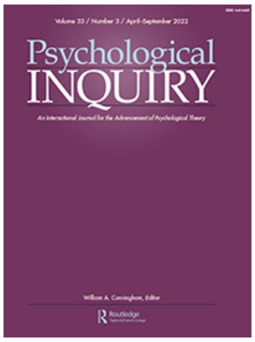

Psychological Inquiry

An International Journal for the Advancement of Psychological Theory

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hpli20

Implicit Bias as Automatic Behavior

Kate A. Ratliff & Colin Tucker Smith

To cite this article: Kate A. Ratliff & Colin Tucker Smith (2022) Implicit Bias as Automatic Behavior, Psychological Inquiry, 33:3, 213-218, DOI: 10.1080/1047840X.2022.2106764

To link to this article: https://doi.org/10.1080/1047840X.2022.2106764

Published online: 19 Sep 2022.

PSYCHOLOGICAL INQUIRY
2022, VOL. 33, NO. 3, 213–218
https://doi.org/10.1080/1047840X.2022.2106764

COMMENTARIES

Implicit Bias as Automatic Behavior

Kate A. Ratliff and Colin Tucker Smith

University of Florida, Gainesville, Florida

Researchers interested in implicit bias agree that no one agrees what implicit bias is. Gawronski, Ledgerwood, and Eastwick (this issue) join a spate of scholars calling for better conceptual clarity around what it means for a construct or a measure to be *implicit* (Corneille & Hütter, 2020; Fazio, Granados Samatoa, Boggs, & Ladanyi, 2022; Schmader, Dennehy, & Baron, 2022; Van Dessel et al., 2020). Some argue we should do away with the term entirely (Corneille & Hütter, 2020), and others argue that authors simply need to do a better job defining how they are idiosyncratically using the term each time they use it (Greenwald & Lai, 2020). In their target article, Gawronski et al. argue for a fundamental redefinition of what it means for bias to be implicit. More specifically, they argue that *implicit bias* (IB) and *bias on implicit measures* (BIM) are conceptually and empirically distinct, and that BIM (defined as "effects of social category membership on behavioral responses captured by measurement instruments conventionally describe as implicit") should not be treated as an instance of IB (defined as "behavioral responses influenced by social category cues when respondents are unaware of the effect of social category cues on their behavioral responses").

We agree that the time has come for our definition of *implicit* to be revamped in light of new findings. In fact, it is past time; we co-chaired a symposium titled "What is implicit about implicit attitudes?" at the Society for Personality and Social Psychology's annual meeting in 2009, more than a decade ago. And we applaud the authors of the target article for taking a bold step toward making a change. Further, we agree with them that *bias is best defined as a behavioral phenomenon* rather than a latent mental construct. This is not a statement we make lightly; it has required some serious scholarly contemplation of the current state of the literature and some serious non-scholarly contemplation of our own egos to reach this conclusion. For some time now we, like most others, have described implicit bias as something that people have–e.g., participants *have* an implicit bias favoring one novel individual over another (Ratliff & Nosek, 2011), *have* an implicit preference favoring White over Black Americans (Chen & Ratliff, 2018), or *have* an implicit positive or negative attitude toward feminists (Redford, Howell, Meijs, & Ratliff, 2018). Many of us are quite invested in this way of thinking. And change is hard! But we recognize that we gain a lot by taking this more functional approach to bias. Most notably, a functional approach allows researchers to circumvent the perplexing situation of using the same name for construct and measure. Further, many of us working in this area are doing so because we hope to provide insights through which people can change their behavior in order to reduce inequality on real life issues that matter. Given that the problem of bias is a behavioral problem (De Houwer, 2019), it makes sense to define bias in behavioral terms. So let us agree to define bias as the influence of social category cues on behavioral responses.[1]

But we are still left, however, with the problem of what makes bias *implicit*. To that end, we would like to raise two concerns about the target article. First, if the authors had proposed that BIM should not *necessarily* be treated as an instance of IB, we would concur; but we disagree with the strong language implying that BIM should *never* be considered an instance of IB; "does not equal" is not the same as "is orthogonal to." Second, we do not agree that *awareness* (which the original authors use interchangeably with *consciousness*) is the only or best factor by which to distinguish implicit from explicit bias. Consciousness is messy business, and it is nearly impossible to delineate whether any given effect is unconscious or conscious as most, maybe all, have aspects of both. We would instead argue for distinguishing between implicit and explicit bias based on features of automaticity (Moors & De Houwer, 2006).

On the Mutual Exclusivity of Bias on Implicit Measures and Implicit Bias

Can bias on implicit measures (BIM) also be implicit bias (IB)? We do not see why not. It is now obvious that it is conceptually and empirically problematic to use the term *implicit* to describe both a latent mental construct and to describe the measures purported to assess that latent construct. However, if one demonstrates BIM, which the original authors define[2] as "effects of social category cues on

CONTACT Kate A. Ratliff ratliff@ufl.edu University of Florida, Gainesville, FL, USA.

[1] We also note our appreciations for the original authors' statement that the phrase *effect of social category cues on behavioral responses* should not be taken to mean that the cause of dominant social group members' racist or sexist behavior is located within a marginalized group member.

[2] Given the authors' careful attention to detail in definition, we are somewhat puzzled by the imprecise and ambiguous nature of this one. We are not confident there is, in fact, an agreed-upon convention. Examples that come easily to mind are speeded self-report and self-reported gut reactions which each load comfortably on a latent factor with measures that are "conventionally described as implicit" (Ratliff (Ranganath), Smith, & Nosek, 2008). The AMP is similarly ambiguous (Bar-Anan & Nosek, 2012; Hughes, Cummins, & Hussey, 2022; Payne et al., 2013). Further, conventions can change over time, as they should as new evidence accrues. For example, the Modern Racism scale was once presented in contrast to "old-fashioned racism" as being "nonreactive" and relatively immune to "faking" (see McConahay, Hardee, & Batts, 1981). Although this is now clearly EB, it could have at one time been considered BIM had the term existed.

behavioral responses captured by measurement instruments conventionally describe as implicit," and if performance on that measure can be considered a behavioral response influenced by social category cues when "respondents are unaware of the effect of social category cues on their behavioral responses" (Gawronski et al., this issue, pp. 139–140) which is the original authors' definition of IB, then why would BIM *not* be an example of IB? The authors briefly raise the possibility that BIM could be IB, but then dismiss it entirely, seemingly because there is *currently* no example that meets their definition of BIM that occurs outside of conscious awareness. We also question whether the authors' argument means that, in those situations where bias on a measurement instrument conventionally described as implicit does not meet the criteria for implicit bias, does that mean it is then explicit bias? Or is it a third category and must we now decide whether any novel measure that is developed is similar enough to those that are currently conventionally described as implicit to fall into the original category versus the new category? In sum, it is not entirely clear to us why we need the category of BIM at all. Using the original authors' terminology, if a social group cue influences a behavior outside of awareness then it is implicit bias; if it influences a social group cue with awareness then it is explicit bias. The nature of the behavior—whether categorizing stimuli or choosing where to sit—seems largely irrelevant for the distinction between implicit and explicit bias.

On Using (Un)Consciousness to Distinguish Between Implicit and Explicit Bias

In the previous section we argued that there may be cases in which BIM is an example of IB if one is unaware of the effect of social category membership on behavioral responses on a measurement instrument conventionally described as implicit. If we accept that IB hinges on awareness, this is likely a moot point, at least for the foreseeable future, as there are few, if any, measures of any type that could meet the burden of proof of producing effects that are entirely outside of conscious awareness. This brings us to our second point—that, if we must distinguish between whether an effect is implicit or explicit bias, (un)consciousness is not the best factor by which to do so because awareness: (a) is complex and nearly impossible to prove, and (b) ignores the importance of an actor's intentions, a feature of automaticity that is likely to be particularly relevant to implicit bias.

The Complexity of (Un)awareness

To their credit, the original authors acknowledge the complexity of defining (un)awareness, but they offer little by way of guidance for how researchers might determine whether a particular effect is or is not conscious. We infer from the examples presented that the original authors favor a strict and binary definition of consciousness, whereby any effect that is not clearly demonstrated to be entirely outside of awareness should be considered conscious (and therefore explicit bias). For example, they note that, because Hugenberg and Bodenhausen (2003) did not confirm unawareness in their studies that demonstrate White participants' greater readiness to perceive anger in African American (compared to White American faces), we cannot interpret the effect to be an example of IB. Of course, they who write the target article get to make the definition, but we do not agree with defining awareness, and thus implicit bias, in such narrow terms.

Nisbett and Wilson recognized the complex nature of awareness in 1977, noting that experimental manipulations may influence verbal reports on cognitive processes without participants' awareness of: (1) the existence of a response (e.g., participants' observable behavior is influenced by an experimental manipulation but they cannot verbally report that it has done so), (2) the existence of a change response (e.g., participants' observable behavior is *changed* by an experimental manipulation but they cannot verbally report that it has done so), or (3) the stimuli that produced the response (e.g., a stimulus presented outside of the threshold of conscious awareness influences a behavior).

Fast-forward nearly two decades; Bargh (1994) argues that there are three ways in which a person may be unaware of a mental process, by being unaware of: (1) the stimulus itself (e.g., subliminal perception), (2) the way in which a stimulus event is interpreted or categorized (e.g., stereotype influencing a judgment without participants' awareness of it having done so), or (3) the determining influences on his or her judgements or subjective feeling states (e.g., seeing a stimulus multiple times leads to more positive evaluations of that stimulus without participants being aware that their liking is based on mere exposure).

Fast forward, again, nearly three decades; Hahn and Goedderz (2020) propose two definitions of unconsciousness: (1) *trait-unconsciousness*, whereby it is impossible for an actor to know about the existence of a cognition unless they are informed of its existence from an outside source, and (2) *state-unconsciousness* (also called *preconsciousness*), whereby an actor does not know about the existence of a cognition at a given moment, because they are not thinking about it, though it is not impossible for them to do so once their attention is appropriately directed. Although Hahn and Goedderz apply their distinction to cognitions, rather than to observable behaviors, the point remains—awareness is complicated and there are multiple ways to define what it means for someone to be (un)aware of the influence of social cues on their thinking, feeling, or behavior. This complexity is particularly salient when considering what the original authors call BIM. Let us take two such measures—the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and the Affective Misattribution Procedure (AMP; Payne, Cheng, Govoron, & Stewart, 2005)—as examples.

The target article suggests that "the IAT … would not qualify as implicit in this sense, because respondents are typically aware of the effects of social category cues on their responses in the task" (Gawronski et al., this issue, p. 140). This and other similar statements treat the conscious nature of the IAT as settled science, but we believe the reality is more complicated. To be clear, we are not arguing that people are *entirely* unaware of the effects of social category cues on their IAT performance. However, people are far from being able to completely and accurately report what those effects are.

The original authors also note Monteith, Voils, and Ashburn-Nardo's (2001) finding which they summarize as "[m]ost people quickly notice that their responses are slower and that they make more errors in the bias-incongruent block than in the bias-congruent block" (Gawronski et al., this issue, p. 140). It is not obvious to us that noticing one's performance in the middle of a particular task "counts" as being aware of the effect. At what point during the process does awareness matter? The first trial? The halfway point? If a participant realizes during the final moments of a task that their performance has been influenced by social category cues, is that effect then explicit? We also note that 36% of the participants in this study were inaccurate even regarding the *direction* of their bias on the IAT. At what point would we say that this is evidence of awareness? We genuinely do not mean to be facetious with these rhetorical questions; our intention is to point out that awareness is a tricky concept and the evidence cited is not as airtight as it is presented in the target article.

Another potential piece of evidence that people are aware of their bias on the IAT is a paper by Hahn and colleagues (2014) in which participants see a set of IAT stimuli for five tasks (*Black* vs. *White*, *Latino* vs. *White*, *Asian* vs. *White*, *Celebrity* vs. *Regular Person*, and *Child* vs. *Adult*), and then predict their IAT performance on those tasks using a scale anchored by, for example, "Sorting pictures of the category BLACK with GOOD (and WHITE with BAD) will be a lot easier" and "Sorting pictures of the category WHITE with GOOD (and BLACK with BAD) will be a lot easier." Awareness is operationalized on the basis of within-person correlations–essentially testing whether people can predict the size of a set of IAT effects relative to one another. While these results are certainly suggestive that people are aware of whatever content forms the basis of their IAT scores, we note that, even if each individual prediction is inaccurate, so long as each inaccurate prediction lines up in the right order as BIM, it is referred to as accuracy. Hahn and colleagues (2014) themselves do not make an argument that people are fully aware, writing "we interpret these results to mean that our participants had *some* awareness of their implicit attitudes" (emphasis ours, p. 26). We agree with this point, but still question how a distinction between IB and EB that relies on awareness can account for the majority of effects that are not "all or nothing." As Hahn and Goedderz (2020) write, "The cognitions reflected on implicit evaluations [BIM] are often referred to as unconscious attitudes. Our analysis suggests that this is an incorrect characterization when the term *unconscious* is used as a trait that describes those cognitions at all times. However, when unconscious is defined as a state in which cognitions can be at specific points in time, then our data are compatible with the nation that the cognitions reflected on implicit measures can be unconscious" (pp. s130–s131).

Although we risk getting too far into the weeds on this point, we also want to note that there is similar ambiguity around performance on the AMP. Evidence for awareness of AMP performance comes from methodologically rigorous studies showing that nearly all participants report at least some trials in which they report consciously using a prime to evaluate the target; in those trials AMP effects are larger, more reliable, and relate more strongly to other measures (Bar-Anan & Nosek, 2012; Hughes et al., 2022).[3] If any conscious awareness of the influence of a social group cue on performance is all that is needed to disqualify a behavior as an example of IB, then the AMP is pretty clearly not IB. However, what about those participants who did not report awareness of using of the primes in their evaluation of the targets? Or those *trials* on which participants did not report using the primes in their evaluation of the target? Bar-Anan and Nosek (2012) write: "Even participants who reported no awareness or control of the priming effect still evaluated targets differently when they followed different prime categories" (p. 1204; Bar-Anan & Nosek, 2012). Hughes et al. (2022) find the same (and we highly recommend their article for nuanced, in-depth discussion of this issue).

Finally, the authors make the point that surprise about one's IAT feedback should not be considered evidence that the IAT effect (the effect of social category cues on their ability to categorize stimuli more easily when paired with "Good" relative to "Bad") is implicit (i.e., outside of awareness). A series of studies using nine different topics show that people respond defensively to feedback about implicit bias to the extent that their self-reported attitudes and their IAT scores are discrepant (Howell, Gaither, & Ratliff, 2015; Howell, Redford, Pogge, and Ratliff, 2017; Howell & Ratliff, 2017). These results replicate when we look specifically at an item about their experience taking the IAT being "eye-opening" (unpublished data). Gawronski et al. (this issue) argue that such surprise might indicate a mismatch between the metric by participants to describe the extremity of their bias and the metric used by researchers to convert numeric IAT scores into verbal feedback (e.g., your performance indicates a preference for X over Y) rather than surprise about the feedback itself. We are not convinced that this is the case and have reason to suspect that people are genuinely surprised.

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

[3] We note that, although these authors describe the effects as being about *awareness* of the influence of primes on targets, Bar-Anan and Nosek (2012) asked participants whether or not they *intentionally* rated the primes instead of the target. In this case, an intentional decision to rate the primes necessarily indicates awareness, so their awareness-based interpretation is not incorrect, but we do argue later in this commentary that we would prefer that intentionality, rather than awareness, be used to distinguish between implicit and explicit bias, and note that some scholars are already doing so and calling it awareness.

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.

To conclude this section: We reiterate our agreement with the original authors that bias is best defined as a behavioral phenomenon rather than a latent mental construct, but we do not agree that BIM can never be IB. While we are not saying that any BIM is definitively IB for any particular task, we are simultaneously unconvinced by the evidence that BIM cannot be an instance of IB.[4] For us, the door is still open, for one reason because we have different interpretations of the evidence regarding awareness of BIM. We also do not believe that asking for an effect to occur entirely outside of conscious awareness in order to be considered IB is a useful requirement given the difficulty of demonstrating (un)awareness. We anticipate that an unconsciousness requirement will lead to authors, reviewers, and editors spending large amounts of time going back and forth about whether some particular effect is implicit or explicit based on the point at which someone decides whether or not a particular effect is unconscious enough, when that effort would be better spent evaluating other questions, such as whether a particular effect (e.g., IAT performance) actually *matters* in some way. As an alternative to awareness, we favor distinguishing between implicit and explicit bias based on an automaticity distinction, which allows accounting for an actor's *intentions*, as we describe in the final section below.

The Importance of Intentions

In the previous section we argue that (un)awareness is simply too messy of a construct to use to distinguish between implicit and explicit bias. We also believe that using a strict definition of consciousness unnecessarily excludes too many phenomenon that might reasonably be considered implicit on the basis of automaticity.

Moors and De Houwer (2006) conceptualize automaticity as a process that influences task performance (i.e., behavior) in a way that has one or more of the following features: unintentional, goal-independent, autonomous, unconscious, efficient, and/or fast (see also De Houwer & Moors, 2007). De Houwer (2019) proposes that, when performance on a task (including those the original authors call BIM) is influenced by one of these processes, it is reasonable to equate *implicit bias* and *automaticity*, using *automatic* as an umbrella term and then specifying the relevant different features of automaticity at play. This is also consistent with the perspective proposed by Fazio et al. (2022) who argue that implicit bias should not be conceptualized as an unconscious construct but instead as an effect of attitudes that are activated automatically from memory, and with Schmader et al. (2022) who argue that implicit bias is discriminatory action or judgment against a group due to biases that the perceiver is unaware of or unable to effectively regulate (i.e., unintentional) in that moment. Corneille and Hütter (2020) and Gawronski et al. (this issue) argue that automaticity creates conceptual ambiguity; however, De Houwer and Moors (2010) and De Houwer (2019) are clear that it is always necessary to specify the particular automaticity features that characterize performance on a given task (De Houwer & Moors, 2010). Further, we believe that not all potential features of automaticity will be as relevant now that they are being applied to an effect rather than to a process or latent mental construct.

Of the particular features of automaticity, intentionality (i.e., whether or not one has control over the startup of a process; Bargh, 1994) and control (i.e., whether or not one can override a process once started) are highly relevant to distinguishing between implicit and explicit bias. For simplicity moving forward, we will use *intentions* to describe both as these are essentially the same thing when applied to bias as a behavioral effect.[5] For distinguishing between implicit and explicit effects of social group cues on behaviors, the key question is "did the actor *intend* to use social category cues in their behavioral response?"

Intentionality may have some overlap with awareness; if social cues influence behavior entirely outside of awareness, it is unlikely that an actor intended for those social cues to influence their behavior. But, taking intentionality into account allows for partial, inaccurate, and fleeting awareness and expands the range of phenomenon that can be considered examples of implicit bias, especially if we do not sort tasks into BIM and non-BIM and instead decide whether any particular effect is implicit or explicit bias regardless of whether the task has conventionally been described as implicit.[6]

Take, for example, a situation in which an actor knows that their behavioral response is impacted by social category cues, *but they are unable to stop it from happening*. Imagine participants choosing a starting salary for a job candidate with candidate gender identity manipulated between subjects. Participants assign a higher starting salary to the man than the woman, and they are not aware that they have used the candidate's gender in their decision. This is a

[4] We caution against treating a procedure as a fixed measure (i.e., referring to "*the* IAT"). It is possible, for example, that IATs for some attitude objects are more likely to show evidence of IB than others.

[5] You may wonder why, if we're going to use automaticity to distinguish between implicit and explicit bias, we do not just call it automatic bias. Touché! But the same point could be made about awareness; that is, we could do away with the term implicit (Corneille & Hütter, 2020) and instead refer to conscious and unconscious bias. But the horse is out of the barn (unless it is an extremely long horse) where implicit bias is concerned and, at this point, it seems like we are best served by collectively deciding on a clear definition than by trying to eliminate the term completely.

[6] The original authors make the point that automaticity cannot be used to distinguish between implicit and explicit bias because unintentionality and uncontrollability do not overlap with unawareness. This argument is only relevant, however, if one accepts their premise that awareness is the feature by which we should distinguish implicit and explicit bias.

demonstration of IB. Now imagine another group of participants who, prior to assigning a starting salary, are reminded that women are often paid less for equal work and are given a strong incentive to avoid inequity. For these participants, the gap between the salary they assign to the man and the women is smaller than it is for those participants who did not receive instructions. This indicates that participants in the second group have consciously taken the candidate's gender identity into account. But what if the gap between the salary they assign to men and women is not eliminated? That is, they treat the candidates differently based on their gender? Using an awareness criterion, this cannot be IB because participants were aware of using gender information. But surely it matters that they produced an *unintended* disparity? It lacks face-validity to consider this second situation to be explicit bias but the first to be implicit bias. It seems well-within reason to consider performance implicit to the extent that one does not want or intend to use social group cues in task performance but does so anyway.

As noted previously, using automaticity (and particularly intentions) to distinguish been implicit and explicit bias expands the range of phenomenon that can be considered examples of implicit bias and would possibly put some so-called BIM into the IB category as there are tasks, such as the IAT, on which people might be made aware of their biased performance during the task (though see caveats above), but are in some cases unable to prevent that biased performance (Hawkins & Ratliff, 2015) without specific instructions about how to do so (see Röhner, Holden, & Schütz, 2022, for a recent review).[7]

There are also pragmatic reasons to take intention into account in distinguishing implicit from explicit bias. First, intentionality is easier to assess than unconsciousness. By definition, people cannot experience something they are unaware of, but people can report whether or not they intended to use some piece of information or produce some effect. In fact, that is the very definition of intention, which is convenient. While of course there are potential pitfalls of motivated reasoning, self-report, and demand characteristics, at least there is the *possibility* of people reporting on what it is that they meant to do, and the same strategies the authors propose to determine lack of awareness could be applied to intentionality as well.

We also note that framing bias as *unconscious* has been shown to lead to reduced accountability for discrimination. Redford and Ratliff (2016a, 2016b) conducted a series of studies in which we told participants about a hiring manager who engaged in discriminatory behavior against Black employees, and found that people held him less morally responsible for his behavior when he was unaware of his racial bias compared to when he was aware. Similarly, Daumeyer, Onyeador, Brown, and Richeson (2019) found that attributions of discrimination to implicit bias, compared to explicit bias, resulted in lower judgements of accountability and, in some cases, lower intention to punish; importantly, the definition of implicit bias they gave to their participants was one hinging on *unawareness.*

In their discussion of the politics of communicating our science to the general public, Fazio et al. (2022) write "It is critically important that we be careful about the language we use in pursuing the science of implicit bias." Daumeyer, Rucker, and Richeson (2017) also caution that the implications of our conceptualization of implicit bias will have consequences for how people reason about discrimination. By presenting bias as unconscious, and thus clearly outside of what people could control, we risk promoting lowered accountability for biased behavior.

You may wonder whether framing implicit bias as unintentional may produce similar effects of lowered accountability. While it is indeed the case that people hold others more morally responsible for behavior that they intended compared to that which is unintended (Malle, Guglielmo, & Monroe, 2014), people also hold others accountable for behavior they *should have foreseen* (Lagnado & Channon, 2008; Redford & Ratliff, 2016). A behavior that one is not aware of is inherently unforeseeable, whereas one that is unintentional at least has the possibility of being viewed as foreseeable, and there is a rich literature on which to draw about when it is that people will weight intent or outcome more heavily in assigning moral blame (e.g., McNamara, Willard, Norenzayan, & Henrich, 2019). And, although the original authors suggest that an argument against using intentionality to differentiate between implicit and explicit bias might be that different strategies are needed to combat the harmful effects of conscious versus unintentional bias, they do not explain why this might be the case, nor do we know whether the potential would outweigh the empirically demonstrated costs of framing implicit bias as unconscious. We also note that, thus far, the work on moral responsibility for implicit bias has generally implied that implicit bias is a *latent mental construct* that produces some behavior. It remains to be seen whether or not defining implicit bias *as the behavior itself* will impact moral attributions.

Conclusion

We want to make it very clear that we appreciate Gawronski et al.'s (this issue) attempt to reclaim the narrative and rally scientists together around a new, more useful definition of implicit bias. There is much to like in their wide-ranging review and we reiterate our agreement with what we see as the most important, fundamental, and radical of their arguments—that implicit bias is a behavioral effect. We object, however, to a strong argument that bias on implicit measures (BIM) cannot be an example of implicit bias (IB) and to requiring that an effect occur entirely outside of conscious awareness to be considered an example of implicit bias. Instead, we see it as more reasonable to distinguish implicit from explicit bias on features of automaticity and propose that intentions are particularly useful to consider. Finally, rather than adding a new term (BIM), we argue that

[7] We recognize that some will disagree with the idea that IAT performance is not entirely controllable. That, of course, is an empirical question and an increased focus on intentionality will likely increase attention to the possibility of IAT "faking". The point is that there are likely *more* tasks that will be considered examples of IB if we use a broader factor (e.g., intentions) to distinguish between implicit and explicit bias than a narrow one (e.g., unconsciousness). What those specific tasks are remains to be seen.

the term implicit bias is sufficient to describe all situations in which the effects of social category cues influence behavioral responses in ways that are automatic, regardless of how the behavioral task has conventionally been classified.

References

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the Affective Misattribution Procedure. *Personality & Social Psychology Bulletin*, *38*(9), 1194–1208. doi:10.1177/0146167212446835

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

Chen, J. M., & Ratliff, K. A. (2018). Psychological essentialism predicts intergroup bias. *Social Cognition*, *36*(3), 301–323. doi:10.1521/soco.2018.36.3.301

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

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

Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2017). Thinking structurally about implicit bias: Some peril, lots of promise. *Psychological Inquiry*, *28*(4), 258–261. doi:10.1080/1047840X.2017.1373556

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

De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 179–194). New York, NY: Guilford Press.

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–193). New York, NY: Guilford Press.

Fazio, R. H., Granados Samayoa, J. A., Boggs, S. T., & Ladanyi, J. (2022). Implicit bias: What is it? In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.). *The Cambridge handbook of implicit bias and racism*. Cambridge, UK: Cambridge University Press.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*, 419–445.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (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., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluation. *Social Cognition*, *38*(Supplement), s115–s134. doi:10.1521/soco.2020.38.supp.s115

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.

Hawkins, C. B., & Ratliff, K. A. (2015). Trying but failing: Implicit attitude transfer is not eliminated by overt or subtle objectivity manipulations. *Basic and Applied Social Psychology*, *37*(1), 31–43. doi:10.1080/01973533.2014.995378

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

Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *The British Journal of Social Psychology*, *56*(1), 125–145. doi:10.1111/bjso.12168

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive responding to IAT feedback. *Social Cognition*, *35*(5), 520–562. doi:10.1521/soco.2017.35.5.520

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*(6), 640–643.

Hughes, S., Cummins, J., & Hussey, I. (2022). Effects on the Affect Misattribution Procedure are strongly moderated by awareness. *Behavior Research Methods*. doi:10.3758/s13428-022-01879-4

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. doi:10.1080/1047840X.2014.877340

McConahay, J. B., Hardee, B. B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution*, *25*(4), 563–579. [Database] doi:10.1177/002200278102500401

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition*, *182*, 95–108. doi:10.1016/j.cognition.2018.09.008

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395–417. doi:10.1521/soco.19.4.395.20759

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. doi:10.1037/0033-2909.132.2.297

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi:10.1037/0033-295X.84.3.231

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012). *Personality & Social Psychology Bulletin*, *39*(3), 375–386.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. doi:10.1037/0022-3514.89.3.277

Ratliff, K. A., & Nosek, B. A. (2011). Negativity and outgroup biases in attitude formation and transfer. *Personality and Social Psychology Bulletin*, *37*(12), 1692–1703.

Ratliff (Ranganath), K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*, 386–396.

Redford, L., Howell, J. L., Meijs, M. H., & Ratliff, K. A. (2018). Implicit and explicit evaluations of feminist prototypes predict feminist identity and behavior. *Group Processes & Intergroup Relations*, *21*(1), 3–18. doi:10.1177/1368430216630193

Redford, L., & Ratliff, K. A. (2016a). Hierarchy-legitimizing ideologies reduce behavioral obligations and blame for implicit attitudes and resulting discrimination. *Social Justice Research*, *29*(2), 159–185. doi:10.1007/s11211-016-0260-3

Redford, L., & Ratliff, K. A. (2016b). Perceived moral responsibility for attitude-based discrimination. *The British Journal of Social Psychology*, *55*(2), 279–296. doi:10.1111/bjso.12123

Röhner, J., Holden, R. R., & Schütz, A. (2022). IAT faking indices revisited: Aspects of replicability and differential validity. *Behavior Research Methods*. doi:10.3758/s13428-022-01845-0

Schmader, T., Dennehy, T. C., & Baron, A. S. (2022). Why anti-bias interventions (need not) fail. *Perspectives on Psychological Science*. doi:10.1177/17456916211057565

Van Dessel, P., Cummins, J., Hughes, S., Kasran, S., Cathelyn, F., & Moran, T. (2020). Reflecting on 25 years of research using implicit measures: Recommendations for their future use. *Social Cognition*, *38*(Supplement), s223–s242. doi:10.1521/soco.2020.38.supp.s223