PSYCHOLOGICAL INQUIRY
2022, VOL. 33, NO. 3, 185–192
https://doi.org/10.1080/1047840X.2022.2106759

COMMENTARIES

What Implicit Measures of Bias Can Do

David E. Melnikoff[a] and Benedek Kurdi[b]

[a]Department of Psychology, Northeastern University, Boston, Massachusetts; [b]Department of Psychology, Yale University, New Haven, Connecticut

Gawronski, Ledgerwood, and Eastwick (this issue; GLE) bring much needed attention to the limitations of currently available implicit measures as tools for studying unconscious bias. We agree with the authors of the target article that the current state of the literature offers little reason to believe that commonly used implicit measures, such as sequential priming (Fazio, Sanbonmatsu, Powell, & Kardes, 1986), the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), and the Affect Misattribution Procedure (Payne, Cheng, Govorun, & Stewart, 2005), capture unconscious influences of social category cues on behavioral responses. If anything, the evidence suggests the opposite: Participants may well be aware of how their responses are influenced by social cues on implicit measures (Hahn, Judd, Hirsh, & Blair, 2014; Hahn & Gawronski, 2019), although to what degree and as a result of what type of process or processes remains to be investigated (Morris & Kurdi, 2022).

Nonetheless, even if the extent of awareness differs depending on the specific conditions of the task, the lack of compelling evidence for the ability of currently available implicit measures to index unconscious bias is surprising. As GLE observe, the concepts of unconscious bias and bias on implicit measures have been, and continue to be, conflated, both in the empirical literature and popular discourse. This conundrum will prompt many readers to wonder: If implicit measures of bias are not useful for measuring unconscious bias, are they useful at all?

They are. Whether or not they shed light on unconscious bias, implicit measures have been, and we believe will remain, essential to the study of social cognition. We suspect that the lead author of the target article, who has used implicit measures of bias to make numerous contributions to the understanding of social information processing, would agree. But what is it, exactly, that implicit measures of bias are good for, if not probing the human unconscious? This is the question we address in the current commentary.

Broadly speaking, implicit measures of bias have been and continue to be critical for addressing two related questions: (*i*) what is the nature of unintentional bias? and (*ii*) what is the cognitive architecture of bias? In what follows, we show how implicit measures fuel progress on both fronts while, crucially, also advancing the translational goal of revealing the nature of, and reducing, group-based inequality.

Using Implicit Measures of Bias to Reveal the Nature of Unintentional Bias

The target article maintains a singular focus on the fact that bias can be unconscious. However, unconsciousness is just one of several independent features of automaticity, along with efficiency, uncontrollability, and unintentionality (Bargh, 1989; Melnikoff & Bargh, 2018; Moors & De Houwer, 2006). Arguably, when it comes to social bias, one of the most critical features of automaticity is unintentionality, perhaps even eclipsing lack of awareness in its importance.

If intent is not a prerequisite for biased responding, then people can perpetuate group-based inequality without wishing to do so—not only indirectly, as when ignorance of structural racism leads to unwitting support for unjust policies (Rucker & Richeson, 2021), but also directly, through biased interpersonal behavior. This possibility, it turns out, is very real: Thanks to implicit measures of bias, there now exists strong evidence that social bias can operate unintentionally. Specifically, someone can lack a goal to respond to a social cue (Bargh, Chaiken, Raymond, & Hymes, 1996; Devine, 1989; Fazio et al., 1986; Greenwald et al., 1998), or even possess a goal to avoid responding to a social cue (Payne et al., 2005), and still respond to it in a biased manner.

This fundamental feature of bias was discovered using sequential priming tasks (Meyer & Schvaneveldt, 1971; Neely, 1976)—a class of implicit measure employed by Devine (1989) to discover unintentional stereotype activation and by Fazio, Jackson, Dunton, and Williams (1995) to discover unintentional forms of evaluative bias. In these tasks, participants see incidental social cues, such as names or pictures of people from different racial groups. The social cues are incidental because they are irrelevant to the primary task goal, which is to respond to nonsocial stimuli, such as by classifying words as positive or negative or by classifying letter strings as words or non-words.

Performance on the focal task is reliably influenced by the incidental social cues. For instance, after seeing a picture of a racial outgroup member, participants are slower to identify positive words as positive, and are likelier to misidentify positive words as negative. Likewise, after seeing a picture of a racial ingroup member, participants are slower to identify negative words as negative, and are likelier to

CONTACT David E. Melnikoff davidemelnikoff@gmail.com Department of Psychology, Northeastern University, Boston, MA 02115, USA.

misidentify negative words as positive. Such influences of social cues on performance appear unintentional; the cues are irrelevant to the focal task, and processing them hinders performance. Subsequent work has confirmed the unintentionally of the effects observed on the sequential priming task by showing that they persist even when participants are incentivized to ignore the social cues (Mann, Cone, Heggeseth, & Ferguson, 2019; Todd, Thiem, & Neel, 2016), and when the social cues are presented subliminally (Cameron, Brown-Iannuzzi, & Payne, 2012; Lowery, Hardin, & Sinclair, 2001; Sinclair, Lowery, Hardin, & Colangelo, 2005).

In short, by quantifying the effects of task-irrelevant social cues on task-relevant performance, sequential priming tasks were used to discover, and are still used to measure, unintentional forms of bias. Most implicit measures of bias used today follow the same basic strategy: From the Implicit Association Test (IAT) to the Affect Misattribution Procedure (AMP), what makes implicit measures of bias implicit is their ability to tap bias that operates unintentionally.

Since revealing the existence of unintentional bias, implicit measures have illuminated much about how this form of bias operates, including insights that advance translational efforts to promote group-based equality. Take the issue of controllability: Is unintentional bias uncontrollable, or can it be stopped or altered by a person's current goal? When unintentional bias was discovered, the prevailing view was that all features of automaticity—unintentional, uncontrollable, unconscious, and efficient—always, or nearly always, cooccur (Kahneman, 2003; Posner & Snyder, 1975; Stanovich & West, 2000). Accordingly, unintentional bias was deemed uncontrollable.

These ideas sparked considerable concern: If bias can be uncontrollable, then, in many cases, bias may be permissible, both morally and legally (Fiske, 1989). Indeed, that people should not be held accountable for outcomes beyond their control is a universally held moral intuition and a core tenet of legal doctrine. A nightmare scenario suddenly seemed plausible: People who egregiously and demonstrably discriminate may evade punishment by claiming their bias was outside of their control (Fiske, 1989).

Fearing this outcome, researchers started using implicit measures of bias to test whether unintentional bias is, in fact, uncontrollable. This work put the nightmare to rest, as it quickly became apparent that unintentional bias is well within people's control (Blair, 2002). For instance, Devine, Plant, Amodio, Harmon-Jones, and Vance (2002) used an implicit measure of bias to demonstrate that unintentional racial bias is dramatically reduced among people who are highly internally motivated, and not externally motivated, to respond without prejudice. Similarly, Moskowitz, Gollwitzer, Wasel, and Schaal (1999) used an implicit measure of bias to show that chronic egalitarian goals can completely eliminate unintentional gender bias. Unintentional bias can also be controlled through the situational formation of if-then plans, known as implementation intentions. For instance, Mendoza, Gollwitzer, and Amodio (2010) manipulated whether or not participants formed an implementation intention of the form, "If I see a person, then I will ignore their race!" Those who did displayed significantly less unintentional pro-White/anti-Black bias.

Unintentional bias can be controlled even without the explicit goal to suppress prejudice—simply wanting to get along with someone is sufficient. In a study by Sinclair et al. (2005), White participants interacted with a White experimenter who either did or not did express anti-racist attitudes. In addition, participants were given a goal to affiliate with the experimenter or not (by having the experimenter behave either likably or rudely). Using a measure of implicit bias, the authors found that unintentional pro-White, anti-Black bias was completely eliminated among participants whose goal was to affiliate with the anti-racist experimenter.

Even seemingly trivial goals, such as a goal to accomplish an arbitrary task, can radically alter unintentional responses to social cues, including under conditions where responding has long been thought to be resistant to any such malleability. Specifically, Melnikoff and Bailey (2018) tested whether unintentional biases in favor of moral over immoral individuals can be eliminated or reversed by task goals (also see Béna, Melnikoff, Mierop, & Corneille, 2022; Melnikoff, Lambert, & Bargh, 2020). These studies constitute a particularly conservative test of the controllability of unintentional bias given that the human preference for morality over immorality is routinely seen as one of the strongest, least malleable, and most pervasive of all known social biases (Fiske, Cuddy, & Glick, 2007; Goodwin, 2015; Goodwin, Piazza, & Rozin, 2014; Landy, Piazza, & Goodwin, 2016).

In the experiments by Melnikoff and Bailey (2018), participants were told that they would play one of two games: one in which moral people are goal-conducive and one in which immoral people are goal-conducive. For instance, a game in which moral people were goal-conducive involved playing the role of defense attorney, and trying to secure a not-guilty verdict by selecting merciful people to serve on the jury; a game in which immoral people were goal-conducive involved playing the role of prosecuting attorney, and trying to secure a guilty verdict by selecting merciless people to serve on the jury. Using implicit measures of bias, the authors found that when immorality was goal-conducive, unintentional pro-morality/anti-immorality biases were completely eliminated, and sometimes even reversed. If an arbitrary task goal, such as winning a game in an anonymous online study, can reverse the unintentional expression of one of the most powerful social biases, it is safe to say that unintentional bias is not just controllable, but highly so.

The controllability of unintentional bias has major implications—moral, legal, and psychological. For one, it confirms that even in the absence of ill intent, bias is not morally permissible. Though moral condemnation is typically reserved for intended harms, it is also considered appropriate for unintended harms that could and should be prevented—that is, harms that are controllable (Malle, Guglielmo, & Monroe, 2014). By the same token, the controllability of unintentional bias implies that perpetrators of unintentional bias may be found liable for discriminatory outcomes in court; people may be held liable for unintended harms resulting from negligence, defined as a failure to prevent harms despite having the obligation and capacity to do so (Banaji, Bhaskar, & Brownstein, 2015; Brownstein, 2016; Fiske, 1989; Oppenheimer, 1993).

In addition to its moral and legal implications, the controllability of unintentional bias has played a major role in revising the more general view of human information processing as involving two fundamentally different types of process: automatic and controlled (Evans, 2003; Kahneman, 2003; Posner & Snyder, 1975; Shiffrin & Schneider, 1977; Stanovich & West, 2000). This so-called dual-process framework cannot account for the existence of bias that is unintentional yet controllable (Bargh, 1989, 1994; Melnikoff & Bargh, 2018); indeed, recent critiques of the framework have leaned heavily on discoveries derived from implicit measures of bias, specifically the finding that intentionality and controllability can dissociate (Ferguson, Mann, & Wojnowicz, 2014; Melnikoff & Bargh, 2018). By showing so clearly that automatic and controlled features of responding to social cues can coexist, implicit measures of bias have played a major role in advancing our understanding not only of bias, but of automaticity more generally.

Ironically, to see the impact of this discovery, one need look no further than GLE's very own critique of implicit measures of bias: In clearly distinguishing unconscious bias from bias characterized by other forms of automaticity, the authors rely on the nuanced understanding of automaticity that implicit measures of bias have played a key role in promoting. It is revealing that implicit measures of bias provided some of the ammunition for GLE's own critique; though this irony in no way undermines the critique, it does reinforce what we take to be the critique's intended interpretation as revealing not the uselessness of implicit measures of bias, but rather the ways in which these otherwise useful tools may have been misapplied or misinterpreted.

Using Implicit Measures to Reveal to Cognitive Architecture of Bias

Implicit measures of bias have many important uses beyond exploring questions of automaticity. They can also help address questions of cognitive architecture: How many cognitive systems produce bias, and what are features of these systems? Here too, today's theoretical landscape would seem all but unrecognizable to the pioneering social cognition researchers of the early 1980s (for more on the history of implicit social cognition, see Kurdi & Banaji, in press-a). This sea change can be attributed, in large part, to work that has leveraged implicit measures of bias.

Research on unintentional bias began at a time when unintentional and intentional processes were thought to emerge from two distinct cognitive systems (Collins & Loftus, 1975; Neely, 1977; Posner & Snyder, 1975). The system considered responsible for intentional processes was thought to perform relatively sophisticated computations capable of implementing "high-level" phenomena such as planning, reasoning, and self-control. In contrast, the system considered responsible for unintentional processes was thought to rely on a simple spreading-of-activation mechanism, limiting it to "low-level" phenomena such as semantic priming and habitual responding.

Naturally, this latter system was deemed responsible for unintentional bias, as it was deemed responsible for all unintentional processes. Under this assumption, unintentional bias occurs when activation spreads across links between conceptual nodes in semantic memory—for instance, between Black People and Violent and White People and Nonviolent. When one node is activated (e.g., upon perception of a Black person), activation unintentionally spreads to the linked node (e.g., Violent), promoting biased responding. The links themselves were considered results of the mind passively registering long-term co-occurrence statistics in the environment, such as repeatedly encountering biased news coverage depicting Black people as more violent than White people.

A critical aspect of the process described above is that it makes no mention of propositional information. In fact, the process is considered effectively blind to any information other than co-occurrence statistics. This idea implies, for instance, that encountering the propositional statement "Black people are not violent" would do nothing to weaken the conceptual connection between Black People and Violent, not even if the proposition was uttered by a trusted source and considered valid by the listener. If anything, the proposition "Black people are not violent" may, ironically, strengthen the link between Black People and Violent because it places these concepts in close proximity. This mode of operation stands in stark contrast to the one attributed to the system deemed responsible for intentional bias, whose relatively sophisticated computations can calibrate cognition and action to propositional content.

The idea that propositional information can alter intentional bias, but not unintentional bias, has important implications both for cognitive architecture and the way in which unwanted social biases can be alleviated. For instance, if unintentional bias reflects exclusively the environmental associations to which a person is exposed, then the only way to escape it is by moving to a desert island or, at the very least, consuming information in a highly selective way. Moreover, moral responsibility for unintentional bias would belong to those who disseminate biased information rather than those unwittingly exposed to it, for anyone exposed to biased information would be doomed to absorb it given the constraints of their cognitive architecture.

To give an illustrative example: In July 2008, at the height of the United States presidential campaign, *The New Yorker* published a cover depicting then-candidate Barack Obama, dressed like Osama bin Laden, fist bumping Michelle Obama, clad in military gear and carrying a weapon, with the American flag burning in the fireplace behind them. Many readers expressed outrage over what they saw as the magazine's pandering to racist stereotypes about Obama being a foreigner, a Muslim, and a criminal. The editor of the magazine, David Remnick, defended the image as well-intentioned satire: “The intention is to satirize not Barack Obama and Michelle Obama, but, in fact, to hold a pretty harsh light up to the rumors, innuendos, lies about the Obamas that have come up—that they are somehow insufficiently patriotic or soft on terrorism” (NPR, 2008).

Relying on dominant theoretical ideas of the time about the nature of unintentional bias, Banaji (2008) chastised Remnick’s response as betraying “absurd naïveté about the basic facts of information transmission.” Specifically, Banaji argued that exposing readers to a visual association between Obama and terrorism creates an inescapable conceptual link between Obama and Terrorist because, “[t]o some part of the cognitive apparatus, that association is for real. Once made, it has a life of its own because of a simple rule of much ordinary thinking: Seeing is believing.” Banaji added that the effects of the visual association would be especially likely to emerge “when conscious cognition isn’t up to policing” such effects, thus expressing some degree of optimism that, although unintentional responding might fall prey to the unwanted association, intentional processes of high-level reasoning would be able to overcome it.

Since 2008, our understanding of the cognitive architecture of bias has changed considerably, thanks in large part to studies relying on implicit measures of bias. Dozens of experiments—enough to merit multiple narrative reviews (Cone, Mann, & Ferguson, 2017; De Houwer, Van Dessel, & Moran, 2020; Kurdi & Dunham, 2020)—have used implicit measures of bias to demonstrate that unintentional bias is, in fact, sensitive to propositional information. In light of this work, new accounts of cognitive architecture have emerged, accounts that reject the view that unintentional bias emerges exclusively, or at least overwhelmingly, from a system sensitive only to non-propositional information (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Some of these new accounts of cognitive architecture posit that intentional and unintentional bias are both subserved by a single system sensitive only to propositional information (De Houwer, 2014); other, less radical accounts suggest that intentional and unintentional bias emerge from distinct but mutually interacting systems, while retaining the core idea that these systems are differentially sensitive to propositional information (Gawronski & Bodenhausen, 2006).

At first blush, these new proposals suggest that the worries expressed by Banaji (2008) over *The New Yorker* cartoon may have been misplaced; those who understood the cartoon’s propositional meaning may have avoided any increase in unintentional bias toward Obama (see Peters & Gawronski, 2011). Yet the results of a recent meta-analysis suggest that this conclusion may be both too simplistic and overly optimistic (Kurdi, Morehouse, & Dunham, 2022). This meta-analysis, relying on 677 effect sizes obtained from close to 25,000 participants across 55 research reports, found that propositional influences on unintentional bias are robust, but not ubiquitous. For example, propositional influences on unintentional bias appear limited to a kind of “critical period”—immediately before and after they are formed, unintentional biases are sensitive to propositional information, but they lose this sensitivity over time, a finding originally reported by Peters and Gawronski (2011). The implication is that viewers of *The New Yorker* cartoon may have avoided unintentional bias toward Obama if they understood the cartoon’s satirical meaning immediately; if they took too long to find the correct interpretation, any unintentional bias induced may have ossified.

The meta-analysis also found that unintentional bias is less sensitive to propositional information when the target of the bias is well-known rather than novel, and when the co-occurrence information itself is negative (e.g., a visual association of Barack Obama with terrorism) rather than positive (e.g., a visual association of Barack Obama with the symbols of the U.S. presidency). As such, although unintentional bias often responds to positional information, such influences appear relatively less potent in consequential situations such as the one discussed by Banaji (2008).

What is more, contrary to the optimistic assessment by Banaji (2008), the meta-analysis provides some reason for pessimism about the power of intentional processing to act as a check on unintentional bias. Specifically, it seems that intentional responses to social cues, like their unintentional counterparts, reflect a mix of propositional and non-propositional information. Why this is the case will be an important avenue for future work. It is conceivable that, just as assumed by early theories, bias can emerge from a simple spreading-of-activation mechanism sensitive only to environmental co-occurrences, except that this mechanism contributes not only to unintentional bias, but intentional bias as well.

Alternatively, or in addition, what looks like sensitivity to non-propositional information may, in fact, reflect misinterpretations of propositional information. Suppose, for instance, that *The New Yorker* cartoon increased intentional or unintentional bias. Would this mean that either of these forms of bias is sensitive to non-propositional information? Not necessarily. It may simply mean that viewers incorrectly inferred that Obama is, in fact, soft on terrorism, and it was this (wrong) propositional reasoning that altered either or both types of bias—not the mere co-occurrence of Obama and terrorist imagery (Kurdi & Dunham, 2021). Such misunderstandings are commonplace, and particularly likely when visual associations (whose propositional content is inherently ambiguous) are presented without verbal commentary, as was done on the cover of *The New Yorker*.

Clearly, many questions about the cognitive architecture of bias remain unanswered. Do unintentional and intentional forms of bias really emerge from different systems, one sensitive to propositional information and one sensitive to non-propositional information? If so, how do these systems interact to allow both types of information to contribute to both types of bias? If not, what single-system account can be offered that accurately predicts when intentional and unintentional forms of bias will dissociate from each other (Kurdi & Dunham, 2020)? Most likely, none of these questions would have been asked, and they would be considerably more difficult to answer, without implicit measures of bias.

Using Implicit Measures of Bias to Tackle Group-Based Inequality

Basic questions of automaticity and the cognitive architecture of social evaluation are inextricably linked to issues of group-based inequality. For instance, as discussed above, the automaticity of social bias bears on who can and cannot be held accountable, morally and legally, for discrimination, and the cognitive architecture of bias bears on whether, and how, social bias can be mitigated. It is no surprise, then, that many insights into the psychological substrates of group-based inequality have come from work employing implicit measures of bias. Much of this work has recently been reviewed elsewhere (Kurdi & Dunham, 2022); as such, here we limit ourselves to a concise summary.

Studies using implicit measures of bias have started revealing the psychological underpinnings of group-based inequality by examining individual-level behavior, regional-level behavior, and even the content of the English corpus. To name just a few examples, relevant work at the individual level has shown that female high school students with higher levels of unintentional math–gender stereotyping do worse in math classes (Steffens, Jelenec, & Noack, 2010), and HR managers with higher levels of unintentional anti-obese stereotyping are less likely to invite overweight applicants for an interview (Agerström & Rooth, 2011; Rooth, 2010). Although the overall meta-analytic effect of unintentional evaluation and stereotyping on downstream behavior is small, much is now understood about the variability of such influences, including conditions under which relatively large effects can emerge (Kurdi, Seitchik, et al., 2019). Moreover, with repeated interactions, even relatively small effects can compound to produce large-scale societal inequalities (Greenwald, Banaji, & Nosek, 2015).

Considerably larger effects are generally observed when implicit measures of bias are correlated with consequential outcomes at the regional, rather than the individual, level (for reviews, see Calanchini, Hehman, Ebert, Wilson, Simon, & Esposito, in press; Hehman, Calanchini, Flake, & Leitner, 2019; Payne, Vuletich, & Lundberg, 2017). Again, just to name a few examples, higher levels of unintentional gender stereotyping have been found to be predictive of larger STEM achievement gaps at the country level (Nosek et al., 2009), and higher levels of unintentional anti-Black evaluation at the level of counties in the United States have been found to be predictive of larger income inequality and lower levels of upward mobility among Black Americans (Chetty, Hendren, Jones, & Porter, 2020).

In addition, it has been shown that regional-level measures of unintentional bias can reflect group-based inequality in situations where the causal link is assumed to point from inequality to bias rather than *vice versa*. For instance, Ofosu, Chambers, Chen, and Hehman (2019) provided evidence for diminished unintentional anti-gay bias at the regional level following the legalization of marriage equality. In a similar vein, Payne, Vuletich, and Brown-Iannuzzi (2019) have demonstrated that the degree of unintentional anti-Black bias in a U.S. county today is highly correlated with the number of enslaved Black individuals in that county 160 years ago. Findings of this sort are many and their numbers are growing.

Finally, remarkable convergence has also been detected between implicit measures of bias and algorithmic measures of bias derived from large amounts of text data. In their seminal study, Caliskan, Bryson, and Narayanan (2017) showed that the semantic associations apparent in a large-scale corpus of the English-language Internet exhibited the same direction, and often even size, as the conceptual content captured by individual-level implicit measures of bias, such as the IAT. In a theoretically driven investigation, Kurdi, Mann, Charlesworth, and Banaji (2019) demonstrated that dissociations between evaluations (e.g., liking one social group more than another) and stereotypes (e.g., considering one social group more competent than another) may uniquely characterize verbal responses captured by survey items, whereas the two constructs seem to converge more readily both on individual-level implicit measures of bias and algorithmic measures of textual bias. A third relevant paper by Charlesworth, Yang, Mann, Kurdi, and Banaji (2021) suggests that the gender stereotypes routinely found on individual-level implicit measures of bias are already present at extremely early ages and reflect input from several different sources, including child–parent conversations, children's books, and audiovisual media.

Further demonstrating the profound connection between theoretically driven social cognition work and questions of group-based inequality, each of these literatures has also provided important insights into and sparked debates about the basic nature of bias. For example, contrary to the expectations of certain dual-process theories, the meta-analysis by Kurdi, Seitchik, et al. (2019) has found that unintentional evaluative bias is equally predictive of relatively controllable and relatively uncontrollable behaviors. The literature on regional-level effects has generated lively discussions about whether unintentional bias is best studied as a feature of individuals, situations, or a combination thereof (Connor & Evers, 2020; Gawronski & Bodenhausen, 2017; Kurdi & Banaji, 2017; Payne et al., 2017, Payne, Vuletich, & Lundberg, 2022). Finally, by revealing a tight coupling between bias on implicit measures, which is relatively unintentional, and bias in verbal behavior, which is relatively intentional, research on word embeddings raises interesting questions about cognitive architecture (see Kurdi & Banaji, in press-b). For example, this convergence may suggest that intentional and unintentional forms of bias may emerge from a common system or distinct but highly interactive systems.

Conclusion

Within a few short decades, implicit measures of bias have catapulted the field of social psychology from not even envisioning the possibility that unintentional forms of bias may exist to grounding unintentional bias in increasingly sophisticated, and ever evolving, accounts of automaticity and cognitive architecture, all with clear translational implications for group-based inequality. Implicit measures of bias are valuable. They are not, however, without limitations. By highlighting one limitation—an apparent inability to tap unconscious bias—GLE have inspired us, and will no doubt inspire many others, to reflect on what it is that implicit measures of bias have to offer, and how the results emerging from such measures should best be interpreted.

At the same time, the target article highlights the importance of further work on the automaticity features of existing implicit measures of bias and of developing new measures that are tailored specifically to index unconscious processes in social evaluation and judgment. We are optimistic that measures of this kind will do for our understanding of unconscious bias what implicit measures have done, and continue to do, for our understanding of unintentional bias, thus revealing a fuller and more accurate picture of how humans automatically relate and respond to their social environments.

Author Note

Benedek Kurdi is a member of the Scientific Advisory Board at Project Implicit, a 501(c)(3) nonprofit organization and international collaborative of researchers who are interested in implicit social cognition.

ORCID

David E. Melnikoff http://orcid.org/0000-0002-7090-9995
Benedek Kurdi http://orcid.org/0000-0001-5000-0584

References

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *The Journal of Applied Psychology*, *96*(4), 790–805. doi:10.1037/a0021594.

Banaji, M. R. (2008, August 1). The science of satire. *The Chronicle of Higher Education*. https://www.chronicle.com/article/the-science-of-satire/?bc_nonce=27j61r3jr5qhjuc37diwib&cid=reg_wall_signup.

Banaji, M. R., Bhaskar, R., & Brownstein, M. (2015). When bias is implicit, how might we think about repairing harm? *Current Opinion in Psychology*, *6*, 183–188. doi:10.1016/j.copsyc.2015.08.017

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York: Guilford 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: Basic processes; Applications* (pp. 1–40). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic attitude activation with a pronunciation task. *Journal of Experimental Social Psychology*, *32*(1), 104–128. doi:10.1006/jesp.1996.0005

Béna, J., Melnikoff, D. E., Mierop, A., & Corneille, O. (2022). Revisiting dissociation hypotheses with a structural fit approach: The case of the prepared reflex framework. *Journal of Experimental Social Psychology*, *100*, 104297. doi:10.1016/j.jesp.2022.104297

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. doi:10.1207/S15327957PSPR0603_8

Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, *7*(4), 765–786. doi:10.1007/s13164-015-0287-7

Calanchini, J., Hehman, E., Ebert, T., Wilson, L., Simon, D., & Esposito, E. (in press). Regional intergroup bias. *Advances in Experimental Social Psychology*, *66*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. doi:10.1126/science.aal4230.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350. doi:10.1177/1088868312440047.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240. doi:10.1177/0956797620963619.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics*, *135*(2), 711–783. doi:10.1093/qje/qjz042

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. doi:10.1037/0033-295X.82.6.407

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in Experimental Social Psychology*, *56*, 131–199. doi:10.1016/bs.aesp.2017.03.001

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. doi:10.1177/1745691620931492.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. doi:10.1111/spc3.12111

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology*, *61*, 127–183. doi:10.1016/bs.aesp.2019.09.004

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. doi:10.1037//0022-3514.56.1.5

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835–848. doi:10.1037//0022-3514.82.5.835

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. doi:10.1016/j.tics.2003.08.012.

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. doi:10.1037//0022-3514.69.6.1013

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

Ferguson, M. J., Mann, T. C., & Wojnowicz, M. T. (2014). Rethinking duality: Criticisms and ways forward. In J. W. Sherman, B. Gawronski & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 578–594). New York: Guilford Press.

Fiske, S. T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. S. Uleman & J. A. Bargh

(Eds.), *Unintended thought* (pp. 253–283). New York: Guilford Press.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. doi:10.1016/j.tics.2006.11.005.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. doi:10.1037/0033-2909.132.5.692

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

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44. doi:10.1177/0963721414550709

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. doi:10.1037/a0034726.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. doi:10.1037/pspa0000016.

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., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. doi:10.1037/pspi0000155.

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. doi:10.1037/a0035028.

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology. General*, *148*(6), 1022–1040. doi:10.1037/xge0000623.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *The American Psychologist*, *58*(9), 697–720. doi:10.1037/0003-066x.58.9.697.

Kurdi, B., & Banaji, M. R. (2017). Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on Payne, Vuletich, and Lundberg. *Psychological Inquiry*, *28*(4), 281–287. doi:10.1080/1047840X.2017.1373555

Kurdi, B., & Banaji, M. R. (in press-a). Implicit person memory: Domain-general and domain-specific processes of learning and change. In G. B. Moskowitz & E. Balcetis (Eds.), *The handbook of impression formation: A social psychological approach*. Milton Park: Taylor & Francis.

Kurdi, B., & Banaji, M. R. (in press-b). Implicit social cognition: A brief (and gentle) introduction. In A. S. Reber & R. Allen (Eds.), *The cognitive unconscious: The first half century*. Oxford: Oxford University Press.

Kurdi, B., & Dunham, Y. (2020). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Social Cognition*, *38*(Supplement), s42–s67. doi:10.1521/soco.2020.38.supp.s42

Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition*, *214*, 104792. doi:10.1016/j.cognition.2021.104792.

Kurdi, B., & Dunham, Y. (2022). What can reading the implicit social cognition literature teach us about implicit social cognition? *Behavioral and Brain Sciences*, *45*, e80. doi:10.1017/S0140525X21000595

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5810. doi:10.1073/pnas.1820240116

Kurdi, B., Morehouse, K. N., & Dunham, Y. (2022). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information [Manuscript submitted for publication]. Department of Psychology, Yale University.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*, *74*(5), 569–586. doi:10.1037/amp0000364.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart. *Personality & Social Psychology Bulletin*, *42*(9), 1272–1290. doi:10.1177/0146167216655984.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*(5), 842–855. doi:10.1037//0022-3514.81.5.842

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

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the Affect Misattribution Procedure. *Journal of Personality and Social Psychology*, *116*(3), 349–374. doi:10.1037/pspa0000146.

Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. immoral traits in others are conditional. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(4), E592–E600. doi:10.1073/pnas.1714945115.

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*(4), 280–293. doi:10.1016/j.tics.2018.02.001.

Melnikoff, D. E., Lambert, R., & Bargh, J. A. (2020). Attitudes as prepared reflexes. *Journal of Experimental Social Psychology*, *88*, 103950. doi:10.1016/j.jesp.2019.103950

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality & Social Psychology Bulletin*, *36*(4), 512–523. doi:10.1177/0146167210362789.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. doi:10.1037/h0031564.

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.

Morris, A., & Kurdi, B. (2022). *Awareness of implicit attitudes: Large-scale investigations of scope and mechanism* [Manuscript in preparation]. Department of Psychology, Harvard University.

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, *77*(1), 167–184. doi:10.1037/0022-3514.77.1.167

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. doi:10.3758/bf03213230.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254. doi:10.1037/0096-3445.106.3.226

Nosek, B. A., Smyth, F. L., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10593–10597. doi:10.1073/pnas.0809921106.

NPR. (2008, July 14). 'New Yorker' editor defends Obama cover. https://www.npr.org/templates/story/story.php?storyId=92529393.

Ofosu, E. K., Chambers, M. K., Chen, J. M., & Hehman, E. (2019). Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(18), 8846–8851. doi:10.1073/pnas.1806000116.

Oppenheimer, D. B. (1993). Negligent discrimination. *University of Pennsylvania Law Review*, *141*(3), 899–972. doi:10.2307/3312446

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.

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, *116*(24), 11693–11696. doi:10.1073/pnas.1818816116

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

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2022). Critique of the bias-of-crowds model simply restates the model: Reply to Connor and Evers (2020). *Perspectives on Psychological Science*, *17*(2), 606–610. doi:10.1177/1745691621997394.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality & Social Psychology Bulletin*, *37*(4), 557–569. doi:10.1177/0146167211400423.

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55–85). Hillsdale: Erlbaum.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534. doi:10.1016/j.labeco.2009.04.005

Rucker, J. M., & Richeson, J. A. (2021). Toward an understanding of structural racism: Implications for criminal justice. *Science*, *374*(6565), 286–290. doi:10.1126/science.abj7779.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. doi:10.1037/0022-3514.91.6.995.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. doi:10.1037/0033-295X.84.2.127

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*(4), 583–592. doi:10.1037/0022-3514.89.4.583.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. doi:10.1207/S15327957PSPR0402_01

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*(5), 645–665. doi:10.1017/s0140525x00003435.

Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math–gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology*, *102*(4), 947–963. doi:10.1037/a0019920

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. doi:10.1207/s15327957pspr0803_1.

Todd, A. R., Thiem, K. C., & Neel, R. (2016). Does seeing faces of young black boys facilitate the identification of threatening stimuli? *Psychological Science*, *27*(3), 384–393. doi:10.1177/0956797615624492