanding unintentional bias (De Houwer & Boddez, this issue; Dovidio & Kunst, this issue; Krajbich, this issue; Melnikoff & Kurdi, this issue; Olson & Gill, this issue; Ratliff & Smith, this issue), we want to reiterate that unintentional is not the same as unconscious, and that there is no conceptual and empirical basis to interpret BIM as unconscious.

A more optimistic view was expressed by Olson and Gill (this issue), who argued that unintentionally activated attitudes may influence basic perceptual processes in a manner that can lead to unconscious effects of social category cues. Similar to arguments we made in our target article, such mechanisms would suggest a potential role for BIM in understanding the mental underpinnings of unconscious biases, but this role does not permit a direct equation of

BIM with unconscious bias. Some commentators also noted the value of implicit measures to prevent effects of self-presentational concerns, given the greater difficulty of controlling responses on implicit compared to explicit measures (Norman & Chen, this issue; Olson & Gill, this issue). We generally agree with this idea. However, the obvious value of implicit measures for studying unintentional effects that are hard to control does not imply that implicit measures are useful for capturing unconscious effects of social category cues. Moreover, when using implicit measures for one or more of these purposes, researchers should take into account that implicit measurement scores show rather low temporal stability (Gawronski et al., 2017), which undermines their suitability for predicting outcomes over time. The low temporal stability of implicit measurement scores is just one among several pieces of evidence that is difficult to reconcile with the dominant narrative suggesting that implicit measures capture trait-like unconscious representations that coexist with functionally independent conscious representations (see Gawronski et al., 2022). Instead, the available evidence aligns better with alternative frameworks that treat responses on implicit measures as the product of dynamic processes that operate on currently activated, consciously accessible information.

Some commentators also noted the potential value of implicit measures for studying biases at the regional level as opposed to the individual level (Cyrus-Lai et al., this issue; Melnikoff & Kurdi, this issue). Although this line of work is still in its infancy, it has already produced a large number of interesting findings (for a review, see Calanchini et al., in press), inspiring the development of novel theories of BIM such as the bias-of-crowds model (Payne et al., 2017). Echoing the main points made by the commentators, we are very curious about where this line of work will lead us. However, to avoid premature conclusions, we would like to highlight two issues in research using implicit measures to study regional bias. First, as we explained in our target article, some of the effects obtained in this line of work may reflect little more than the statistical truism that aggregation reduces measurement error (Connor & Evers, 2020). Whereas aggregation of data across individuals isolates situation-related variance by eliminating effects of person-related factors, aggregation of data across situations isolates person-related variance by eliminating effects of situation-related factors (see Gawronski & Bodenhausen, 2017). Second, when aggregating data across individuals to obtain indicators of regional bias, the common dissociations between implicit and explicit measures tend to disappear, in that bias on the two kinds of measures shows high correlations and high overlap in their functional properties (Calanchini et al., in press). Thus, it remains unclear whether implicit measures provide any insights for understanding regional bias that could not be gained from explicit measures.

Moving Forward

Despite the disagreements on specific points addressed in this reply, we feel encouraged by the commentaries that the field might be ready to move on and to overcome the issues raised in our target article. The following list of recommendations, which integrates the key points of our target article and the valuable insights provided by the commentaries, may provide some guidance in this endeavor (see Table 1):

Table 1. Recommendations for advancing research on unconscious bias inspired by the commentaries on the target article by Gawronski, Ledgerwood, and Eastwick (this issue).

1. Clearly distinguish between biased behavior as a phenomenon that needs to be explained and mental constructs proposed to explain biased behavior.
2. Clearly distinguish between effects of social category cues on behavioral responses as an empirical phenomenon and moral evaluations of such effects as desirable or undesirable.
3. Avoid using the ambiguous terms *implicit* and *automatic*. Instead use the more precise terms *unconscious* and *unintentional*, and clearly distinguish between the two.
4. Be clear that, although bias on implicit measures is unintentional, there is no conceptual or empirical basis to describe it as unconscious.
5. Redirect attention from bias on implicit measures to actual instances of unconscious bias and potential underlying mechanisms (e.g., biased interpretation, biased weighting).
6. Use stringent methods to empirically corroborate claims about the presumed unconsciousness of bias.
7. Use sample sizes that provide sufficiently high power and follow open and inclusive science practices, including preregistration, open data, open materials, and inclusive research teams.

1. Expanding on earlier concerns in the attitude literature and building on the current discussion on how to define *bias*, we would encourage bias researchers to clearly distinguish between instances of biased behavior that need to be explained (*explanandum*) and the mental constructs that are proposed to explain biased behavior (*explanans*). Doing so not only increases conceptual precision; it also avoids logical fallacies and circular explanations in the interpretation of empirical findings. Our definition of bias as an effect of social category cues on behavioral responses meets this criterion by providing a purely behavioral definition that does not invoke any reference to underlying mental constructs. Although mental constructs (e.g., attitudes, beliefs, stereotypes, motivation) are undeniably important for understanding biased behavior, they should not be conflated with the behavior they are proposed to explain.
2. Similar to the distinction between behavioral phenomena and explanatory mental constructs, bias researchers should be mindful of the difference between effects of social category cues on behavioral responses and evaluations of such effects as desirable or undesirable. By defining *bias* as an effect of social category cues on behavioral responses, we leave space for the important question of whether specific instances of such effects are beneficial or harmful, and for the answer to this question to change depending on the social context and perceivers' goals and values. Moral evaluations of bias invoke crucial considerations pertaining to extant power structures and histories of oppression and unfair treatment that cannot and should not be ignored. This does not mean that bias researchers should refrain from participating in societal discourses about bias or that they should avoid taking a stance on these issues in their

scientific publications—quite the contrary. Yet, when they do so, they should clearly distinguish between empirical effects and moral evaluations of these effects, and be explicit about the assumptions and values underlying the latter. We believe that explicit discussions of both aspects are superior for advancing applications of bias research than tacitly assuming that everyone shares one's values or that one is more objective by hiding one's personal vantage point and assumptions.

3. Expanding on the debate about the meaning of the term *implicit*, we discourage using the term *implicit* in reference to bias. Use of the term *implicit* is just too flexible and inconsistent to ensure conceptual precision. Greater precision could be easily achieved by using the terms *unconscious* and *unintentional*, and by clearly distinguishing between the two instead of lumping them under the imprecise umbrella term *automatic*.
4. Although it seems empirically justified to describe the biases captured by implicit measures as unintentional, researchers should not describe them as unconscious. Moreover, if researchers want to claim that the biases observed in their studies are unconscious, they should provide empirical evidence that supports their claims. We hope we were able to convince our readers that there is no conceptual and empirical basis for describing the biases captured by implicit measures as unconscious. Biases on implicit measures clearly tend to be unintentional, but that is not same as unconscious.
5. Given the potential significance of unconscious bias for the perpetuation of social disparities and the sphere of inattention for unconscious bias caused by the dominant focus on implicit measures, we would encourage researchers who are interested in unconscious bias to shift their attention from bias on implicit measures to studying unconscious effects of social category cues. Of course, some researchers may not feel particularly strongly about whether the biases in their studies are conscious or unconscious. That is perfectly legitimate. However, in such cases, it would seem appropriate not to make unfounded claims about unawareness.
6. If researchers want to make claims about unawareness, they should use stringent methods to establish unawareness and follow current best practices in research on unconscious mental processes. Ideally, research on unconscious bias would combine multiple approaches to test for unawareness, compensating for idiosyncratic limitations of particular approaches.
7. Field research suggests that effects of social category cues in real-world settings are frequent and often quite strong (e.g., Bertrand & Mullainathan, 2004; Bordieri et al., 1997; Moss-Racusin et al. 2012). It is conceivable that at least some of these effects are driven by unconscious mechanisms involving biased interpretations of ambiguous information or biased weighting of mixed information. However, the number of studies suggesting unconscious effects of social category cues is still very small and a considerable portion of these studies were conducted prior to the adoption of current best practices. We therefore deem it especially important to use sample sizes that provide sufficiently high power and to follow open and inclusive science practices, including preregistered analysis plans, open data, open materials, and research teams that include multiple vantage points (see Ledgerwood et al., in press). Registered reports with peer review prior to the collection of data would be especially valuable.

We hope that these recommendations are helpful in moving the field forward. Theoretically, it seems very plausible that unconscious effects of social category cues contribute to social disparities in a significant manner. Yet, based on the currently available evidence, any such claims are premature, partly because the widespread equation of IB and BIM has distracted the field from studying actual instances unconscious bias. If we care about social disparities and the possibility that they are perpetuated by unconscious bias, it seems prudent to go back where we stopped more than 25 years ago when research on BIM took over. Research using implicit measures clearly has taught us a lot, but counter to the dominant narrative it has not taught us much about unconscious bias.

Funding

Preparation of this article was supported by National Science Foundation Grant BCS-1941440. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID

Bertram Gawronski http://orcid.org/0000-0001-7938-3339
Alison Ledgerwood http://orcid.org/0000-0002-4535-6276
Paul W. Eastwick http://orcid.org/0000-0001-8512-8721

References

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. New York: Delacorte Press.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. doi:10.1257/0002828042002561

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679.

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*(1), 5–25.

Bordieri, J. E., Drehmer, D. E., & Taylor, D. W. (1997). Work life for employees with disabilities: Recommendations for promotion. *Rehabilitation Counseling Bulletin*, *40*, 181–191.

Calanchini, J., Hehman, E., Ebert, T., Esposito, E., Simon, D., & Wilson, L. (in press). Regional intergroup bias. *Advances in Experimental Social Psychology*.

Calanchini, J., Sherman, J., Klauer, C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality & Social Psychology Bulletin*, *40*(10), 1285–1296.

Cervone, D., Shadel, W. D., & Jencius, S. (2001). Social-cognitive theory of personality assessment. *Personality and Social Psychology Review*, *5*(1), 33–51. doi:10.1207/S15327957PSPR0501_3

Cikara, M., Martinez, J. E., & Lewis, N. A. (in press). Moving beyond social categories by incorporating context in social psychological theory. *Nature Reviews Psychology*.

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science*, *15*(6), 1329–1345.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The quad-model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487.

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*(3), 212–232. doi:10.1177/1088868320911325

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

Correll, J., Wittenbrink, B., Crawford, M., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate complex visual stimuli. *Journal of Personality and Social Psychology*, *108*(2), 219–233. doi:10.1037/pspa0000015

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, *84*, 103812. doi:10.1016/j.jesp.2019.04.010

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science*, *14*(5), 835–840. doi:10.1177/1745691619855638

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*(1), 252–287. doi:10.1080/10463283.2014.892320

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584. doi:10.1037/0022-3514.63.4.568

Eagly, A. H., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition*, *25*(5), 582–602. doi:10.1521/soco.2007.25.5.582

Elder, J., Wilson, L., & Calanchini, J. (2022). *Estimating the stability of implicit processes: Implicit attitudes reflect the contribution of both state and trait-like cognitive processes*. *PsyArXiv*. doi:10.31234/osf.io/sm2yt

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology*, *54*(1), 297–327. doi:10.1146/annurev.psych.54.101601.145225

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. doi:10.1037/0022-3514.50.2.229

Fleeson, W., & Jayawickreme, E. (2021). Whole traits: Revealing the social-cognitive mechanisms constituting personality's central variable. *Advances in Experimental Social Psychology*, *63*, 69–128.

Fryberg, S. A., & Stephens, N. M. (2010). When the world is color-blind, American Indians are invisible: A diversity science approach. *Psychological Inquiry*, *21*(2), 115–119. doi:10.1080/1047840X.2010.483847

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*(4), 574–595. doi:10.1177/1745691619826015

Gawronski, B., & Bodenhausen, G. V. (2012). Self-insight from a dual-process perspective. In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 22–38). New York: Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2015). Social-cognitive theories. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 65–83). New York: Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry*, *28*(4), 268–272. doi:10.1080/1047840X.2017.1373546

Gawronski, B., Brownstein, M., & Madva, A. (2022). How should we think about implicit measures and their empirical "anomalies."? *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(3), e1590.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). New York: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition*, *38*(Supplement), s1–s25. doi:10.1521/soco.2020.38.supp.s1

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

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2020). Implicit bias and anti-discrimination policy. *Policy Insights from the Behavioral and Brain Sciences*, *7*(2), 99–106. doi:10.1177/2372732220939128

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality & Social Psychology Bulletin*, *43*(3), 300–312.

Goedderz, A., & Hahn, A. (2022). Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions. *Journal of Experimental Social Psychology*, *102*, 104374. doi:10.1016/j.jesp.2022.104374

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*(9), 861–871.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*(4), 945–967. doi:10.2307/20439056

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. doi:10.1037/0022-3514.74.6.1464

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology. General*, *143*(3), 1369–1392.

Helms, J. E., Jernigan, M., & Mascher, J. (2005). The meaning of race in psychology and how to change it: A methodological perspective. *American Psychologist*, *60*(1), 27–36.

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: Defensive responses to IAT feedback among whites, blacks, and biracial black/whites. *Social Psychological and Personality Science*, *6*(4), 373–381. doi:10.1177/1948550614561127

Hudson, S. K. T. J., Cikara, M., & Sidanius, J. (2019). Preference for hierarchy is associated with reduced empathy and increased counter-empathy towards others, especially out-group targets. *Journal of Experimental Social Psychology*, *85*(103871), 103871. doi:10.1016/j.jesp.2019.103871

Jones, J. M. (1998). Psychological knowledge and the new American dilemma of race. *Journal of Social Issues*, *54*(4), 641–662. doi:10.1111/j.1540-4560.1998.tb01241.x

Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology*, *13*(1), 1–44. doi:10.1002/ejsp.2420130102

Ledgerwood, A., da Silva Frost, A., Kadirvel, S., Maitner, A., Wang, Y. A., & Maddox, K. B. (in press). Methods for advancing an open, replicable, and inclusive science of social cognition. In D. E. Carlston, K. Johnson, & K. Hugenberg (Eds.), *The Oxford handbook of social cognition*. New York: Oxford University Press.

Ledgerwood, A., Eastwick, P. W., & Smith, L. K. (2018). Toward an integrative framework for studying human evaluation: Attitudes towards objects and attributes. *Personality and Social Psychology Review*, *22*(4), 378–398. doi:10.1177/1088868318790718

Ledgerwood, A., Pickett, C., Navarro, D., Remedios, J., & Lewis, N. (2022). The unbearable limitations of solo science: Team science as a path for more rigorous and relevant research. *Behavioral and Brain Sciences*, *45*, E81. doi:10.1017/S0140525X21000844

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, *82*(1), 5–18.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. doi:10.1037/0022-3514.37.11.2098

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*(6149), 976–980.

Mora, Y., Klein, O., Leys, C., & Smeding, A. (2022). External validity of social psychological experiments is a concern, but these models are useful. *Behavioral and Brain Sciences*, *45*, E84.

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*, 16474–16479.

Mueller, J. C. (2020). Racial ideology or racial ignorance? An alternative theory of racial cognition. *Sociological Theory*, *38*(2), 142–169. doi:10.1177/0735275120926197

Norman, E. (2010). The unconscious" in current psychology. *European Psychologist*, *15*(3), 193–201. doi:10.1027/1016-9040/a000017

Okonofua, J. (2022). Controlled lab experiments are one of many useful scientific methods to investigate bias. *The Behavioral and Brain Sciences*, *45*, E85. doi:10.1017/S0140525X21000650

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. doi:10.1037/0022-3514.81.2.181

Payne, B. K., & Banaji, M. R. (2022). Two thousand years after Archimedes, psychologist finds three topics that will simply not yield to the experimental method. *Behavioral and Brain Sciences*, *45*, E87.

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology*, *62*, 1–50.

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). New York: Guilford Press.

Payne, B. K., Shimizu, Y., & Jacoby, L. J. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentification. *Journal of Experimental Social Psychology*, *41*(1), 36–47. doi:10.1016/j.jesp.2004.05.001

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. doi:10.1080/1047840X.2017.1335568

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Reddy, G., & Amer, A. (2022). *Precarious engagements and the politics of knowledge production: Listening to calls for reorienting hegemonic social psychology*. Unpublished manuscript.

Redford, L., & Ratliff, K. A. (2016). Perceived moral responsibility for attitude-based discrimination. *British Journal of Social Psychology*, *55*(2), 279–296.

Salter, P. S., Adams, G., & Perez, M. J. (2018). Racism in the structure of everyday worlds: A cultural-psychological perspective. *Current Directions in Psychological Science*, *27*(3), 150–155. doi:10.1177/0963721417724239

Schmader, T., Dennehy, T. C., & Baron, A. S. (in press). Why anti-bias interventions (need not) fail. *Perspectives on Psychological Science*.

Sidanius, J., Pratto, F., Van Laar, C., & Levin, S. (2004). Social dominance theory: Its agenda and method. *Political Psychology*, *25*(6), 845–880. doi:10.1111/j.1467-9221.2004.00401.x

Stern, C. (2022). Political ideology and social categorization. *Advances in Experimental Social Psychology*, *65*, 167–233.

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 21–46). Oxford, UK: Oxford University Press.

Yi, J., Neville, H. A., Todd, N. R., & Mekawi, Y. (in press). Ignoring race and denying racism: A meta-analysis of the associations between colorblind racial ideology, anti-Blackness, and other variables antithetical to racial justice. *Journal of Counseling Psychology*.