https://doi.org/10.1038/s44159-023-00248-y

The case against implicit bias fatalism

Benedek Kurdi & Eric Mandelbaum

The standard associative account of implicit bias posits that the mind unavoidably mirrors the biased co-occurrences that are present in the environment. The resulting fatalistic view of implicit bias as inevitable and immutable is both scientifically unwarranted and societally counterproductive.

Since the 1980s, implicit bias — including attitudes (for example, positive evaluations of white people and negative evaluations of Black people) and stereotypes (for example, attributing the trait 'professional' to men and 'nurturing' to women) — has been a central topic of inquiry in experimental social psychology. What makes implicit biases implicit is the fact that they are activated automatically in response to social entities (most notably, social group members) and can deviate from consciously endorsed (explicit) views about all humans' inherent equality. The notion of implicit bias has now been popularized to such a degree that it is used routinely and prominently in public discourse, including in presidential debates and US Supreme Court opinions, to explain why societal inequalities persist even though explicit views about the worth and capabilities of social groups have become considerably more egalitarian over time[1].

Understanding the learning and memory processes that give rise to implicit bias is a fundamental question of mental architecture. At the same time, this understanding can also be harnessed to shift implicit biases towards alignment with explicit egalitarian views and thereby to help to curb their pernicious effects on judgements and decisions involving other people in domains including employment, healthcare and criminal justice. Unfortunately, the standard associative view of implicit bias, which has dominated the field since its inception, provides little grounds for optimism that implicit biases can change in meaningful ways. However, emerging evidence calls the standard associative account into question and suggests that a new outlook on implicit bias can inform future research and lead to successful interventions.

The standard associative account

According to the standard associative account of implicit bias, people catch implicit biases in much the same way they catch colds — simply by going about their business in the world[2]. At the core of this view is associationism, or the idea that implicit biases form and shift in response to co-occurrences in the environment. Many such co-occurrences (such as bread and butter or thunder and lightning) are innocuous. Others, such as the overrepresentation of Black Americans in news stories about crime or frequent mentions of male breadwinners and female homemakers in conversations, movies and books[3], have social ramifications. Critically, according to the associative account, people's cognitive responses to these co-occurrences (such as whether one endorses or denies ideas such as 'Black people are dangerous' or 'women should stay in the kitchen') are inconsequential: simply being exposed to co-occurrences automatically creates and cements the corresponding mental associations. Implicit bias is therefore conceptualized as an indelible carbon copy of the biased environment.

This view has both theoretical and practical consequences. First, if implicit bias is structured associatively and is acquired in the way described above, then it will respond only to the long-term co-occurrence statistics of the environment. Consequently, any attempt to change implicit biases experimentally will be futile (with the potential exception of paradigms involving the rote learning of vast numbers of co-occurrences). Indeed, the literature has little to offer in terms of successful experimental attempts at long-term change at the individual or the organizational level, which has (in our view, erroneously) been interpreted as supporting the standard associative account. Second, if implicit biases are an inescapable consequence of how human minds respond to biased social environments, then people (even those genuinely concerned with inequality) have little choice but to respond with complacency[4].

Evidence against the associative account

Three lines of work provide clear evidence against the standard associative account. These lines of work suggest that implicit biases are more malleable than the standard associative account predicts. Moreover, implicit biases are also sensitive to information to which they should not be sensitive if they had the purely associative structure posited by the standard account.

First, implicit biases respond to evidence beyond environmental co-occurrences[5]. For example, implicit attitudes towards a drug are more negative when participants believe that the drug causes rather than prevents a patient's symptoms, even though the drug and symptoms co-occur with equal frequency in both cases. Similarly, implicit attitudes towards a social group tend to be negative if the group is portrayed as responsible for oppressing another group and positive if the group is portrayed as the blameless target of oppression; here again, both groups are equally associated with the act of oppression.

Perhaps most strikingly, the updating of implicit attitudes is sensitive to whether participants make errors in reasoning about evidence involving co-occurrences[6]. Specifically, participants who correctly conclude that the combination of a conditional statement ('If you see a blue square, then X is sincere') and a disambiguating stimulus (a green circle) warrants no valid inferences about X exhibit no change in implicit attitudes towards X. By contrast, participants who commit the error known as 'denying the antecedent' and conclude that X is insincere show implicit attitude updating in line with the error. In short, implicit bias is sensitive to logical considerations[7] and is therefore not purely associative.

Second, implicit attitudes can rapidly update in response to epistemic factors, such as how diagnostic a piece of evidence is. For example, one extremely diagnostic piece of information — that is, information that provides crucial insight into someone's moral character — can reverse implicit attitudes that were formed on the basis of previously encountered co-occurrences. For example, learning

that someone is a child molester can immediately undo the effects of learning that the same person has performed 100 mildly positive behaviours, such as volunteering to tutor disadvantaged students or paying for their parents' anniversary trip[8]. Such updating can endure beyond a single experimental session[9].

Third, implicit attitudes can exhibit substantial change not only in tightly controlled experimental paradigms with questionable generalizability to real-world settings but also in the real world. In fact, relevant research has found massive cultural-level shifts in implicit biases towards consequential social categories: biases based on properties such as sexuality, race and skin tone have moved substantially towards neutrality over the past 15 years[1]. Notably, this change is unlikely to be explained solely by cohort replacement (that is, younger, more progressive generations replacing older, less progressive ones). Instead, despite the recalcitrance of systemic forms of bias[10], broad-based implicit attitude change seems to have occurred in millions of individual minds.

Outlook

Twenty years ago, it may have been reasonable to assume that short experimental interventions and corporate implicit bias training fail to produce long-term change in implicit attitudes because of the inherent intransigence of the associative mental structures from which they emerge. But given the three lines of evidence described above, this explanation seems difficult to defend today.

Rather, to understand why single-dose interventions are bound to fail, it is important to consider the broader societal context in which they are embedded. Specifically, following quick experimental studies or even rare instances of well-designed implicit bias education, people return to their usual social environments, which are replete with reminders of old biases and have little to offer to facilitate the consolidation of counter-attitudinal updating. Implicit biases therefore often bounce back to their baselines not because implicit attitude change is cognitively impossible but rather, in large part, because ecologies are biased[9].

Moving beyond the fatalistic view of implicit bias has important practical implications for bias reduction. Specifically, individuals have ample opportunities to harness the finding that the way in which people interact with and reason about their environments — and not merely the co-occurrences that characterize those environments — matters for the formation and maintenance of implicit biases. When exposed to content that they deem problematic, such as the umpteenth mention of a Black criminal on cable news, people can actively negate that content and even supplement that negation with an affirmation of the opposite, counter-attitudinal statement. Dozens of studies have now shown that self-generated interventions of this kind can have a meaningful effect on implicit bias[5].

A non-fatalistic view of implicit bias also suggests exciting avenues for future research. First, given all the evidence that implicit bias is not inherently unchangeable at the cognitive level, resources should be devoted to studying how long-term changes produced with novel experimental targets (for example, fictitious individuals such as 'Bob' and fictitious social groups such as 'Laapians') can generalize to social groups of consequence. Discovering how to counteract the effects of ubiquitous negative environmental reminders will probably be instrumental to long-term implicit bias mitigation. Second, although long-term changes in implicit attitudes at the cultural level are now well documented[1], a mechanistic understanding of what enables them is lagging. A better grasp of the processes that underlie such change will enable the design of more effective experimental interventions.

Needless to say, even if all implicit bias were magically eliminated, group-based disparities would still persist. The psychological processes that maintain group-based inequality are manifold, and psychological factors are not solely responsible for such inequality[2,10]. Structural interventions — ranging from decision blinding at the organizational level to anti-discrimination protections at the macro-level of society — have a critical role to play[2], along with psychological interventions, including those targeting implicit bias. Given the multifaceted nature of group-based inequality, any search for a silver bullet is destined to fail. However, fatalism about the potential for progress is both dangerous and unwarranted — even when it comes to changing individuals' long-standing implicit biases.

Benedek Kurdi [1] ✉ **& Eric Mandelbaum**[2,3,4]

[1]Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL, USA. [2]Department of Philosophy, Baruch College, New York, NY, USA. [3]Department of Philosophy, CUNY Graduate Center, New York, NY, USA. [4]Department of Psychology, CUNY Graduate Center, New York, NY, USA.

✉e-mail: kurdi@illinois.edu

Published online: 16 October 2023

References

1. Charlesworth, T. E. S. & Banaji, M. R. Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychol. Sci.* **30**, 174–192 (2019).
2. Greenwald, A. G. et al. Implicit-bias remedies: treating discriminatory bias as a public-health problem. *Psychol. Sci. Publ. Int.* **23**, 7–40 (2022).
3. Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B. & Banaji, M. R. Gender stereotypes in natural language: word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol. Sci.* **32**, 218–240 (2021).
4. Daumeyer, N. M., Onyeador, I. N., Brown, X. & Richeson, J. A. Consequences of attributing discrimination to implicit vs. explicit bias. *J. Exp. Soc. Psychol.* **84**, 103812 (2019).
5. Kurdi, B., Morehouse, K. N. & Dunham, Y. How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *J. Pers. Soc. Psychol.* **124**, 1174–1202 (2023).
6. Kurdi, B. & Dunham, Y. Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition* **214**, 104792 (2021).
7. Mandelbaum, E. Attitude, inference, association: on the propositional structure of implicit bias. *Noûs* **50**, 629–658 (2016).
8. Cone, J. & Ferguson, M. J. He did *what*? The role of diagnosticity in revising implicit evaluations. *J. Pers. Soc. Psychol.* **108**, 37–57 (2015).
9. Kurdi, B., Mann, T. C., Axt, J. & Ferguson, M. J. The fragility of implicit attitude updating: the role of cognitive and ecological constraints. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/mwfah (2023).
10. Skinner-Dorkenoo, A. L., George, M., Wages, J. E. III, Sánchez, S. & Perry, S. P. A systemic approach to the psychology of racial bias within individuals and society. *Nat. Rev. Psychol.* **2**, 392–406 (2023).

Competing interests

B.K. is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition. E.M. declares no competing interests.