The Science of Implicit Race Bias: Evidence from the Implicit Association Test

Kirsten N. Morehouse & Mahzarin R. Banaji

Beginning in the mid-1980s, scientific psychology underwent a revolution – the implicit revolution *– that led to the development of methods to capture implicit bias: attitudes, stereotypes, and identities that operate without full conscious awareness or conscious control. This essay focuses on a single notable thread of discoveries from the Race Attitude Implicit Association Test (RA-IAT) by providing 1) the historical origins of the research, 2) signature and replicated empirical results for construct validation, 3) further validation from research in sociocognitive development, neuroscience, and computer science, 4) new validation from robust association between regional levels of race bias and socially significant outcomes, and 5) evidence for both short- and long-term attitude change. As such, the essay provides the first comprehensive repository of research on implicit race bias using the RA-IAT. Together, the evidence lays bare the hollowness of current-day actions to rectify disadvantage experienced by Black Americans at individual, institutional, and societal levels.*

The science of implicit race bias emerged from a puzzle. By the 1980s, laboratory experiments and surveys revealed clear and noteworthy reductions in expressions of racial animus by White Americans toward Black Americans.[1] But on every dimension that determines life's opportunities and outcomes – housing, employment, education, health care, treatment by law and law enforcement – the presence of widespread racial inequality remained. Further, on surveys asking even slightly indirect questions, such as attitudes toward federal support for racial equality in employment, attitudes appeared to have regressed, with 38 percent support in 1964 dropping to 28 percent in 1996.[2] These inconsistencies demanded an answer from science.

In their search for an explanation, experimental psychologists recalled an interesting *dissociation* or disparity in beliefs recorded decades ago. During his travels through the Jim Crow South, Gunnar Myrdal, a Swedish economist engaged by the Carnegie Corporation to conduct a study on interracial relations in America,

https://doi.org/10.1162/daed_a_02047

encountered an unexpected dilemma. The data from surveys and interviews of White Americans confirmed expected expressions of racism. And yet as Myrdal noted, other sentiments from the very same individuals spoke to their uneasy acknowledgment of a disparity between the cherished national ideal of equality and the history of slavery and the realities of racism, even decades after emancipation. These dissonant cognitions, expressed inside quiet homes and noisy factories, struck Myrdal as distinctive enough to serve as the motif for his classic treatise, *An American Dilemma : The Negro Problem and Modern Democracy*.[3]

Four decades later, psychologists responded to receding levels of "old-fashioned racism" by generating theories of "aversive racism" and measures of "modern racism."[4] These ideas emerged as necessary acknowledgment that although race bias persists, modern racism manifests in more indirect and subtle ways than before. Indeed, experimental data emerging in the 1980s further highlighted the presence of automatic race bias in the minds of honest race egalitarians.[5] With accumulating evidence demonstrating that many judgments and decisions could operate outside conscious awareness or control, social psychologists Anthony G. Greenwald and Mahzarin R. Banaji proposed the idea of *implicit bias* and suggested that a tractable measure of implicit cognition was needed.[6] This essay reports on a thread of the development and discoveries of a singularly important test : the Race Attitude Implicit Association Test (hereafter, RA-IAT), a measure designed to capture differential automatic attitudes, such as associations of "good" and "bad" with White and Black Americans.[7]

In 1967, Martin Luther King Jr. gave the keynote address at the annual meeting of the American Psychological Association (APA), only months before his assassination. He seemed to be aware that his audience of largely White Americans was eager to learn how they could contribute to the success of the civil rights movement. But King's speech clearly conveyed his perspective regarding the responsibility of the APA's scholars and clinicians. If they wished to support the movement, they should simply "'tell it like it is.'"[8] This essay is a response to that call from more than fifty years ago, to emphasize the strength and pervasiveness of anti-Black bias today. We tell it like it is, believing that empirical knowledge production is indeed the responsibility of scientists with expertise in psychological and other sciences. However, the responsibility of addressing challenges to the ideal of racial justice sits squarely at the feet of the nation. In fact, it would be ill-advised to expect scientists – who generally lack knowledge of history, law, policy development, organizational behavior, and the modes of societal change – to be primarily responsible for imagining and constructing paths to social change. By telling it like it is, and remaining focused on the evidence itself, this report can, should the will exist, serve as a foothold to move America toward a solution to racial inequality.

History and Definitions

The science of implicit bias is rooted in experimental psychology. At the core of a particular family of measures is the concept of *mental chronometry*: studying the mind by measuring the time course of human information processing.[9] That is, rather than analyzing participants' responses to a question, the critical unit of measurement is the response latency or the time it takes to react to a stimulus. In the 1970s, researchers conducted the first robust studies testing the automaticity of *semantic memory*. These studies indexed the strength of association between two concepts by using precisely timed stimuli and measuring an individual's response latencies on the order of tens of milliseconds.[10] These procedures were soon adapted to test another important dimension of word meaning: *valence*, that is, the *good-bad* or *pleasant-unpleasant* dimension. Evidence soon emerged that, like semantic meaning, word or concept valence could be automatically extracted by relying on response latencies.[11] Today, this result is received wisdom, and evaluative priming is regarded to be a standard method to measure *automatic attitudes*.[12]

This class of experimental procedures captured the attention of psychologists concerned with the limitation of self-report measures of racism: individuals can withhold their true beliefs in favor of more socially desirable responses. Moreover, even if the desire to speak forthrightly is assured, self-report measures are limited because humans have a desire to present a positive view of themselves, not just to others but even to themselves. Finally, even if such concerns about self and social desirability were removed, a great deal of research had demonstrated that access to mental content and process is vastly limited, making the problem less an issue of motivation and more one of inaccessibility.[13] These considerations, especially the latter, led psychologists to adapt mental chronometry to study automatic or implicit forms of bias. Race was a natural domain for exploration because of the inconsistency between conscious values in aspirational documents like the U.S. Constitution and the history of American racism.

A harbinger of the breakthrough to come appeared in a paper by psychologists John F. Dovidio, Nancy Evans, and Richard Tyler.[14] Diverging notably from previous research methods, these researchers sat their subjects before a computer screen on which the category labels "Black" or "White" appeared. After each of these primes, target words that represented positive and negative stereotypes of these groups (such as ambitious, sensitive, stubborn, lazy) appeared on the screen, and subjects were asked to decide rapidly if each stereotypic word could "ever be true" or was "always false" of the group. The results were clear: participants classified words more quickly when positive words followed "White" and when negative words followed "Black" primes, suggesting that the category White was more positive than Black in participants' implicit cognition. Although this method lacked the components that are characteristic of standard measures of implicit

cognition today (the response task still required deliberation), this study pointed toward the potential of nonreactive measurement of race bias.

Social psychologist Patricia Devine's dissertation experiments hammered a second stake into the ground.[15] She subliminally presented words that captured negative Black stereotypes (in the experimental condition) or neutral words (in the control condition) and then requested evaluations of an ambiguously described person. Remarkably, those who were subliminally exposed to Black stereotypes as primes were more likely to view the ambiguously described person as hostile than those in the control condition. Equally remarkable, the degree of race bias on this more automatic measure of stereotypes was similar *regardless* of consciously reported levels of anti-Black prejudice.

Devine's research demonstrated the first classic dissociation between more deliberate or explicit race attitudes and more automatic or implicit race attitudes, and it prompted a shift in thinking about the nature of race bias. If bias were hidden, even to the person who carried it, that would explain how racial animus could decrease on survey measures while bias embedded in individual minds, institutions, and long-standing societal structures persisted. The two were *dissociated*. From a research standpoint, it was clear that to gain access to race bias in all forms, experimental psychologists would need to develop and sharpen measures of implicit race bias.

Several measures of implicit cognition emerged, among them the Implicit Association Test (IAT).[16] The IAT followed in the tradition of its predecessors by relying on a single fundamental idea: when two things become paired in our experience (for instance, *granny* and *cookies*), evoking one (granny) will automatically activate the other (cookies). In the context of race bias, the speed and accuracy with which we associate concepts like *Black* and *White* with attributes like *good* and *bad* provides an estimate of the strength of their mental association, in this case, an implicit attitude.

Today, decades after the first uses of terms such as *implicit bias*, *implicit attitude*, and *implicit stereotype*, these concepts have permeated scientific and scholarly writing as well as the public's consciousness so effectively that they are rarely accompanied by a definition or explanation.[17] The earliest formal definition of implicit cognition reads: "The signature of implicit cognition is that traces of past experience affect some performance, even though the influential earlier experience is not remembered in the usual sense – that is, it is unavailable to self-report or introspection."[18]A more colloquial definition of implicit bias has emerged as "a form of bias that occurs automatically and unintentionally, that nevertheless affects judgments, decisions, and behaviors."[19]

Both definitions are quite general, and wisely so, to be inclusive of any domain under investigation (such as self-perception, health decisions, and financial decisions). However, despite its generality, the greatest empirical attention has been

devoted to one particular family of biases: those that concern attitudes (valence) and stereotypes (beliefs) about *social groups* (such as by age, gender, sexuality, race, ethnicity, social class, religion, or nationality). Among these, the test that has garnered the greatest scientific and public interest is the race test (as seen in the scientific record and from completion rates of the test online, where the RA-IAT outstrips all other tests in public interest).[20] Unsurprisingly, and for the same reasons, some resistance to the science of implicit race bias has also emerged, but such criticisms remain minor (2 percent of thousands of Google Alerts analyzed include any critical commentary).[21]

Scope of the Essay

Although full-fledged research on implicit social cognition began only in the 1990s, thousands of research articles on implicit bias have since been published. In fact, Google Scholar returns over sixty-five thousand results in response to a query of *implicit bias* as of January 2024. This prolificacy, while notable, renders any complete review of the literature impossible. As such, this essay constrains coverage in four ways. First, we report research on implicit race attitudes, setting aside all other social categories (such as gender, age, sexuality, disability) with a focus on construct validity. Second, we highlight research on *attitudes*, setting aside research on race *stereotypes*. Third, we focus almost entirely on a single method, the IAT, because 1) it is the most widely used measure of implicit bias today (the original report by Greenwald, Debbie McGhee, and Jordan L. K. Schwartz has recorded over seventeen thousand citations on Google Scholar as of January 2024), and 2) the online presence and popularity of the RA-IAT at Project Implicit offer an unparalleled source of data to explore implicit race attitudes.[22] Surprisingly, the signature results from this most popular IAT over the last twenty-five years have not been presented in a single location before. We synthesize them here. Fourth and finally, given the mission of *Dædalus* to explore the frontiers of knowledge on issues of public importance, we prioritize coverage of questions about the *nature* of implicit race bias and its interpretation rather than questions of primarily scientific interest, such as the nature of the psychological *processes* underlying implicit bias, like whether the underlying representation is best viewed as associative or propositional in nature.[23]

With these constraints and opportunities in mind, we introduce 1) streams of research from other sciences, notably cognitive development, neuroscience, and computer science, to provide convergent validation for the RA-IAT data; 2) new research providing predictive validity by demonstrating robust covariation between regional RA-IAT and racial disparities in health care, education, business, and treatment by law enforcement; and 3) evidence demonstrating the RA-IAT's malleability at the individual level (change within one person) and population

level (change within the United States). Together, the data offer confidence in the concept of *implicit race bias* for use in two ways: as a foothold to an effort for broad-based programs and procedures to ensure racial equality, and as the basis for teaching about implicit bias in all educational settings, including schools, colleges, and the workplace.

The Race Attitude IAT: Early Discoveries and Signature Results Providing Validation

Evidence of implicit race bias using the IAT first emerged in the mid-1990s from small-scale, highly controlled experiments administered to college students, as was characteristic of research at that time. These initial experiments were important for benchmarking data that would soon arrive from exponentially larger and more diverse internet-based samples. In 1998, Yale University hosted a test of implicit race attitude, the RA-IAT, among a few other IATs, and the site was immediately bombarded with participants. The RA-IAT was immediately the most popular test, and it remains so twenty-five years later. Today, the amount of research conducted and the diversity of empirical results obtained may appear insurmountable to the general reader. Here, we have created the first repository of the basic discoveries and signature results of the RA-IAT in easy-to-access percentages, histograms, and inferential statistics.

Implicit Social Cognition Terminology and IAT Components

The RA-IAT, following the general IAT procedure, consists of items that appear on a computer screen belonging to a pair of target categories (such as *Black* and *White*) and a pair of target attributes (such as *Good* and *Bad*). At the most basic level, the RA-IAT provides an index of implicit race bias by measuring the relative speed (on the order of milliseconds) it takes participants to sort stimuli when White and Good share a response key (and Black and Bad share a different response key), relative to when Black and Good share a response key (and White and Bad share a different response key).[24] The IAT score is captured by the statistic D, which is a measure of effect size, computed by taking the difference between response latencies in the two critical conditions (that is, Black + Good/White + Bad, and Black + Bad/White + Good) and divided by the standard deviation across all blocks of the test.

Uninitiated readers may wish to take the test at https://implicit.harvard.edu/implicit/selectatest.html. Additionally, in Table 1, we provide descriptions and examples of the core terminology of implicit social cognition and the IAT more generally, even though our focus in this essay will remain on the concept of the attitude.

Table 1
Core Terminology of Implicit Social Cognition Theory and the Implicit Association Test

Term	Description	Labels (examples)	Stimuli (examples)
Concept Category	The concept or category of scientific interest: that is, the target object toward which a measure of attitude or stereotype is sought, such as race, gender, age, sexuality	Black, White, Asian, Latinx (race) Male, Female, Nonbinary (gender) Elderly, Young (age)	Photos/pictures to represent the concept (such as faces of Black and White individuals) Names or other words to represent the concept (such as John or Jane to represent gender) Faces or images to represent age
Attribute Category	The psychological process of scientific interest such as attitude, stereotype, identity; the attribute is the category whose strength of association to the concept category is tested	**Attitude:** Good-Bad, Pleasant-Unpleasant **Stereotype:** Strong-Weak, Smart-Dumb, Honest-Lying **Identity:** Me-Not Me, Me-Other	**Good:** Love, peace, joy **Bad**: Devil, awful, failure **Strong:** Powerful, sturdy, robust **Weak:** Fragile, delicate, frail **Me:** Me, Myself, I, Mine **Not Me**: Not Me, They, Them, Other
Attitude	Evaluative or valence dimension	Good-Bad, Pleasant-Unpleasant, Positive-Negative	See "Attribute Category" row for example stimuli
Stereotype	Beliefs about social groups	Strong-Weak, Smart-Dumb, Honest-Lying	See "Attribute Category" row for example stimuli
Identity	Attitudes and beliefs about oneself	Me-Not Me, Me-Other	See "Attribute Category" row for example stimuli

Source: Descriptions and definitions by the authors.

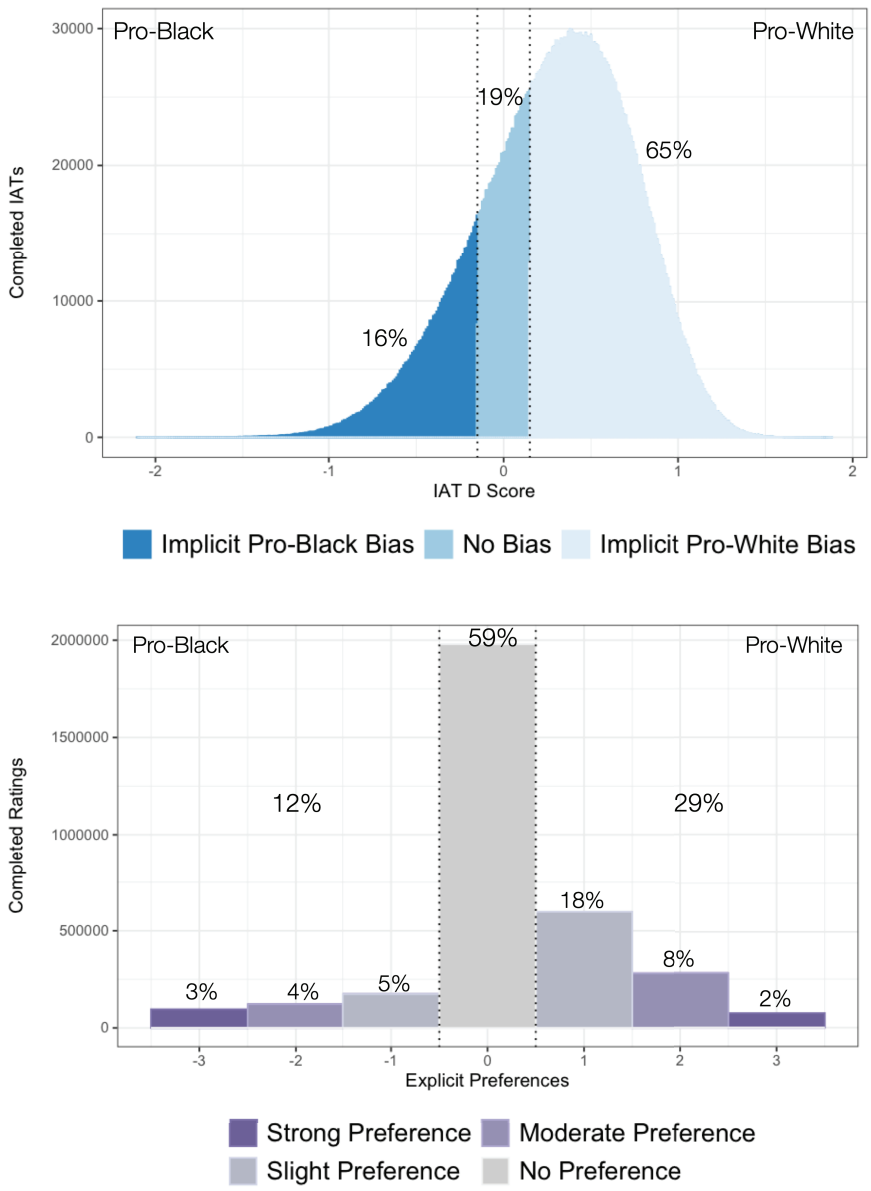
Overall Levels of Explicit and Implicit Race Attitudes and Their Dissociation

An analysis of Project Implicit data from 3.3 million American respondents who completed the RA-IAT across fourteen years (2007–2020) shows robust evidence of implicit race bias: overall, 65 percent of respondents displayed a meaningful association of White with good relative to Black with good ("implicit pro-White bias"), whereas 19 percent of respondents displayed no preference (see Figure 1; for corresponding effect sizes, see Table 2).[25] That is, 2.1 of 3.3 million respondents automatically associated the attribute "Good" (relative to "Bad") more so with White than Black Americans. By contrast, across all fourteen years, only 29 percent of respondents *explicitly* reported a preference for White over Black, and 60 percent of respondents reported equal liking for both groups. As the reader may anticipate, these overall scores are strongly modulated by the social group of the respondent; those data are presented in the next section.

This divergence between mean levels of implicit and explicit race attitudes is striking and bolstered by a dissociation between implicit and explicit race attitudes within a single person. Specifically, modest correlations between implicit and explicit attitudes are typically observed across all participants (for example, $r = 0.30$ [95% CI: 0.308, 0.310]), and even weaker correlations often emerge for Black Americans (see Table 2).[26] Additional support for this dissociation has been derived from latent variable modeling. Unlike variables that can be directly observed or measured (like temperature), latent variables refer to constructs – such as race attitudes – that are inferred indirectly and can possess a degree of measurement error. These latent modeling techniques indicate that implicit and explicit attitudes are related, but distinct. That is, although the latent implicit and explicit attitude variables are correlated ($r = 0.47$), a confirmatory factor analysis suggests that a two-factor solution fits the data better than a single-factor solution with a single latent "attitude" variable.[27] In other words, this technique indicated that implicit and explicit attitudes are related, but psychometrically distinct.

Together, this pattern of data – low levels of explicit race bias but high levels of implicit bias – is considered a key result of implicit intergroup cognition. The data also provide a conceptual replication of Devine's early discovery that implicit race bias can emerge in defiance of stated egalitarian values.[28] However, unlike Devine's work with subliminally presented stimuli, the IAT does not hide its intent; the two racial categories are in full view and the test is announced as one of race bias. Moreover, the IAT components are not shrouded in mystery and completing the task is so simple that even a child can participate. These features contribute to the surprise that often accompanies the IAT: if the task itself is easy, why can I not control my responses?

Figure 1
Distributions of Implicit and Explicit Race Attitudes



IAT D scores range from -2.0 to 2.0, with 0 ± 0.15 serving as the null interval ("Little or No Bias"). Source: Created by the authors using Project Implicit data.

Table 2
Implicit and Explicit Race Attitudes by Participants' Race/Ethnicity

Demographic Subgroup		Implicit		Explicit		
	N	IAT D	Cohen's *d*	Mean	Cohen's *d*	E-I Correlation
Overall	3,325,990	0.29	0.66	0.20	0.19	0.30 [0.308, 0.310]
White	1,881,719	0.36	0.85	0.42	0.51	0.21 [0.211, 0.214]
Asian (East and South)	176,218	0.30	0.70	0.28	0.29	0.27 [0.261, 0.271]
Hispanic	335,780	0.25	0.57	−0.02	−0.02	0.27 [0.275, 0.281]
Multiracial	43,650	0.15	0.34	−0.19	0.62	0.28 [0.268,0.285]
Black	290,837	−0.05	−0.11	−1.07	−0.84	0.17 [0.164, 0.171]

IAT D scores range from −2 to +2, with positive values indicating an implicit pro-White bias. Explicit preferences ranged from −3 ("I strongly prefer African Americans to White Americans") to +3 ("I strongly prefer White Americans to African Americans"). The column "E-I" represents the correlation between IAT D scores and explicit preferences, with 95 percent confidence intervals reported in brackets. Source : Compiled by the authors using Project Implicit data.

Nevertheless, after nearly a century of work based on almost purely explicit measures, these results lay bare the full extent of the challenge we face when confronting the status of race in America today.[29] Recall in Myrdal's interviews during Jim Crow that respondents revealed a disparity between two consciously held beliefs : the American ideal of liberty and equality and America's history of bondage and inequality. In a sense, that conflict is psychologically simple because both cognitions are conscious. By contrast, the dissociation between explicit and implicit race attitudes is especially challenging because implicit attitudes operate largely outside the purview of conscious awareness and control, and therefore may unwittingly produce behaviors that conflict with consciously held values and beliefs.

Explicit and Implicit Race Bias by Racial/Ethnic Group

Among psychology's most ubiquitous results is the demonstration of in-group bias. Irrespective of whether the groups involved are "minimal" (based on a "minimal" preference, such as for the artist Klee over Kandinsky) or real, research has overwhelmingly demonstrated that humans show a preference for their own group relative to the out-group.[30] For example, Japanese Americans and Korean Americans, Yankee and Red Sox fans, and Yale and Harvard students all display clear and symmetric in-group preferences.[31] However, as visualized in Figure 2, the data across White and Black Americans paint a much more complex picture.

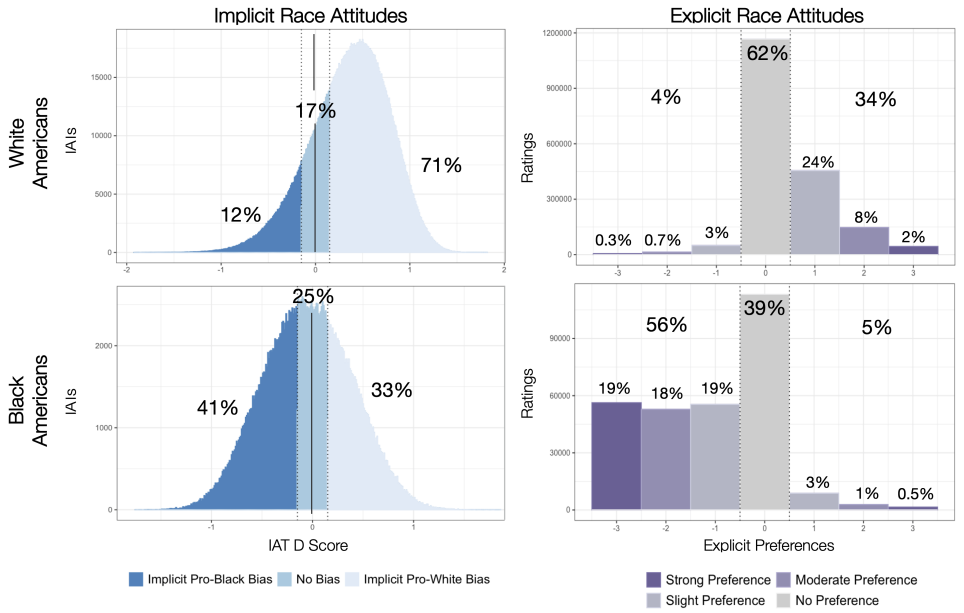
Specifically, 71 percent of White Americans displayed an implicit pro-White bias, whereas only 33 percent of Black Americans displayed an implicit pro-Black bias. These data are in contrast with the robust in-group preferences among Japanese and Korean Americans, Red Sox and Yankee fans, and Yale and Harvard students, in which each group showed an equally robust preference for its own group. This lack of in-group preference among Black Americans is a second signature result and it extends beyond Black Americans to other less advantaged groups. That is, unlike members of socially advantaged groups, who consistently display implicit in-group preferences, members of socially disadvantaged groups typically do not.

On the measure of *explicit* bias, an almost opposite pattern emerges, making these data among the clearest examples of mental dissociation: the lack of consistency between two measures *of the same concept*, within the same mind. Only 34 percent of White Americans displayed an explicit pro-White bias, whereas 56 percent of Black Americans displayed an explicit pro-Black bias. These data highlight the role conscious values play on responses. White Americans, likely being aware of the history of race relations in America, report a far more muted in-group preference. Black Americans, equally likely aware of the history of race relations in America, report an overwhelming in-group preference.

When taken together, the data for White and Black Americans showed a double dissociation. On the one hand, White Americans report little in-group preference on the explicit measure but strong in-group preference on the implicit measure. On the other hand, Black Americans show a strong in-group preference on the explicit measure but no in-group preference on the implicit measure. We regard this result to be sufficiently important that we recommend that it play a role in any discussion of policies to ensure racial equality. Conscious attitudes need not follow such a pattern, but to the extent that attitudes and behavior are driven by both *explicit and implicit* cognition, the balance sheet of intergroup liking shows a striking lack of parity.

Interestingly, when third-party groups are tested (such as Asian Americans taking a White-Black IAT), they consistently show an implicit pro-White bias (see Ta-

Figure 2
Distributions of Implicit and Explicit Race Attitudes for White and Black Americans



IAT D scores range from -2.0 to 2.0, with 0 ± 0.15 serving as the null interval ("Little or No Bias"). Source: Created by the authors using Project Implicit data.

ble 2). That is, rather than associating both out-groups with good equally, third-party respondents display an implicit preference for the socially dominant group. In fact, rivaling the degree of bias among White Americans, 65 percent of Asian Americans and 60 percent of Latinx Americans display an implicit pro-White preference.

Similar patterns also emerge on measures of implicit *stereotyping*. As one example, Morehouse and Banaji, with Keith Maddox, found that White Americans and third-party participants associate human (versus nonhuman attributes like "animal" and "robot") more with their group, whereas nondominant groups (like Black Americans) display no "human = own group" bias.[32] This striking absence of in-group preference in members of disadvantaged groups points to the power of the social standing of groups in society, and has been interpreted to be consistent with system justification tendencies.[33]

Explicit and Implicit Race Bias by Other Demographic Variables

Beyond race/ethnicity, do other demographic variables modulate the strength of implicit race bias? That is, will men and women, liberals and conservatives, or older and younger respondents show different levels of implicit race bias? To test this question, variation across five additional demographic characteristics was examined: religion, level of education, age, gender, and political ideology. Implicit race bias was largely stable across respondents' religious affiliation and level of education. However, differences emerged across age, gender, and political ideology. Implicit pro-White preferences increased with age (each five-year increase translating roughly to a 3 percent increase in IAT D scores), and respondents over age sixty displayed levels of bias that were 15 percent stronger than individuals under age twenty. Further, the incidence of pro-White bias was 20 percent higher among self-identified conservatives relative to self-identified liberals, and 7 percent higher among men relative to women.

These results show how group membership is related to variation in implicit and explicit race attitudes. Later in this essay, we explore another potential determinant of attitude strength – participants' local environment – and the relationship between regional levels of implicit race attitudes and socially significant outcomes (such as lethal use of force by police or health outcomes).

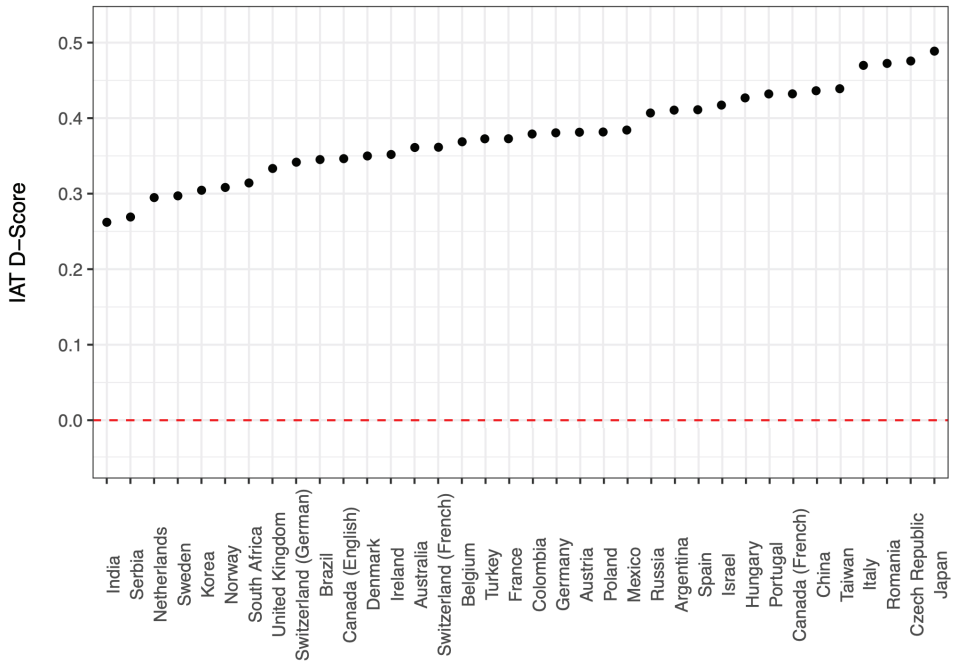
Origins of Implicit Race Bias: Evidence for Developmental Invariance

Over the past twenty years, researchers have gained a new understanding about the surprisingly early precursors of race encoding and race preference in infants and young children. Although far from biological and social maturity, infants and children show evidence of a mind that is already attuned to race but has the capacity to set racial groupings aside, even when attending to other social categories like gender and age, in other situations.[34]

Human groups across the world, as much as they differ by language, culture, preferences, beliefs, and values, are all members of the same species. Is implicit bias a core capacity that unifies us as humans? If we look cross-culturally, a recent analysis of implicit race attitudes from thirty-four countries revealed that an implicit preference for White over Black appears in every country sampled (see Figure 3).[35]

Another way to test whether a particular attitude is fundamental is to observe whether it is present in infants and young children. Our interest here is not in children qua children, but rather in developing minds. Is implicit race bias present even in early stages of cognitive-affective development? The obvious prediction would be that, of course, given the massively different levels of personal experience and knowledge of the culture that children have acquired relative to adults,

Figure 3
Implicit Race Attitudes by Country



Country-level RA-IAT scores expressed in Cohen's *d* effect sizes, with positive effect sizes representing an implicit pro-White bias. For comparison, the average IAT D-Score for the United States for the same period (2009–2019) was 0.30. Source : Adapted from Tessa Charlesworth, Mayan Navon, Yoav Rabinovich, Nicole Lofaro, and Benedek Kurdi, "The Project Implicit International Dataset : Measuring Implicit and Explicit Social Group Attitudes and Stereotypes across 34 Countries (2009–2019)," *Behavioral Research Methods* 55 (3) (2023) : 1413–1440.

implicit race bias should differ based on age. But to the extent that the data show the opposite – similar patterns of implicit race bias in adults and children – we would learn that such biases require little time and experience in a culture to be acquired.

Much has been written about the development of race cognition in infancy.[36] From this work, we know that even infants prefer faces of members of their own group, an effect that likely emerges out of familiarity with their caregivers. For example, three-month-old Ethiopian infants in Ethiopia prefer African over European faces, Ashkenazi babies in Israel prefer European over African faces, and ba-

bies of Ethiopian Jews who have *immigrated* to Israel and have caregivers of both groups show no race preference.[37] Importantly, these preferences are early emerging but not hard-wired; they are absent at birth but present by three months of age.[38] In other words, these data show that the human brain is attuned to features, like race and gender, in the environment that can differentiate between in-group and out-group members.

Work with toddlers has been especially fruitful because the *same method* used to measure implicit race bias in adults could be adapted to measure implicit race bias in children. Specifically, psychologist Andrew Scott Baron and Banaji created a child version of the RA-IAT.[39] Given that children's experiences and knowledge of racial groups vastly differ from adults', the authors expected stark differences in the degree of implicit race bias expressed by children and adults. However, this is not what they found. The surprising result, now replicated many times, is that White six-year-olds, ten-year-olds, and adults show identical levels of implicit race bias.

Notably, and further mirroring the results obtained in adult samples, children's implicit race bias was qualified by social status. By age three, White American children show an in-group preference, whereas Hispanic and Black American children show no in-group preference.[40] This result is remarkable because it teaches us that implicit attitudes are absorbed from the culture and into the minds of even young children. It also challenges the theoretical intuition that implicit attitudes are learned slowly over time. (For further discussion of the development of implicit racial bias, see Andrew N. Meltzoff and Walter S. Gilliam's contribution to this volume.)[41]

Converging Evidence from Neurons and Natural Language

Understanding how the mind works is not for the meek. The Nobel Prize–winning physicist Murray Gell-Mann seemed to understand this when he reputedly said, "Think how hard physics would be if particles could think." Not only are beings who can think the object of our study, but the thinking under consideration is not easily available to their own conscious awareness. As such, building a case for an imperceptible yet consequential bias requires a multipronged, continuous, and iterative process of validation.

There is already deep and broad evidence for the construct validity of the IAT. For example, providing face validity, we know *a priori* that the concept "flower" is more positive than "insect," and the IAT detects this implicit pro-flower preference in most humans.[42] Further evidence can be obtained by studying groups who are known to differ in attitude and observing whether expected differences emerge. Indeed, we have already reported that Black and White Americans show diverging implicit race attitudes, providing additional evidence for construct validity. As a

third route, construct validation has been obtained by demonstrating that findings derived on the IAT are related to (but not redundant with) conceptually similar constructs. Indeed, we have shown that although implicit and explicit race attitudes are modestly correlated, latent variable modeling suggests that a two-factor solution (with "implicit bias" and "explicit bias" as separate latent factors) provided the best fit to the data. In fact, providing discriminant validation, implicit insect-flower attitudes did not hang together with implicit intergroup attitudes.

In the following sections, we will encounter construct validation in several new ways. In particular, we show that methods from other fields (including neuroimaging and word embeddings) also demonstrate evidence of implicit race bias. Moreover, we explore the origins and consequences of implicit race bias to push the engine of construct validity further. Together, these various approaches have not only created a strong foundation for understanding the concept of implicit race bias, but have produced unexpected empirical findings that challenged and refined existing theory.

The Neural Basis of the RA-IAT

When the first pre-IAT measures of implicit attitudes were introduced, little discussion ensued about whether these alien measures should be considered measures of *attitude*.[43] However, when the IAT was introduced, the question of construct validity appeared immediately.[44] It became obvious that measures that directly interrogated the brain, especially those regions that had long been identified as playing a role in emotional learning (such as Pavlovian conditioning), could prove useful if correlations between IAT behavior and brain activation patterns in regions known to be evolved in emotional learning could be observed.

Research with neuroimaging methods like fMRI has long demonstrated that the amygdala, a subcortical brain structure, is involved in the continuous evaluation and integration of sensory information, with a special role for assigning values for valence and intensity.[45] Crucially, neuroscientist Elizabeth A. Phelps and colleagues showed that amygdala activation to Black faces of unknown individuals (relative to White) was significantly correlated with implicit race bias; no such correlation was observed with explicit race bias as measured by the Modern Racism Scale.[46] This suggested that whatever the RA-IAT detects has a core valence component, in line with the idea of "attitudes" as measuring evaluations or the dimension of *positive* and *negative*. A second study suggested that race-based responding is modulated by experience: when the faces of famous and generally liked Black (Denzel Washington) and White (Jerry Seinfeld) faces were used, this activation-implicit bias correlation disappeared. Put differently, this result indicated that familiarity can interrupt this relationship, providing two-pronged convergence.

In the decades that have followed, a plethora of evidence has linked implicit attitudes with neural responses to race-based in-group and out-group faces and more downstream decision-making to test the ability to control default, biased responding.[47] Results of relevance demonstrate that 1) the neural representation of race-based attitudes involve a range of overlapping and interacting brain systems, 2) race-based processing of in-group and out-group faces occurs early in the information-processing sequence starting at one hundred milliseconds upon encountering a face, 3) implicit bias observed in brain activity is malleable and responsive to task demands and context, and 4) individual differences exist in the ability to exert control over biased responses, and this control itself can be initiated without awareness as well as involve both inhibition of unwanted responses and the initiation and application of intentional behavior.[48] Crucially, this last piece of evidence highlights the need for proactive interventions. If bias can creep in, even during early visual processing, then it is unrealistic to expect even well-intentioned individuals to prevent bias from impacting their behavior in the moment. Instead, changes that alter the choice structure and prevent bias from entering the decision-making process are more likely to succeed.

Overall, neuroscientific evidence provided important construct validity for the IAT and its presumed measurement of expressions of value along a good-bad dimension. Moreover, it indicated that implicit race bias converges with multiple levels of information processing from the earliest stages of face detection to judgments of behavior.

Word Embeddings Based on Massive Language Corpora Converge with IAT Data

A long history of research on natural language processing (NLP) coupled with the availability of massive language corpora (such as the Common Crawl and Google Books) have created the opportunity to learn how social groups are represented in language on an unprecedented scale. Specifically, mirroring the logic of the IAT, computer scientist Aylin Caliskan and colleagues used word embeddings – a technique that maps words or phrases to a high-dimensional vector space – to understand the relative associations between targets (such as Black and White people) and attributes (such as Good and Bad).[49] Creating a parallel measure, the Word Embeddings Association Test (WEAT), they performed tests of group-attribute associations in language on a trained dataset of eight hundred and forty billion tokens from the internet. In doing so, they replicated the classic implicit race bias finding: European American names were more likely than African American names to be closer (semantically similar) to pleasant words than to unpleasant words.

These approaches have also enabled researchers to ask questions about human attitudes that are beyond the scope of behavioral tools. Experimental psycholo-

gist Tessa Charlesworth, Caliskan, and Banaji used trained databases of historical texts to demonstrate that attitudinal biases toward racial/ethnic groups have remained stable over the course of two centuries (1800–1999).[50] Moreover, just as neuroimaging data showed convergence between theoretically identified brain regions like the amygdala and the RA-IAT but *not* with explicit race bias, analyses of the biases embedded in language suggest that they are related to IATs but not self-report data.[51] In other words, linguistic patterns represent a reservoir for collectively held or culturally imprinted beliefs.[52]

In fact, recent work indicates that algorithms are even capable of refracting beliefs about racial purity.[53] Specifically, information scientist Robert Wolfe, Caliskan, and Banaji showed that CLIP, an algorithm that relies on both image and text data, has learned the one-drop rule or hypodescent (that is, a legal principle prominent even in the twentieth century that held that a person with just one Black ancestor is to be considered Black).[54] Overall, these findings add to the burgeoning evidence that implicit bias embedded in human minds exists in language and that algorithms trained on these databases will carry, amplify, and even reproduce bias.[55]

Covariation between Regional Implicit Race Bias and Socially Significant Outcomes

A growing number of "audit studies" have demonstrated group-based discrimination in controlled field settings.[56] These studies, typically conducted by economists and sociologists, create highly standardized but naturalistic situations to explore how specific variables (such as race/ethnicity) influence behavior. For example, economist Marianne Bertrand and computation and behavioral scientist Sendhil Mullainathan sent roughly five thousand fictitious résumés to employers in Boston and Chicago.[57] The résumés were identical in all ways except that the applicant's name was either a White- or Black-sounding name. Despite their identical qualifications, résumés with White names received 50 percent more callbacks than résumés with Black names. In another example in the domain of employment, Devah Pager and colleagues demonstrated that, despite having equivalent résumés and being actors trained to respond identically to interview questions, Black applicants were half as likely to receive a callback than White applicants.[58] In fact, in an even more stunning demonstration of race bias, Black applicants were just as likely to receive a callback as White applicants with a felony record. These individual studies mirror a larger trend observed in a meta-analysis : hiring discrimination against African Americans remained stable over a twenty-five-year period (1989–2015).[59]

These audit studies, like the perplexing disconnect between consciously reported prejudice and observed inequalities in society, require an explanation. How is it that the same résumé or qualifications can be evaluated more positively

if they are attributed to a White person? We posit that implicit bias is the most likely explanation. The difficulty was that, until recently, no direct link between measures of implicit bias and large-scale race-based discrimination was available. However, a new line of research, now reaching a substantial number of demonstrations, provides the first persuasive evidence that implicit bias is indeed correlated with racial discrimination on socially significant outcomes (SSB) in domains like employment, health care, education, and law enforcement.[60]

Specifically, a mounting body of research across laboratories and disciplines within the social sciences shows that U.S. regions with stronger implicit race bias (measured by the RA-IAT and stereotype IATs) also have larger Black-White disparities in SSBs. In fact, this research has demonstrated covariation between regional implicit race bias and SSBs in four prominent domains: 1) education (including suspension rates and Black-White gaps in standardized test scores);[61] 2) life and economic opportunity (adoption rates and upward mobility);[62] 3) law enforcement (Black-White disparities in traffic stops and the use of lethal force);[63] and 4) health care (Medicaid spending and Black-White gaps in infant birth weight and preterm births).[64] These studies show that implicit bias, measured at the level of individual minds but aggregated across geographic space, reflects race discrimination that cannot otherwise be explained.

Evidence and Interventions for Implicit Attitude Change: Early Evidence of Malleability

With hindsight, we know that implicit bias is malleable. However, this was not always received knowledge or even expected. In the early years of research on implicit bias using the IAT, many primary investigators believed that implicit bias was intractable.[65] Yet even early work raised the possibility that implicit race attitudes were sensitive to perceivers' motivations, goals, and strategies, as well as contextual manipulations.[66] For example, social psychologist Bernd Wittenbrink and colleagues found that negativity toward Black individuals was lower after watching a movie clip depicting Black Americans in a positive setting (relative to a negative setting).[67] Similarly, social psychologist Brian Lowery and colleagues demonstrated that White Americans displayed lower levels of negativity toward Black individuals in the presence of a Black (rather than White) experimenter.[68]

Extending this work, psychologist Calvin Lai and colleagues conducted an important study exploring the comparative efficacy of seventeen interventions designed to reduce implicit race bias.[69] Although these interventions were roughly five-minutes long and only administered once, eight of the seventeen interventions were effective in reducing implicit race bias. The most effective interventions invoked high self-involvement and/or linked Black people with positivity and White people with negativity.[70] By contrast, interventions that required perspective-

taking, asked participants to consider egalitarian values, or induced a positive emotion were ineffective. When participants' attitudes were tested even a few hours after the intervention, none of the eight previously effective interventions produced a continued reduction in implicit race bias.[71] Of course, this temporary (but not durable) change is to be expected; implicit bias should snap back, rubber band–like, to some stable individual, situational, or broader cultural default. In fact, that single presentations of short interventions can produce *any* change is surprising.

But many "light" interventions, often involving a few counterattitudinal associations or a hypothetical written scenario (a paragraph long) presenting counterattitudinal information, do not show long-term change. To us, the lack of long-term change is hardly surprising given the weakness of the interventions. In fact, in such a case, implementing flimsy interventions and looking for long-term effects is a fool's errand; yet well-intentioned investigators with the hope that a sentence or two should wipe out a lifetime of learning have tried them.

Change at the Societal Level

These laboratory studies provide excellent tests of specific interventions, but they are less equipped to test whether implicit bias has changed over the course of years or decades. As such, the key question of whether long-term change was possible remained. However, recent analyses by Charlesworth and Banaji challenged this idea.[72] Specifically, using time-series modeling, they traced almost three million Americans' implicit race attitudes over the course of fourteen years (2007–2020). Crucially, they found evidence of pervasive change: across all participants, implicit race bias decreased by 26 percent, making it the second fastest changing implicit attitude after sexuality attitudes (anti-gay bias), which saw a dramatic 65 percent reduction during the same period.[73] In fact, if trends continue, implicit race attitudes could first touch neutrality in 2035.

Moreover, this change was not restricted to only certain segments of society (for instance, younger and more liberal participants). Rather, pointing to *widespread societal change*, men and women, older and younger, liberal and conservative, and more- and less-educated participants alike all moved toward neutrality.[74] The only exception was that, unlike White participants, who recorded a 27 percent reduction in implicit bias (IAT D score reduced by 0.11 points), Black participants' implicit attitudes remained relatively stable, only changing 0.03 IAT D score points over the fourteen-year period (see Table 3).

This widespread change is remarkable, especially when one considers that not all implicit biases are changing. For example, implicit anti-elderly, anti-disability, and anti-fat biases remained relatively stable over the fourteen-year period. This change toward some social categories but not others begs an important question: what is the *source* of this change?

Table 3
Change in Implicit Race Attitudes by Participants' Race/Ethnicity

Demographic Subgroup	Start Value (2007)	End Value (2020)	Raw Change	% Change
Overall	0.33	0.24	–0.09	–27
White	0.41	0.30	–0.11	–27
Hispanic	0.29	0.18	–0.11	–38
Asian (East and South)	0.32	0.23	–0.09	–28
Black	–0.09	–0.06	0.03	33

"Start Value" refers to the mean IAT D score recorded in January 2007; "End Value" refers to the mean IAT D score recorded in December 2020. Source: Compiled by the authors using Project Implicit data.

We pose this question because of its relevance to the different claims about how to reduce bias, and where resources earmarked for attitude change should be directed. On the one hand, some researchers and practitioners have criticized a focus on change at the individual level (such as deploying appeals of equality to change individual minds). On the other hand, past interventions targeting structural-level change have not eradicated racial inequalities as expected.[75] In fact, change through laws and acts of Congress, if resisted by individuals, may actually prompt reactance and undo progress.[76]

We noted above that implicit anti-gay bias dropped dramatically (64 percent) between 2007 and 2020. What caused this surprising and especially rapid change? We propose that anti-gay bias may possess unique features that allowed such change. For one, sexuality is more easily concealed than a person's race/ethnicity, gender, age, or weight. But we argue that another explanation warrants further investigation: anti-gay interventions occurred at three levels within the same fourteen-year period.

First, change occurred at the *individual level* as children (and adults of all ages) came out to parents, grandparents, friends, neighbors, and coworkers. Love, already in place, trumped even implicit bias. In other words, the concealable nature

of sexuality forced individuals to reconcile their anti-gay attitudes with their positive feelings toward their loved ones ; this choice architecture was not in place for attitudes about other social groups. Second, change occurred at the *institutional level*. Of course, such change was not adopted everywhere, and some organizations were directly hostile to nonheterosexual employees. However, many institutions, like the U.S. military, enacted policies that affirmed the status of same-sex relationships (such as extending health benefits to same-sex partners) even before the country did. Third, change occurred at the *macro level*. Massachusetts and other states legalized same-sex marriages in the early 2000s, and the Supreme Court of the United States followed suit in 2015. In our estimation, it is rare for interventions at all three levels – individual, institutional, and societal – to occur within a short period of time. To our knowledge, change at all three levels within a short time frame has not eventualized for other social groups.

Implicit race bias exists. Support for its presence is undergirded by evidence from other areas of psychology (cognitive, developmental, neuroscience) as well as other behavioral sciences using quite different methods. New evidence shows that regional implicit bias predicts socially significant outcomes of Black-White disparity along several important dimensions that determine life's opportunities and outcomes. To bring hope, data also reveal that implicit bias is malleable. Overall, these data represent one of many robust streams of scientific evidence available today. Together, they call for a nationwide undertaking for change – at the individual, institutional, and societal levels.

ABOUT THE AUTHORS

Kirsten N. Morehouse is a PhD candidate in psychology at Harvard University. She uses computational and behavioral tools to study when and why humans harbor implicit associations that are in conflict with ground truth data and consciously held beliefs. She has published in such journals as *Proceedings of the National Academy of Sciences*, *Current Research in Ecological and Social Psychology*, and *Journal of Personality and Social Psychology*.

Mahzarin R. Banaji, a Fellow of the American Academy since 2008, is the Richard Clarke Cabot Professor of Social Ethics in the Department of Psychology and the first Carol K. Pforzheimer Professor at the Radcliffe Institute for Advanced Study at Harvard University ; and the George A. and Helen Dunham Cowan Chair in Human Dynamics at the Santa Fe Institute. She is the author of *Blindspot : Hidden Biases of Good People* (with Anthony G. Greenwald, 2013).

ENDNOTES

[1] Howard Schuman, Charlotte Steeh, and Lawrence Bobo, *Racial Attitudes in America : Trends and Interpretations* (Cambridge, Mass. : Harvard University Press, 1985).

[2] Howard Schuman, Charlotte Steeh, Lawrence D. Bobo, and Maria Krysan, *Racial Attitudes in America : Trends and Interpretations*, rev. ed. (Cambridge, Mass. : Harvard University Press, 1997).

[3] Gunnar Myrdal, *An American Dilemma : The Negro Problem and Modern Democracy*, volumes 1 and 2 (Oxford : Harper, 1944).

[4] For aversive racism, see John F. Dovidio and Samuel L. Gaertner, "Prejudice, Discrimination, and Racism : Historical Trends and Contemporary Approaches," in *Prejudice, Discrimination, and Racism*, ed. John F. Dovidio and Samuel L. Gaertner (San Diego : Academic Press, 1986), 1–34. For so-called modern racism, see John B. McConahay, "Modern Racism, Ambivalence, and the Modern Racism Scale," in ibid.

[5] Patricia G. Devine, "Stereotypes and Prejudice : Their Automatic and Controlled Components," *Journal of Personality and Social Psychology* 56 (1989) : 5–18, https://doi.org/10.1037/0022-3514.56.1.5.

[6] Anthony G. Greenwald and Mahzarin R. Banaji, "Implicit Social Cognition : Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* 102 (1) (1995) : 4.

[7] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition : The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998) : 1464–1480, https://doi.org/10.1037/0022-3514.74.6.1464.

[8] "King's Challenge to the Nation's Social Scientists," *The APA Monitor* 30 (1) (1999), https://www.apa.org/topics/equity-diversity-inclusion/martin-luther-king-jr-challenge.

[9] R. Duncan Luce, *Response Times : Their Role in Inferring Elementary Mental Organization* (New York : Oxford University Press, 1986) ; and Michael I. Posner, *Chronometric Explorations of Mind* (Oxford : Lawrence Erlbaum, 1978).

[10] David E. Meyer and Roger W. Schvaneveldt, "Facilitation in Recognizing Pairs of Words : Evidence of a Dependence between Retrieval Operations," *Journal of Experimental Psychology* 90 (1971) : 227–234, https://doi.org/10.1037/h0031564 ; and James H. Neely, "Semantic Priming and Retrieval from Lexical Memory : Roles of Inhibitionless Spreading Activation and Limited-Capacity Attention," *Journal of Experimental Psychology : General* 106 (3) (1977) : 226–254, https://doi.org/10.1037/0096-3445.106.3.226.

[11] Russell H. Fazio, David M. Sanbonmatsu, Martha Powell, and Frank R. Kardes, "On the Automatic Activation of Attitudes," *Journal of Personality and Social Psychology* 50 (1986) : 229–238, https://doi.org/10.1037/0022-3514.50.2.229.

[12] For a fuller treatment of the "implicit revolution," see Anthony G. Greenwald and Mahzarin R. Banaji, "The Implicit Revolution : Reconceiving the Relation between Conscious and Unconscious," *American Psychologist* 72 (9) (2017) : 861–871, https://doi.org/10.1037/amp0000238.

[13] Greenwald and Banaji, "Implicit Social Cognition" ; and Richard E. Nisbett and Timothy D. Wilson, "Telling More than We Can Know : Verbal Reports on Mental Processes," *Psychological Review* 84 (1977) : 231–259, https://doi.org/10.1037/0033-295X.84.3.231.

[14] John F. Dovidio, Nancy Evans, and Richard B. Tyler, "Racial Stereotypes : The Contents of Their Cognitive Representations," *Journal of Experimental Social Psychology* 22 (1) (1986) : 22–37, https ://doi.org/10.1016/0022-1031(86)90039-9.

[15] Devine, "Stereotypes and Prejudice."

[16] For comprehensive reviews of measures of implicit cognition, see Bertram Gawronski and Jan De Houwer, "Implicit Measures in Social and Personality Psychology," in *Handbook of Research Methods in Social and Personality Psychology*, ed. Harry T. Reis and Charles M. Judd (Cambridge : Cambridge University Press, 2014), 283–310 ; Brian A. Nosek, Carlee Beth Hawkins, and Rebecca S. Frazier, "Implicit Social Cognition : From Measures to Mechanisms," *Trends in Cognitive Sciences* 15 (4) (2011) : 152–159, https ://doi.org/10.1016/j .tics.2011.01.005 ; and Bertram Gawronski, "Automaticity and Implicit Measures," *Handbook of Research Methods in Social and Personality Psychology*, ed. Reis and Judd.

[17] Mahzarin R. Banaji, Curtis Hardin, and Alexander J. Rothman, "Implicit Stereotyping in Person Judgment," *Journal of Personality and Social Psychology* 65 (1993) : 272–281, https ://doi.org/10.1037/0022-3514.65.2.272 ; and Greenwald and Banaji, "Implicit Social Cognition."

[18] Greenwald and Banaji, "Implicit Social Cognition," 4–5.

[19] "Implicit Bias," National Institutes of Health, https ://web.archive.org/web/2022071 6115620/https ://diversity.nih.gov/sociocultural-factors/implicit-bias (accessed January 26, 2024).

[20] For an analysis of Google alerts on "implicit bias," see Kirsten N. Morehouse, Swathi Kella, and Mahzarin R. Banaji, "Implicit Bias in the Public Eye : Using Google Alerts to Determine Public Sentiment" (in preparation).

[21] Jennifer L. Howell and Kate A. Ratliff, "Not Your Average Bigot : The Better-than-Average Effect and Defensive Responding to Implicit Association Test Feedback," *British Journal of Social Psychology* 56 (1) (2017) : 125–145, https ://doi.org/10.1111/bjso.12168 ; and Alexander M. Czopp, Margo J. Monteith, and Aimee Y. Mark, "Standing up for a Change : Reducing Bias through Interpersonal Confrontation," *Journal of Personality and Social Psychology* 90 (5) (2006) : 784–803, https ://doi.org/10.1037/0022-3514.90.5.784.

[22] Greenwald, McGhee, and Schwartz, "Measuring Individual Differences in Implicit Cognition." As of May 2023, over thirty million completed IATs have been sampled and over seventy million tests have been at least partially sampled on Project Implicit. See Project Implicit, https ://implicit.harvard.edu (accessed May 1, 2023).

[23] For more on the psychological processes, see Benedek Kurdi, Kirsten N. Morehouse, and Yarrow Dunham, "How Do Explicit and Implicit Evaluations Shift ? A Preregistered Meta-Analysis of the Effects of Co-Occurrence and Relational Information," *Journal of Personality and Social Psychology* 124 (6) (2022), https ://doi.org/10.1037/pspa0000329 ; and Benedek Kurdi and Mahzarin R. Banaji, "Implicit Person Memory : Domain-General and Domain-Specific Processes of Learning and Change," PsyArXiv, October 18, 2021, last edited November 18, 2021, https ://doi.org/10.31234/osf.io/hqnfy.

[24] For a detailed review of the IAT, see Kate A. Ratliff and Colin Tucker Smith, "The Implicit Association Test," *Dædalus* 153 (1) (Winter 2024) : 51–64, https ://www.amacad.org/pub lication/implicit-association-test.

[25] Standard interpretations regard 0 ± 0.15 as the null (no bias) interval. When using any deviation away from zero as the cutoff, 75 percent of respondents displayed an implicit

White + Good/Black + Bad association. Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: IV. Change and Stability from 2007 to 2020," *Psychological Science* 33 (9) (2022), https://doi.org/10.1177/09567976221084257.

[26] Brian A. Nosek, Frederick L. Smyth, Jeffrey J. Hansen, et al., "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes," *European Review of Social Psychology* 18 (1) (2007): 36–88, https://doi.org/10.1080/10463280701489053.

[27] William A. Cunningham, John B. Nezlek, and Mahzarin R. Banaji, "Implicit and Explicit Ethnocentrism: Revisiting the Ideologies of Prejudice," *Personality and Social Psychology Bulletin* 30 (10) (2004): 1332–1346, https://doi.org/10.1177/0146167204264654.

[28] Devine, "Stereotypes and Prejudice."

[29] Mahzarin R. Banaji and Anthony G. Greenwald, *Blindspot: Hidden Biases of Good People* (New York: Delacorte Press, 2013).

[30] Henri Tajfel, Michael Billig, Robert P. Bundy, and Claude Flament, "Social Categorization and Intergroup Behaviour," *European Journal of Social Psychology* 1 (2) (1971): 149–178, https://doi.org/10.1002/ejsp.2420010202.

[31] Steven A. Lehr, Meghan L. Ferreira, and Mahzarin R. Banaji, "When Outgroup Negativity Trumps Ingroup Positivity: Fans of the Boston Red Sox and New York Yankees Place Greater Value on Rival Losses than Own-Team Gains," *Group Processes & Intergroup Relations* 22 (1) (2019): 26–42, https://doi.org/10.1177/1368430217712834; Kristin A. Lane, Jason P. Mitchell, and Mahzarin R. Banaji, "Me and My Group: Cultural Status Can Disrupt Cognitive Consistency," *Social Cognition* 23 (4) (2005): 353–386, https://doi.org/10.1521/soco.2005.23.4.353; and Greenwald, McGhee, and Schwartz, "Measuring Individual Differences in Implicit Cognition."

[32] Kirsten N. Morehouse, Keith Maddox, and Mahzarin R. Banaji, "All Human Social Groups Are Human, but Some Are More Human than Others: A Comprehensive Investigation of the Implicit Association of 'Human' to U.S. Racial/Ethnic Groups," *Proceedings of the National Academy of Sciences* 120 (22) (2023): e2300995120, https://doi.org/10.1073/pnas.2300995120.

[33] John T. Jost, "A Quarter Century of System Justification Theory: Questions, Answers, Criticisms, and Societal Applications," *British Journal of Social Psychology* 58 (2) (2019): 263–314, https://doi.org/10.1111/bjso.12297; John T. Jost, Mahzarin R. Banaji, and Brian A. Nosek, "A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo," *Political Psychology* 25 (6) (2004): 881–919, https://doi.org/10.1111/j.1467-9221.2004.00402.x; and John T. Jost and Mahzarin R. Banaji, "The Role of Stereotyping in System-Justification and the Production of False Consciousness," *British Journal of Social Psychology* 33 (1) (1994): 1–27, https://doi.org/10.1111/j.2044-8309.1994.tb01008.x.

[34] For a review, see Tessa Charlesworth and Mahzarin R. Banaji, "The Development of Social Group Cognition in Infancy and Childhood," in *The Oxford Handbook of Social Cognition*, 2nd edition, ed. Donal E. Carlston, K. Johnson, and Kurt Hugenberg (Oxford: Oxford University Press, in press).

[35] Tessa Charlesworth, Mayan Navon, Yoav Rabinovich, Nicole Lofaro, and Benedek Kurdi, "The Project Implicit International Dataset: Measuring Implicit and Explicit Social Group Attitudes and Stereotypes Across 34 Countries (2009–2019)," PsyArXiv, December 11, 2021, last edited March 21, 2022, https://doi.org/10.31234/osf.io/sr5qv.

[36] For a review, see Charlesworth and Banaji, "The Development of Social Group Cognition in Infancy and Childhood." See also Talee Ziv and Mahzarin R. Banaji, "Representations of Social Groups in the Early Years of Life," in *The SAGE Handbook of Social Cognition*, ed. Susan Fiske and C. Macrae (London : SAGE Publications, 2012), 372–389, https://doi.org/10.4135/9781446247631.n19.

[37] Yair Bar-Haim, Talee Ziv, Dominique Lamy, and Richard M. Hodes, "Nature and Nurture in Own-Race Face Processing," *Psychological Science* 17 (2) (2006) : 159–163, https://doi.org/10.1111/j.1467-9280.2006.01679.x.

[38] David J. Kelly, Paul C. Quinn, Alan M. Slater, et al., "Three-Month-Olds, but Not Newborns, Prefer Own-Race Faces," *Developmental Science* 8 (6) (2005) : F31–F36, https://doi.org/10.1111/j.1467-7687.2005.0434a.x.

[39] Andrew Scott Baron and Mahzarin R. Banaji, "The Development of Implicit Attitudes : Evidence of Race Evaluations from Ages 6 and 10 and Adulthood," *Psychological Science* 17 (1) (2006) : 53–58, https://doi.org/10.1111/j.1467-9280.2005.01664.x.

[40] Yarrow Dunham, Andrew Scott Baron, and Mahzarin R. Banaji, "Children and Social Groups : A Developmental Analysis of Implicit Consistency in Hispanic Americans," *Self and Identity* 6 (2–3) (2007) : 238–255, https://doi.org/10.1080/15298860601115344.

[41] For a further discussion of the development of implicit race bias, see Andrew N. Meltzoff and Walter S. Gilliam, "Young Children & Implicit Racial Biases," *Dædalus* 153 (1) (Winter 2024) : 65–83, https://www.amacad.org/publication/young-children-implicit-racial-biases.

[42] To take a flower-insect IAT, visit https://outsmartingimplicitbias.org/module/iat.

[43] Fazio, Sanbonmatsu, Powell, and Kardes, "On the Automatic Activation of Attitudes."

[44] Russell H. Fazio (in a personal communication, May 1, 2023) confirmed the easy acceptance of results from semantic priming methods that demonstrated automatic attitudes. The reason the IAT was held to higher standards is likely because its chosen attitude objects were not nonsocial entities like *clouds* and *pizza* but rather social categories like race, gender, sexuality, and age. It is likely that discovery of bias on these topics was simply less palatable, including to psychologists who were not familiar with the research tradition on implicit memory from which these measures were derived.

[45] Joseph E. LeDoux, "Emotion and the Amygdala," in *The Amygdala : Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (New York : Wiley-Liss, 1992), 339–351 ; and Goran Šimić, Mladenka Tkalčić, Vana Vukić, et al., "Understanding Emotions : Origins and Roles of the Amygdala," *Biomolecules* 11 (6) (2021), https://pubmed.ncbi.nlm.nih.gov/34072960.2023.

[46] Elizabeth A. Phelps, Kevin J. O'Connor, William A. Cunningham, et al., "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation," *Journal of Cognitive Neuroscience* 12 (5) (2000) : 729–738, https://doi.org/10.1162/089892900562552.

[47] For reviews, see David M. Amodio and Mina Cikara, "The Social Neuroscience of Prejudice," *Annual Review of Psychology* 72 (1) (2021) : 439–469, https://doi.org/10.1146/annurev-psych-010419-050928 ; Inga K. Rösler and David M. Amodio, "Neural Basis of Prejudice and Prejudice Reduction," *Biological Psychiatry : Cognitive Neuroscience and Neuroimaging* 7 (12) (2022) : 1200–1208, https://doi.org/10.1016/j.bpsc.2022.10.008 ; Jennifer T. Kubota, Mahzarin R. Banaji, and Elizabeth A. Phelps, "The Neuroscience of Race," *Nature Neuroscience* 15 (7) (2012) : 940–948, https://doi.org/10.1038/nn.3136 ; Pascal Mo-

lenberghs, "The Neuroscience of In-Group Bias," *Neuroscience & Biobehavioral Reviews* 37 (8) (2013): 1530–1536, https://doi.org/10.1016/j.neubiorev.2013.06.002; and Jennifer T. Kubota, "Uncovering Implicit Racial Bias in the Brain: The Past, Present & Future," *Dædalus* 153 (1) (Winter 2024): 84–105, https://www.amacad.org/publication/uncovering-implicit-racial-bias-brain-past-present-future.

[48] Amodio and Cikara, "The Social Neuroscience of Prejudice"; and Rösler and Amodio, "Neural Basis of Prejudice and Prejudice Reduction."

[49] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science* 356 (6334) (2017): 183–186, https://doi.org/10.1126/science.aal4230.

[50] The top ten traits associated with White (versus Black): critical, polite, hostile, decisive, friendly, diplomatic, understanding, philosophical, able, and belligerent. The top ten traits associated with Black (versus White): earthy, lonely, cruel, sensual, lifeless, deceitful, helpless, rebellious, meek, and lazy. Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji, "Historical Representations of Social Groups across 200 Years of Word Embeddings from Google Books," *Proceedings of the National Academy of Sciences* 119 (28) (2022): e2121798119, https://doi.org/10.1073/pnas.2121798119.

[51] Sudeep Bhatia and Lukasz Walasek, "Predicting Implicit Attitudes with Natural Language Data," *Proceedings of the National Academy of Sciences* 120 (25) (2023): e2220726120, https://doi.org/10.1073/pnas.2220726120.

[52] For an exploration of gender biases embedded in internet texts, see Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, et al., "Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (New York: Association for Computing Machinery, 2022), 156–170, https://doi.org/10.1145/3514094.3534162.

[53] Arnold K. Ho, Jim Sidanius, Daniel T. Levin, et al., "Evidence for Hypodescent and Racial Hierarchy in the Categorization and Perception of Biracial Individuals," *Journal of Personality and Social Psychology* 100 (3) (2011): 492–506, https://doi.org/10.1037/a0021562.

[54] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan, "Evidence for Hypodescent in Visual Semantic AI," in *2022 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2022), 1293–1304, https://doi.org/10.1145/3531146.3533185.

[55] See also Darren Walker, "Deprogramming Implicit Bias: The Case for Public Interest Technology," *Dædalus* 153 (1) (Winter 2024): 268–275, https://www.amacad.org/publication/deprogramming-implicit-bias-case-public-interest-technology; and Alice Xiang, "Mirror, Mirror, on the Wall, Who's the Fairest of Them All?" *Dædalus* 153 (1) (Winter 2024): 250–267, https://www.amacad.org/publication/mirror-mirror-wall-whos-fairest-them-all.

[56] For reviews, see S. Michael Gaddis, "An Introduction to Audit Studies in the Social Sciences," in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis (Cham: Springer International Publishing, 2018), 3–44, https://doi.org/10.1007/978-3-319-71153-9_1; and S. Michael Gaddis, "Understanding the 'How' and 'Why' Aspects of Racial/Ethnic Discrimination: A Multi-Method Approach to Audit Studies," SSRN, July 25, 2019, https://doi.org/10.2139/ssrn.3426846.

[57] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94 (4) (2004): 991–1013, https://doi.org/10.1257/0002828042002561.

[58] Devah Pager, Bruce Western, and Bart Bonikowski, "Discrimination in a Low-Wage Labor Market: A Field Experiment," *American Sociological Review* 74 (5) (2009): 777–799, https://doi.org/10.1177/000312240907400505.

[59] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen, "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time," *Proceedings of the National Academy of Sciences* 114 (41) (2017): 10870–10875, https://doi.org/10.1073/pnas.1706255114.

[60] For a review, see Tessa E.S. Charlesworth and Mahzarin R. Banaji, "The Relationship of Implicit Social Cognition and Discriminatory Behavior," prepublication chapter to appear in *Handbook of Economics of Discrimination and Affirmative Action*, ed. Ashwini Deshpande, https://tessaescharlesworth.files.wordpress.com/2021/05/charlesworth_econ-handbook_final.pdf. For a discussion of the practical significance of these relationships, see Jerry Kang, "Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally," *Dædalus* 153 (1) (Winter 2024): 193–212, https://www.amacad.org/publication/little-things-matter-lot-significance-implicit-bias-practically-legally; and Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-cognitive-manifestation-systemic-racism.

[61] Travis Riddle and Stacey Sinclair, "Racial Disparities in School-Based Disciplinary Actions Are Associated with County-Level Rates of Racial Bias," *Proceedings of the National Academy of Sciences* 116 (17) (2019): 8255–8260, https://doi.org/10.1073/pnas.1808307116; and Mark J. Chin, David M. Quinn, Tasminda K. Dhaliwal, and Virginia S. Lovison, "Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes," *Educational Researcher* 49 (8) (2020): 566–578, https://doi.org/10.3102/0013189X20937240.

[62] Sarah Beth Bell, Rachel Farr, Eugene Ofosuc, et al., "Implicit Bias Predicts Less Willingness and Less Frequent Adoption of Black Children More than Explicit Bias," *The Journal of Social Psychology* 163 (4) (2023): 554–565, https://doi.org/10.1080/00224545.2021.1975619; and Raj Chetty, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter, "Race and Economic Opportunity in the United States: An Intergenerational Perspective," *The Quarterly Journal of Economics* 135 (2) (2020): 711–783, https://doi.org/10.1093/qje/qjz042.

[63] B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi, "Historical Roots of Implicit Bias in Slavery," *Proceedings of the National Academy of Sciences* 116 (24) (2019): 11693–11698, https://doi.org/10.1073/pnas.1818816116; and Eric Hehman, Jessica K. Flake, and Jimmy Calanchini, "Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents," *Social Psychological and Personality Science* 9 (4) (2018): 393–401, https://doi.org/10.1177/1948550617711229.

[64] Jordan B. Leitner, Eric Hehman, and Lonnie R. Snowden, "States Higher in Racial Bias Spend Less on Disabled Medicaid Enrollees," *Social Science & Medicine* 208 (2018): 150–157, https://doi.org/10.1016/j.socscimed.2018.01.013; and Jacob Orchard and Joseph Price, "County-Level Racial Prejudice and the Black-White Gap in Infant Health Outcomes," *Social Science & Medicine* 181 (2017): 191–198, https://doi.org/10.1016/j.socscimed.2017.03.036.

[65] Mahzarin R. Banaji, "The Opposite of a Great Truth Is Also True: Homage of Koan #7," in *Perspectivism in Social Psychology: The Yin and Yang of Scientific Progress* (Washington, D.C.: American Psychological Association, 2004), 127–140, https://doi.org/10.1037/10750-010.

[66] For a review, see Irene V. Blair, "The Malleability of Automatic Stereotypes and Prejudice," *Personality and Social Psychology Review* 6 (3) (2002): 242–261, https://doi.org/10.1207/S15327957PSPR0603_8.

[67] Bernd Wittenbrink, Charles M. Judd, and Bernadette Park, "Evaluative versus Conceptual Judgments in Automatic Stereotyping and Prejudice," *Journal of Experimental Social Psychology* 37 (3) (2001): 244–252, https://doi.org/10.1006/jesp.2000.1456.

[68] Brian S. Lowery, Curtis D. Hardin, and Stacey Sinclair, "Social Influence Effects on Automatic Racial Prejudice," *Journal of Personality and Social Psychology* 81 (2001): 842–855, https://doi.org/10.1037/0022-3514.81.5.842.

[69] Calvin K. Lai, Maddalena Marini, Steven A. Lehr, et al., "Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions," *Journal of Experimental Psychology: General* 143 (4) (2014): 1765–1785, https://doi.org/10.1037/a0036260.

[70] Past research suggests that exposure to only positive Black figures may be less effective at changing implicit racial attitudes than exposure to both positive Black and negative White exemplars. Jennifer A. Joy-Gaba and Brian A. Nosek, "The Surprisingly Limited Malleability of Implicit Racial Evaluations," *Social Psychology* 41 (3) (2010): 137–146, https://doi.org/10.1027/1864-9335/a000020.

[71] Calvin K. Lai, Allison L. Skinner, Erin Cooley, et al., "Reducing Implicit Racial Preferences: II. Intervention Effectiveness across Time," *Journal of Experimental Psychology: General* 145 (8) (2016): 1001–1016, https://doi.org/10.1037/xge0000179.

[72] Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016," *Psychological Science* 30 (2) (2019): 174–192, https://doi.org/10.1177/0956797618813087; Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: IV. Change and Stability From 2007 to 2020," *Psychological Science* 33 (9) (2022), https://doi.org/10.1177/09567976221084257; Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes II. Long-Term Change and Stability, Regardless of Group Membership," *American Psychologist* 76 (6) (2021): 851–869, https://doi.org/10.1037/amp0000810; and Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Stereotypes III: Long-Term Change in Gender Stereotypes," *Social Psychological and Personality Science* 13 (1) (2022): 14–26, https://doi.org/10.1177/1948550620988425.

[73] Explicit race attitudes recorded a 98 percent reduction, shifting from a "'slight' preference for White Americans over Black Americans" to neutrality in the span of fifteen years, and making it the fastest changing explicit bias. Charlesworth and Banaji, "Patterns of Implicit and Explicit Attitudes II."

[74] Ibid.

[75] *Brown v. Board of Education of Topeka*, 347 U.S. 483 (1954), https://www.oyez.org/cases/1940-1955/347us483. See also Alexandra Kalev and Frank Dobbin, "Retooling Career Systems to Fight Workplace Bias: Evidence from U.S. Corporations," *Dædalus* 153 (1) (Winter 2024): 213–230, https://www.amacad.org/publication/retooling-career-systems-fight-workplace-bias-evidence-us-corporations.

[76] For a discussion of why legislation is often inadequate, see Wanda A. Sigur and Nicholas M. Donofrio, "Implicit Bias versus Intentional Belief: When Morally Elevated Leadership Drives Transformational Change," *Dædalus* 153 (1) (Winter 2024): 231–249, https://www.amacad.org/publication/implicit-bias-versus-intentional-belief-when-morally-elevated-leadership-drives.