Implicit social cognition: A brief (and gentle) introduction

Chapter to appear in A. S. Reber & R. Allen (Eds.), *The cognitive unconscious: The first half-century*. Oxford, UK: Oxford University Press.

Benedek Kurdi
Yale University, New Haven, Connecticut

Mahzarin R. Banaji
Harvard University, Cambridge, Massachusetts

Author Note

Benedek Kurdi, Department of Psychology, Yale University; Mahzarin R. Banaji, Department of Psychology, Harvard University.

Benedek Kurdi is a member of the Scientific Advisory Committee of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition. Mahzarin R. Banaji served as interim Board member of the same organization from September 2020 to January 2021.

Correspondence concerning this chapter should be addressed to Benedek Kurdi, Department of Psychology, Yale University, 2 Hillhouse Ave, New Haven, CT 06511, email: benedek.kurdi@yale.edu.

Implicit social cognition: A brief (and gentle) introduction

Among the central features of the mind that drive social behavior are the attitudes and beliefs that humans bring to their engagement with other humans. Accordingly, ever since the inception of the field, research in social psychology has been guided by an overarching interest in these two distinct but related constructs. The concept of *attitude* (Eagly & Chaiken, 1993) refers to a feeling toward or affective evaluation of a social entity along a general good–bad dimension and has traditionally been measured via items such as “How much do you love the Red Sox?” or “How warmly do you feel toward African Americans?”. The concept of *belief* (in the context of social groups, a *stereotype*; Hamilton & Sherman, 1994) refers to specific semantic knowledge about a social entity, which may also be infused with affective value but cannot be fully reduced to it (“How strong is the Red Sox defense this year?”; “Are men more skilled at performing surgery than women?”).

The birth of social psychology is synonymous with attempts to understand and measure attitudes and stereotypes in an intergroup context. In a classic study, LaPiere (1934) first asked restaurant owners about their willingness to serve Chinese customers in their establishments and then measured how these attitudes predicted real-world behavior toward a specific Chinese patron. To study stereotypes, Katz and Braly (1933) asked undergraduates at Princeton University to characterize different racial and ethnic groups by selecting traits from a list. In this first documented demonstration of endorsed stereotypes about racial and ethnic groups in American society, participants showed no trepidation expressing that Germans were stolid and industrious, Italians were artistic and dirty, Negroes were musical and lazy, Jews were sly and shrewd, and so on.

A methodological feature shared by these early studies, and thousands of others published since, is the use of verbal self-report to capture social attitudes and stereotypes. Notably, these measures rely on participants' capacity for introspective access. That is, they reflect mental content that is (*a*) accessible to conscious awareness and (*b*) deemed appropriate to express given the individual's personal values and social norms. However, as the middle of the 20th century approached, reports of openly negative attitudes and stereotypes toward gender and race categories started to dissipate (Crosby, Bromley, & Saxe, 1980). Consequently, questions emerged about the validity of self-report measures in contexts where social desirability norms are prominent. These developments pushed social psychologists to consider indirect methods of measuring social attitudes (see Banaji & Greenwald, 2013, Appendix A for a review of the early approaches to measurement).

Two parallel developments concerning semantic and episodic memory facilitated this initial, and ever increasing, shift in attention toward the implicit (less conscious, less intentional, and less controllable) aspects of attitudes and stereotypes. Beginning in the 1970s and in full force by the 1980s, the precise recording of response times using personal computers allowed psychologists to explore the nature and organization of semantic representations without reliance on self-report. For instance, early investigations demonstrated that participants were faster to respond to a word, such as "nurse", if it was preceded by a semantically related word, such as "doctor", rather than a semantically unrelated word, such as "bread" (Meyer & Schvaneveldt, 1971; Neely, 1976). Findings of this kind were widely interpreted to indicate that concepts are organized into semantic networks in the human mind (Collins & Loftus, 1975): The closer two concepts are to each other in this network, the more one (such as "nurse") facilitates processing of the other (such as "doctor").

Second, research on episodic memory started uncovering intriguing dissociations between different memory processes. Specifically, it was shown that explicit (verbalizable) forms of memory need not be aligned with implicit (unconscious) forms of memory, which reveal themselves indirectly in a person's behavior, sometimes in the complete absence of any conscious recollection (Sherry & Schacter, 1987). In an illustrative study, Weiskrantz and Warrington (1979) used a distinctive-looking apparatus to deliver mildly unpleasant air puffs to the eyes of two amnesic patients. Crucially, each air puff was preceded by a tone. Over time, both participants developed a robust conditioned response: Merely upon hearing the tone, they closed their eyes, presumably to avoid the unpleasant sensation that they were anticipating. This response even persisted over a 24-hour delay. However, crucially, when participants were asked to describe the purpose of the apparatus or to recall what happened during the experiment the previous day (or even 10 minutes before), they were at a loss to report any memory for the events.

This case study, in combination with myriad other experiments involving neurologically intact individuals (e.g., Jacoby, 1991; Schacter & Graf, 1986), provided convincing evidence that episodic memory can operate outside the boundaries of conscious recollection. Excited by these new developments, social psychologists started probing whether automatic processes of conceptual association and implicit memory may also unfold in the domain of social attitudes and stereotypes.

Fazio, Sanbonmatsu, Powell, and Kardes (1986), Devine (1989), and Gaertner and McLaughlin (1983) reported the first empirical demonstrations of the automaticity of race-based attitude and stereotype activation (see also Bargh, this volume). Together, these investigations provided evidence that the social categories "White" and "Black" can facilitate the processing of evaluatively and semantically congruent words (such as "delightful" and "ambitious" vs. "awful" and “lazy”). Banaji, Hardin, and Rothman (1993) studied the selective application of activated knowledge to individuals depending on their social group membership. For example, when participants were unobtrusively primed with the trait “dependent,” they subsequently judged a female (but not a male) target as more dependent. These authors used the term “implicit stereotype” to name the resulting bias in person judgment, directly borrowing from Schacter and Graf’s use of the term “implicit memory” (see also Reber, 1967). Although the term “implicit social cognition” was not coined until a decade following these initial findings (Greenwald & Banaji, 1995), in retrospect it is clear that the seeds of a new approach to studying attitudes and stereotypes had already been planted.

Greenwald and Banaji (1995) took inspiration from the work of Nisbett and Wilson (1977) and synthesized several early and emerging discoveries as evidence of implicit social cognition. They defined implicit social cognition as the “introspectively unidentified (or inaccurately identified) traces of past experience” (p. 8) that mediate feelings of favorability (attitudes) or attribution of qualities to members of a social category (stereotypes) and used this definition to provide a unifying framework for a host of phenomena previously regarded as disconnected. Among them were discoveries concerning differential assessments of individuals marked by social group membership (such as age, gender, sexual orientation, race, ethnicity, or social class). This effect, operating without the perceiver’s awareness, was assumed to reflect an *implicit bias*. Today, this term has transcended academic psychology and entered the public discourse with daily use of the term in the media to refer to group-based discrimination in areas such as employment, education, housing, healthcare, law, and law enforcement.

Over the past three decades, implicit social cognition research has flourished and has produced myriad novel insights into the automatic operation of social attitudes and stereotypes. In this chapter, we provide an overview of what we regard to be significant and settled issues as well as the most pressing open questions that remain. We address (*a*) basic findings, such as mean levels of and demographic variation in implicit bias; (*b*) the relationship of implicit attitudes and stereotypes with other measures, including explicit attitudes and stereotypes and other forms of intergroup behavior; (*c*) the neural underpinnings of implicit bias; (*d*) questions of stability and change at different levels of analysis, including developmental stability, situational malleability, the prospect of long-term change within a single individual, and societal-level change; and (*e*) ongoing work and stimulating new developments, including aggregate-level analyses, the role of language, and questions about the representational format of implicit attitudes and stereotypes.

The implicit social cognition literature is vast, with hundreds of new papers published every year. Therefore, this chapter must necessarily be selective and, as such, will only touch on some of the most notable advances. We primarily highlight findings obtained with one particular measure of implicit cognition called the Implicit Association Test or IAT (Greenwald, McGhee, & Schwartz, 1998; for a quick demonstration, see Figure 1) given its widespread use and our own familiarity with it. Moreover, our discussion focuses on the largest databases, which happen measure race attitudes and gender stereotypes, with occasional mention of tests involving other social categories.

How large and pervasive are implicit attitudes and stereotypes?

Much of the early research in social psychology was concerned with documenting the existence, magnitude, and pervasiveness of social attitudes and stereotypes using what we now would refer to as explicit measures of self-report (Allport, 1935; Katz & Braly, 1933; LaPiere, 1934). Following the introduction of the IAT, similarly extensive efforts were made to document the magnitude and pervasiveness of implicit attitudes and stereotypes. This endeavor, which is still ongoing, has been greatly facilitated by the Project Implicit educational website (http://implicit.harvard.edu), which has collected data from over 25 million volunteer participants since its launch in 1998. An initial systematic summary of the insights emerging from the large-scale data collected via Project Implicit was first published in 2007 (Nosek et al., 2007; see also Ratliff et al., 2020). For the sake of brevity, here we focus on data obtained using the White/Black–good/bad attitude IAT and the male/female–career/home stereotype IAT during the periods for which data have been continuously collected and made publicly available (2004–2016 for the former and 2007–2018 for the latter).

Figure 2 shows explicit and implicit White/Black–good/bad race attitude data from 1,321,761 White American and 218,739 Black American visitors to the Project Implicit website from 2004–2016. The differences between explicit and implicit attitudes are striking. On the explicit measure, 39 percent of White Americans expressed preference for their own group, 58 percent expressed neutrality, and the remaining 3 percent expressed preference for the outgroup (i.e., Black Americans). By contrast, Black Americans showed considerably more marked explicit ingroup favoritism: Here, the majority (53 percent) expressed ingroup preference, 42 percent expressed neutrality, and the remaining 5 percent expressed preference for the outgroup (i.e., White Americans). Overall, explicit attitudes among both groups seemed to exhibit a mix of ingroup-favoring tendencies (Tajfel, 1982) and the desire to appear unbiased (Plant & Devine, 1998). Given the history of race relations in the United States, it is perhaps not surprising that social desirability concerns appear to have shaped attitudinal reports of White Americans to a greater extent than those of Black Americans.

On the IAT, a very different picture emerges. Specifically, the majority of White participants (73 percent) exhibited preference for their own group, and strikingly, this includes 69 percent of participants who reported no ingroup preference on the explicit measure. Only 16 percent of White participants were neutral on the implicit measure and the remaining 11 percent demonstrated a preference for the outgroup. By contrast, the implicit attitudes of Black participants were considerably closer to an even split, with 42 percent showing an implicit preference for their ingroup, 25 percent showing neutrality, and the remaining 34 percent showing a preference for the White outgroup.

The explicit–implicit dissociation is notable: White participants exhibited ingroup favoritism on both measures, although this tendency was more pronounced on implicit than explicit measures; Black participants exhibited stronger ingroup preference than White participants on the explicit measure, but showed intergroup neutrality on the implicit measure, challenging theories that assume that ingroup-favoring attitudes are ubiquitous. Although multiple interpretations of this dissociation are possible, the most parsimonious explanation appears to be that Black Americans have internalized the negative attitude toward their group in American society.

Figure 3 demonstrates data from 199,879 male and 473,240 female participants from 2007–2018 probing explicit and implicit gender stereotypes associating men with the concept of career and women with the concept of home. These findings differ in interesting ways from the White/Black attitude results discussed above. Most notably, male and female participants and explicit and implicit measures converge in showing the same stereotypic male–career and female–home association. In fact, the mean levels of explicit (Cohen’s d = 0.76) and implicit (Cohen’s d = 1.00) stereotypes were quite similar to each other.

Overall, these data (along with data from myriad additional tests not discussed here) allow for certain conclusions to be drawn regarding the basic nature of implicit social cognition. First, implicit measures of attitudes and stereotypes can be used to document the presence of pervasive social group biases. For instance, on the tests discussed above, the overwhelming majority of participants exhibited an attitudinal association of White Americans with positivity and Black Americans with negativity as well as a stereotypic association of male with career and female with home. Second, implicit social group biases can emerge even in the absence of any explicit bias: Most participants expressing explicit neutrality (65 percent on the attitude measure and 67 percent on the stereotype measure) showed an implicit bias favoring the dominant group. Third, participants' group membership can modulate implicit biases and these effects do not always mirror the effects obtained with explicit measures. For example, Black Americans (and other stigmatized groups) have been shown to exhibit neutrality on the race attitude IAT, in opposition to ingroup-favoring attitudes that they show on explicit measures of attitudes.

How are implicit attitudes and stereotypes related to other measures?

How are implicit attitudes and stereotypes related to their explicit counterparts?

These findings suggest that implicit attitudes and stereotypes are neither fully independent of their explicit counterparts nor fully redundant with them: Patterns of means at the group (or subgroup) level can be similar but not identical (as they were for the gender stereotype) or quite different (as they were for the race attitude). However, mean levels need not be instructive about the implicit–explicit relationship at the level of individuals, the question to which we turn next. For example, for the race attitude test discussed above, explicit–implicit correlation at the individual level was *r* = .32 although mean levels were different across the two measures; for the gender stereotype test, the correlation was substantially weaker (*r* = .16) despite similar mean levels.

Echoing the insights derived from the two tests introduced above, the overall conclusion from the research investigating the implicit–explicit relationship at the individual level is that explicit and implicit measures of attitude tend to reflect separate but related constructs, especially in the social domain (Bar-Anan & Vianello, 2018; Cunningham, Nezlek, & Banaji, 2004b; Nosek & Smyth, 2007). The most comprehensive test in this area was conducted by Bar-Anan and Vianello (2018) who had over 24,000 participants complete different explicit and implicit measures of attitude toward race, politics, and the self in a so-called multitrait–multimethod design (Campbell & Fiske, 1959). When the data were analyzed using structural equation modeling methods, evidence emerged for separate explicit and implicit attitude constructs in all three domains. Correlations between the explicit and implicit constructs were uniformly positive, but quite different in magnitude across content areas, ranging from *r* = .29 for self-esteem to *r* = .69 for race attitudes and *r* = .91 for political attitudes.

The finding by Bar-Anan and Vianello (2018) confirms the results of an early investigation by Nosek (2005). This paper calculated explicit–implicit correlations for 57 attitude object pairs, which ranged from nonsignificant and negative (*r* = -.05) for attitudes toward men vs. women to almost perfect and positive (*r* = .70) for attitudes toward pro-choice vs. pro-life political positions. Based on this data-driven approach, as well as a meta-analytic review, authors have identified several theoretical predictors of when explicit and implicit attitudes should or should not be associated with each other (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Nosek, 2007). Perhaps most notably, explicit and implicit measures tend to be more dissociated in areas in which expressing a negative attitude may be seen as socially sensitive, such as the

White/Black–good/bad attitude test discussed in detail above. When such concerns are absent, explicit and implicit attitudes are usually more closely aligned.

How is implicit social cognition related to intergroup behavior?

If implicit attitudes and stereotypes are related but not identical to their explicit counterparts, then they should both independently predict intergroup behavior, such as hiring and promotion, healthcare decisions, and classroom interactions with members of different social groups. Indeed, this is precisely the finding that emerged from a recent large-scale meta-analysis relying on data from over 35,000 participants across 217 published and unpublished research reports (Kurdi et al., 2019b). The average independent contributions of explicit ($\beta = .11$) and implicit ($\beta = .14$) measures to predicting intergroup behavior were both small but statistically significant and almost identical in size. Importantly, the implicit–behavior correlations included in this meta-analysis were all calculated at the level of individual participants rather than at the level of larger geographic units (Kurdi & Banaji, 2017), a point to which we return below.

Beyond the overall size of the relationship, the meta-analysis also investigated a large number of potential moderators of the implicit–behavior correlation. Importantly, the relationship was found to be robust and significant for virtually all types of behavior (ranging from social affiliation to person perception), all target groups (including race, gender, and sexual orientation), and settings (including lab, real-world, and online). In addition, the meta-analysis identified methodological features of the studies that produced smaller or larger magnitudes of the implicit–behavior correlation: Studies that were weak on these variables produced a diminishingly small effect of $r = .02$, whereas studies that were strong on these variables produced a medium-sized correlation of $r = .37$. Curiously, theoretically derived predictions suggesting that implicit attitudes should be especially effective in predicting socially sensitive and difficult-to-control behaviors that are low in conscious awareness were not borne out by the data: None of these variables moderated the implicit–behavior relationship.

What neural substrates underlie implicit social cognition?

Compared with the work investigating explicit–implicit relationships and the association between implicit attitudes and behavior, the literature on the neural correlates of implicit social cognition is considerably less voluminous (for reviews see Kubota, Banaji, & Phelps, 2012; Stanley, Phelps, & Banaji, 2008). However, this literature has produced some important insights regarding the basic nature of implicit social attitudes, and especially implicit White/Black race attitudes, worth addressing here.

In an early study, Phelps et al. (2000) found that the magnitude of the response to Black relative to White faces in a subcortical structure (amygdala) was correlated with White participants' implicit but not explicit race attitudes. This result was important for two reasons: First, it established a relationship between IAT performance and neural activation in a brain area known to be implicated in the processing of emotionally relevant stimuli. Second, it provided yet another indication of a dissociation between explicit and implicit attitudes, this time in terms of their neural substrates.

In follow-up work, the difference in amygdala activation in response to White relative to Black faces was especially strong when faces were presented very briefly, thus precluding conscious processing (Cunningham et al., 2004a). In contrast, when White participants viewed faces for longer durations and thus became consciously aware of them, areas of the neocortex known to be involved in top-down processing and conflict monitoring, including dorsolateral prefrontal cortex (dlPFC) and the anterior cingulate cortex (ACC), took over. In this condition, activation in dlPFC and ACC was negatively correlated with the strength of the amygdala response, suggesting suppression of the initial negative response to Black faces via control processes in line with participants' consciously endorsed egalitarian ideals. Further testing will be necessary to obtain better evidence on the temporal nature of such modulation.

Perhaps reflecting a general waning of enthusiasm for the use of imaging techniques, such as functional magnetic resonance imaging (fMRI), to understand cognitive processes (Niv, 2020), studies investigating the relationship between neural activation and IAT performance have become few and far between. However, an exciting new direction seems to have emerged in this area: Some investigators have started using methods, such as transcranial magnetic stimulation (TMS), that can temporarily disrupt the functioning of specific brain areas to understand the neural basis of implicit social cognition (Marini, Banaji, & Pascual-Leone, 2018). Unlike correlation-based techniques such as fMRI, this approach has the potential to establish causal relationships between neural processes unfolding in the human brain and the automatic expression of attitudes and stereotypes on implicit measures such as the IAT.

Is implicit social cognition amenable to change?

The correlational studies discussed above have provided important insights into the basic nature of implicit social cognition, including the nonzero but nonredundant relationship between implicit attitudes and stereotypes and their explicit counterparts, the unique contributions of each to predicting behavior in intergroup contexts, and the neural correlates of each. However, as experimental psychologists, we subscribe to the adage attributed to social psychologist to Kurt Lewin, "If you want truly to understand something, try to change it." In this spirit, below we review evidence on the situational malleability of implicit social cognition, its stability or change across the lifespan, the prospects for durably shifting implicit social cognition in adults via targeted interventions, and societal-level changes in implicit social cognition over cultural time.

Is implicit social cognition situationally malleable?

At the outset of implicit social cognition research, implicit attitudes were often characterized as a “cognitive monster” (Bargh, 1999): impenetrable, recalcitrant, unwilling to budge. Presumably, this view was, at least in part, inherited from spreading activation models (Collins & Loftus, 1975) under which associations between conceptual nodes in long-term memory can be updated only in a slow, painstaking, step-by-step manner. However, contrary to this view, the past two decades have produced a remarkable amount of evidence for the malleability of implicit social cognition in the face of immediate situational affordances (e.g., Blair, 2002; Lai et al., 2014).

To name a few early examples: Implicit race bias among White American participants has been shown to temporarily decrease in the presence of a Black rather than White experimenter (Lowery, Hardin, & Sinclair, 2001) and when Black faces are presented in the context of a family barbecue rather than a dilapidated street corner (Wittenbrink, Judd, & Park, 2001). In a study by Blair, Ma, and Lenton (2001), participants exhibited lower levels of implicit gender stereotyping after a mental imagery exercise during which they were asked to think about a strong woman.

Lai et al. (2014) provided a systematic test of the effectiveness of multiple interventions against implicit race bias in a research competition format. They found that eight of the 17 manipulations submitted by researchers shifted responding on a subsequent IAT, some of them considerably. The most effective intervention prompted participants to imagine a vivid and self-relevant counterattitudinal scenario in which they were viciously attacked by a White man and a heroic Black man came to their rescue. In addition, implicit bias toward Black Americans also showed a temporary decrease after White participants (*a*) completed a game with Black teammates and White competitors; (*b*) practiced the IAT using counterstereotypic exemplars; (*c*) were trained to use implementation intentions ("I will think 'good' when I see a Black face"); (*d*) were primed with multiculturalism; or (*e*) completed one of several variants of evaluative conditioning pairing Black faces with positive and White faces with negative stimuli.

Overall, contrary to pervasive early ideas about the "cognitive monster", this work has robustly demonstrated high levels of situational malleability in implicit social cognition. What is being signaled by entities and events in the immediate environment is capable of shaping even mental representations that were assumed to be long in developing and possessing sufficient stability that they could not be penetrated by conscious will. However, important questions have yet to be answered. First, in the absence of targeted and sufficiently well-powered studies, it is unclear whether and under what conditions immediate changes in implicit social cognition can facilitate immediate changes in intergroup behavior. Second, no unifying theoretical framework exists that can account for both the remarkable malleability of implicit social cognition and its occasional recalcitrance in the face of certain manipulations. Third, as discussed in more detail below, it remains to be seen whether the interventions used to create temporary malleability in implicit social cognition can be harnessed to achieve long-term change.

Does implicit social cognition change over the lifespan?

Given the remarkable malleability of implicit social cognition from context to context, it would seem prudent to assume that, in the aggregate, small situational changes would eventually combine to cause larger shifts in implicit bias over a person's lifespan. Specifically, in line with the often repeated saying according to which no one is born with prejudice, young children may presumably start life out as blank slates with regard to social group attitudes and stereotypes and then slowly shift away from neutrality in the direction of cultural beliefs. However, as reasonable as this view sounds, it was demonstrated to be incorrect in the first study conducted in this domain (Baron & Banaji, 2006) and in every relevant study performed since.

Specifically, in study after study, the development of implicit social cognition over the lifespan seems to be characterized by two striking features (Dunham, Baron, & Banaji, 2008): Implicit attitudes emerge as early as social group concepts themselves and, at least in terms of their magnitude, they are invariant over development. That is, children by the age of 6 (the earliest age at which the IAT can be administered) seem to exhibit the same magnitudes of implicit social group bias as adult samples recruited from the same communities. Other methods capable of eliciting reliable data from younger children confirm the presence of implicit intergroup bias in samples from the North America, East Asia, and Africa already at age 4 (Dunham, Chen & Banaji, 2013; Qian et al., 2016)

The features of early emergence and stability over the lifespan seem to characterize implicit attitudes not only among members of dominant groups, such as White Americans, but even among stigmatized groups, such as Black Americans (Dunham et al., 2013). What is more, children already seem to exhibit spontaneous gender biases in their speech patterns even at ages when they are too young to complete an IAT (Charlesworth, Yang, Mann, Kurdi, & Banaji, in press). Overall, this pattern of results seems to be fundamentally incompatible with the idea of incremental change and suggests that social group attitudes are a candidate for a core system in human development (Spelke & Kinzler, 2007).

However, similar to the domain of situational malleability, important open questions remain. First, developmental stability in the magnitude of implicit attitudes and stereotypes need not imply stability in the psychological mechanisms maintaining such attitudes and stereotypes (Baron, 2015). Theoretically guided intervention work involving children, which would be needed to address the issue of underlying mechanisms, is still in its infancy (but see Charlesworth, Kurdi, & Banaji, 2019; Gonzalez, Dunlop, & Baron, 2016). Second, and most important, as of now, no longitudinal work tracking implicit social attitudes or stereotypes within the same individuals over longer periods of time has been conducted. As such, existing studies are subject to the known limitations of cross-sectional approaches, which provide a snapshot of different age groups at a particular point in time.

What are the prospects for creating durable change in implicit social cognition?

Implicit social cognition is malleable from one moment to the next. At the same time, it seems remarkably stable in the face of information to which individuals are spontaneously exposed over development. However, such naturally occurring developmental stability need not imply that targeted interventions designed to shift implicit social biases toward neutrality are doomed to fail. After all, the stability of implicit social cognition over development may reflect the stability of social information to which individuals are exposed over time rather than a lack of flexibility in psychological processes.

And yet, especially relative to studies aiming to create short-term malleability, the number of papers attempting to achieve long-term change in implicit attitudes or stereotypes is modest. Of the studies that do exist, some have managed to produce change in implicit social cognition going beyond the immediate context of the experiment (Kurdi & Banaji, 2019; Mann & Ferguson, 2017; Mann, Kurdi, & Banaji, 2020); however, these studies involved previously unknown individuals or novel groups as the targets of intervention. When interventions target well-known social categories, such as race, the effects do not seem to reliably persist over time (Forscher, Mitamura, Dix, Cox, & Devine, 2017; Lai et al., 2016).

From our perspective, these findings are not particularly surprising and, if anything, they demonstrate the adaptive nature of the human mind. Indeed, it would be shocking if a single-shot 5-minute (or even 60-minute) intervention were able to undo decades of learning about well-known social categories, including from everyday language and the media (Caliskan, Bryson, & Narayanan, 2017; Charlesworth et al., in press). As such, we are skeptical as to whether the interventions currently used to create short-term malleability, notable as they are in teaching us about the nature of implicit cognition, will be able to serve as an effective jumping-off point for creating long-term change. Rather, achieving long-term change may be predicated on more fundamental changes to the types of social information that we consume on a daily basis (see below). Moreover, attempts for change at the individual level may be especially potent if accompanied by change initiated at the level of mezzo-level institutions (such as schools and work environments) and macro-level forces such as laws and governmental policies.

Are Americans changing in the magnitude of implicit intergroup bias over time?

When looking for hints about the potential for and drivers of durable shifts in implicit social cognition, societal patterns of change and stability over cultural time may be more instructive than experimental studies aimed at creating short-term malleability. In recent work in this area, Charlesworth and Banaji (2019) have analyzed the trajectory of implicit attitudes toward six social categories, including sexual orientation, race, skin tone, disability, age, and body weight in a sample of over 4 million participants from the United States.

Over the time period investigated, spanning 10 years from 2007 to 2016, implicit attitudes toward sexual orientation (gay/straight), race, and skin tone have all become considerably less biased. Having shown a drop of 33% over the decade, implicit sexual orientation attitudes are predicted to reach neutrality within nine years. Race and skin tone attitudes have also shifted toward neutrality, but with a drop of about 15% over the same time period. By contrast, implicit attitudes toward disability and age were found to be stable across time. Implicit bias against overweight people, in turn, may even have increased in strength.

In follow-up work, Charlesworth and Banaji (in press-b) also found evidence for change in implicit gender stereotypes at the rate of about 15%, mirroring the trajectory of some of the implicit attitudes mentioned above. What is more, change in implicit attitudes and stereotypes seems truly ubiquitous (Charlesworth & Banaji, in press-a): Members of different social groups (including the elderly and young, women and men, individuals of different races and ethnicities, those high and low in education and income, and those living in different geographic regions of the U.S.) demonstrate uniform shifts toward neutrality, with minimal variation. At the same time, on the fastest changing sexual orientation attitudes, younger individuals and those who identify as politically liberal are changing faster than other groups. In fact, these populations have already reached attitudinal neutrality. The finding that change on implicit attitudes is observed over the short time frame of a mere decade was unpredicted but welcome. However, the specific causes of change in sexual orientation and race attitudes (and stability in others) still remain to be discovered.

What do we not (yet) know about implicit social cognition?

We conclude by discussing three areas of implicit social cognition research in which remarkable developments have taken place over the last few years, which we expect to continue and even intensify in the near future. These areas include the role of aggregate-level analyses and language in the study of implicit social cognition, as well as vigorous theoretical debates regarding basic mechanisms of acquisition, change, and representation.

What can aggregate-level studies teach us about the nature of implicit social cognition?

A relatively small but steadily growing set of studies have investigated the relationship between implicit attitudes and stereotypes and consequential societal outcomes at the regional level (for reviews see Hehman, Calanchini, Flake, & Leitner, 2019; Payne, Vuletich, & Lundberg, 2017). Just to name a few examples, it has been shown that higher levels of regional implicit race bias are associated with a higher number of police killings of Black Americans (Hehman, Flake, & Calanchini, 2017), more racial discrimination in disciplinary actions in the school system (Riddle & Sinclair, 2019), larger disparities in health outcomes among adults (Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016) and infants (Orchard & Price, 2017), and lower levels of upward mobility among Black individuals (Chetty, Hendren, Jones, & Porter, 2019). Although currently no meta-analysis of these aggregate-level studies exists, it seems that the magnitude of the correlation between implicit bias and behavior in these studies can vastly exceed the more modest effect sizes found in individual-level studies.

This intriguing pattern of findings has prompted some theorists to conclude that implicit bias should be understood as primarily operating at the level of contexts, such as a county or a city, rather than at the level of individuals (Payne et al., 2017). Although we are sympathetic to this argument, we believe that much work remains to be done to establish that the difference in size between individual-level and aggregate-level correlations between implicit bias and behavior is not due to a statistical artifact or a difference in the nature of the outcome measures investigated (Connor & Evers, 2020; Kurdi & Banaji, 2017). Specifically, aggregate-level studies combine data from tens of thousands of participants for each unit of observation, which results in considerably less noisy measurements (and, therefore, higher observed correlations) than in individual-level studies using a single measurement for each person. Moreover, individual-level studies often rely on blatant measures of intergroup discrimination in a laboratory context. The obvious nature of the behaviors, in turn, may result in participants strategically modulating their responses to appear unbiased, thus depressing correlations both with implicit and explicit measures. By contrast, aggregate-level studies use behaviors that are by definition ecologically valid, consequential, and not easily modifiable at will.

Can language provide an independent window into implicit social cognition?

In a separate but related line of work, computer scientists first connected individual-level implicit biases to aggregate-level implicit biases emerging from large repositories of text produced online. Specifically, Caliskan, Bryson, and Narayanan (2017) demonstrated that individual-level implicit attitudes toward ten social categories indexed using the IAT are closely aligned with societal-level semantic associations derived from the vast Common Crawl corpus via so-called word embeddings. (Word embeddings are a family of machine learning algorithms used to produce a measure of semantic relatedness between concepts such as MALE and FEMALE and CAREER and HOME; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2017). We find such convergence remarkable given the obvious differences between the two measures both in methodological features and in units of analysis.

In recent work, word embeddings have been used both to provide convergent evidence for findings obtained using the IAT at the level of individual participants (Kurdi, Mann, Charlesworth, & Banaji, 2019a) and to detect the presence of social group biases in everyday interactions between children and parents, which would not have been amenable to quantitative analyses even a few years ago (Charlesworth et al., in press). Notably, studies focusing on basic mechanisms of acquisition and change have also demonstrated time and again the power of language in shifting implicit social attitudes, which often exceeds that of interventions relying on direct experience (De Houwer, 2006; Mann et al., 2020). Overall, these developments suggest that language-based approaches will play a major role in implicit social cognition research in the coming years.

What mechanisms of learning and representation support implicit social cognition?

Despite remarkable advances in understanding implicit social cognition over the past three decades, what we see as the most fundamental questions about this area are far from having satisfactory answers. Specifically, it is unclear what types of information support the acquisition and updating of implicit attitudes and stereotypes and even how the space of different types of information is best carved up in a theoretically informative manner. Moreover, the nature of the knowledge structures encoding implicit attitudes and stereotypes in the human mind is also relatively poorly understood.

The two major theoretical competitors in this area have been associative (e.g., McConnell & Rydell, 2014) and propositional (e.g., De Houwer, 2014) approaches to implicit social cognition. The former argue that implicit attitudes respond primarily to repeated pairings of stimuli in the environment (such as Black men appearing with guns or women appearing in kitchens) and are represented via conceptual associations (such as BLACK–DANGEROUS or WOMEN–HOME). By contrast, the latter posit that, just like their explicit counterparts, implicit attitudes can change rapidly in response to verbal information and are represented propositionally (such as “Black men are dangerous”). It is clear that verbal information is remarkably powerful in flexibly modulating implicit social cognition (De Houwer, 2006; Mann et al., 2020), thus supporting the propositional perspective. However, if both explicit and implicit attitudes are encoded via the same propositional representations, then how are we to explain clear explicit–implicit dissociations, such as the one that we observed in racial attitudes at the beginning of this chapter?

To summarize, although the propositional approach has reinvigorated research on the mechanisms of implicit attitude acquisition and change over the past decade, it seems clear that without auxiliary assumptions it cannot account for the totality of existing data (Kurdi & Dunham, 2020). Hybrid theories, such as the associative–propositional model (Gawronski & Bodenhausen, 2006), offer more flexibility but are also more difficult to falsify. There remains much conceptual and empirical work to be done in this area, with advances hopefully producing positive ripple effects across the whole of implicit social cognition research. Specifically, we hope that new theoretical approaches will be better integrated with several areas of empirical findings discussed above. These areas include the theoretical conditions under which implicit attitudes should be (*a*) more or less highly correlated with intergroup behavior and (*b*) more or less amenable to temporary malleability and long-term change at the individual and aggregate levels.

Seven takeaways to remember

In this chapter we have provided a brief, even gentle, introduction to the vast literature on implicit social cognition. For those who wish to immerse themselves more deeply in this work, the recent special issue of *Social Cognition* marking 25 years of implicit social cognition research (Gawronski, De Houwer, & Sherman, 2020) may serve as a useful starting point. In any case, we do hope that interested readers will come away with explicit recollection of a few basic findings from this enormous and burgeoning field:

(1) Implicit measures often reveal higher levels of social group biases than their explicit counterparts, including in participants endorsing egalitarian values.

(2) Explicit and implicit attitudes are distinct but related, with the strength of their relationship depending on factors such as the domain under investigation, including its social sensitivity.

High social sensitivity tends to produce dissociations (e.g., racial attitudes); low social sensitivity tends to produce convergence (e.g., political attitudes).

(3) Explicit and implicit measures are each modestly related to intergroup behavior at the level of individuals, with stronger relationships emerging in aggregate investigations of geographic units. Better approximating the methodological strengths of aggregate-level studies may eventually produce stronger relationships at the individual level.

(4) Correlations between subcortical brain activation and implicit behavioral measures suggest that the IAT can detect automatic expressions of group-based attitudes.

(5) Young children show evidence of implicit attitudes that mirror those of the adults of their social groups. As such, implicit attitudes seem to emerge from early learning that requires minimal input.

(6) Findings of stability versus change in implicit social cognition depend on the timescale investigated. Individual-level malleability at the scale of minutes and societal-level change at the scale of years are both well-documented. However, for now, both longitudinal studies establishing spontaneous change within a single individual and experimental demonstrations of durable intervention effectiveness are absent.

(7) Presumably, studies attempting to uncover mechanisms of durable change will be more successful once a better theoretical understanding of the learning processes and mental representations underlying implicit social cognition is achieved.

Given the rate at which research is accumulating from a variety of labs around the world, we stand to learn a great deal from the next decades of research on implicit social cognition.

References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Banaji, M. R., & Greenwald, A. G. (2016). *Blindspot: Hidden biases of good people*. New York, NY: Bantam Books.

Banaji, M. R., Hardin, C. D., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*(2), 272–281. http://doi.org/10.1037//0022-3514.65.2.272

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264–1272. http://doi.org/10.1037/xge0000383

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York, NY: Guilford Press.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes. Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, *17*(1), 53–58. http://doi.org/10.1111/j.1467-9280.2005.01664.x

Baron, A. S. (2015). Constraints on the development of implicit intergroup attitudes. *Child Development Perspectives*, *9*(1), 50–54. http://doi.org/10.1111/cdep.12105

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

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841. http://doi.org/10.1037//0022-3514.81.5.828

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. http://doi.org/10.1037/h0046016

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. http://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S., & Banaji, M. R. (in press-a). Patterns of implicit and explicit attitudes II. Long-term change and stability, regardless of group membership. *American Psychologist*.

Charlesworth, T. E. S., & Banaji, M. R. (in press-b). Patterns of implicit and explicit stereotypes III. Long-term change in gender–science and gender–career stereotypes. *Social Psychological and Personality Science.*

Charlesworth, T. E. S., Kurdi, B., & Banaji, M. R. (2019). Children's implicit attitude acquisition: Evaluative statements succeed, repeated pairings fail. *Developmental Science*, *14*(3), 464–10. http://doi.org/10.1111/desc.12911

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (in press). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of 65+ million words. *Psychological Science*.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2019, December). *Race and economic opportunity in the United States: An intergenerational perspective* (NBER Working Paper Series No. 24441). http://doi.org/10.3386/w24441

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

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science*, *116*(24), 174569162093149–17. http://doi.org/10.1177/1745691620931492

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin*, *87*(3), 546–563. http://doi.org/10.1037/0033-2909.87.3.546

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004a). Separable neural components in the processing of Black and White faces. *Psychological Science*, *15*(12), 806–813. http://doi.org/10.1111/j.0956-7976.2004.00760.x

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004b). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, *30*(10), 1332–1346. http://doi.org/10.1177/0146167204264654

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. http://doi.org/10.1016/j.lmot.2005.12.002

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

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

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences*, *12*(7), 248–253. http://doi.org/10.1016/j.tics.2008.04.006

Dunham, Y., Chen, E. E., & Banaji, M. R. (2013). Two signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. *Psychological Science*, *24*(6), 1–27. http://doi.org/10.1177/0956797612463081

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

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

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, *72*, 133–146. http://doi.org/10.1016/j.jesp.2017.04.009

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*(1), 23–30. http://doi.org/10.2307/3033657

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

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

Gonzalez, A. M., Dunlop, W. L., & Baron, A. S. (2016). Malleability of implicit associations across development. *Developmental Science*, *19*(9), 1–13. http://doi.org/10.1111/desc.12481

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. http://doi.org/10.1037//0033-295X.102.1.4

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

Hamilton, D. L., & Sherman, J. W. (1994). Stereotypes. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition: Basic processes; Applications* (pp. 1–68). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Hehman, E., Flake, J. K., & Calanchini, J. (2017). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *2*, 194855061771122–9. http://doi.org/10.1177/1948550617711229

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. http://doi.org/10.1177/0146167205275613

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. http://doi.org/10.1016/0749-596X(91)90025-F

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology*, *28*(3), 280–290. http://doi.org/10.1037/h0074049

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Neuroscience*, *15*(7), 940–948. http://doi.org/10.1038/nn.3136

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

Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology*, *116*(5), 681–703. http://doi.org/10.1037/pspa0000151

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

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019a). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *21*, 201820240–10. http://doi.org/10.1073/pnas.1820240116

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

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. http://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. http://doi.org/10.1037/xge0000179

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces*, *13*(2), 230–237. http://doi.org/10.2307/2570339

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Racial bias is associated with ingroup death rate for Blacks and Whites: Insights from Project Implicit. *Social Science & Medicine*, *170*(C), 220–227. http://doi.org/10.1016/j.socscimed.2016.10.007

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

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, *68*(C), 122–127. http://doi.org/10.1016/j.jesp.2016.06.004

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), 1169–1192. http://doi.org/10.1037/xge0000701

Marini, M., Banaji, M. R., & Pascual-Leone, A. (2018). Studying implicit social cognition with noninvasive brain stimulation. *Trends in Cognitive Sciences 22*(11), 1050–1066. http://doi.org/10.1016/j.tics.2018.07.014

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 204–217). New York, NY: Guilford Press.

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

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017, December 26). *Advances in pre-training distributed word representations*. arXiv. https://arxiv.org/abs/1712.09405

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. http://doi.org/10.3758/BF03213230

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

Niv, Y. (2020, October 22). *The primacy of behavioral research for understanding the brain*. PsyArXiv. https://psyarxiv.com/y8mxe

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. http://doi.org/10.1037/0096-3445.134.4.565

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69. http://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., & Smyth, F. L. (2007). A multitrait–multimethod validation of the Implicit Association Test. *Experimental Psychology*, *54*(1), 14–29. http://doi.org/10.1027/1618-3169.54.1.14

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. http://doi.org/10.1080/10463280701489053

Orchard, J., & Price, J. (2017). County-level racial prejudice and the Black–White gap in infant health outcomes. *Social Science & Medicine*, *181*, 191–198. http://doi.org/10.1016/j.socscimed.2017.03.036

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

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. http://doi.org/10.1162/089892900562552

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. http://doi.org/10.1037//0022-3514.75.3.811

Qian, M. K., Heyman, G. D., Quinn, P. C., Messi, F. A., Fu, G., & Lee, K. (2016). Implicit racial biases in preschool children and adults from Asia and Africa. *Child Development*, *87*(1), 285–296. http://doi.org/10.1111/cdev.12442

Ratliff, K. A., Lofaro, N., Howell, J. L., Conway, M. A., Lai, C. K., O'Shea, B., et al. (2020). *Documenting bias from 2007–2015: Pervasiveness and correlates of implicit attitudes and stereotypes II.* PsyArXiv. https://osf.io/jeyc7

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. http://doi.org/10.1016/S0022-5371(67)80149-X

Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences*, *88*, 201808307–6. http://doi.org/10.1073/pnas.1808307116

Schacter, D. L., & Graf, P. (1986). Effects of elaborative processing on implicit and explicit memory for new associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 432–444. http://doi.org/10.1037/0278-7393.12.3.432

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439–454. http://doi.org/10.1037/0033-295X.94.4.439

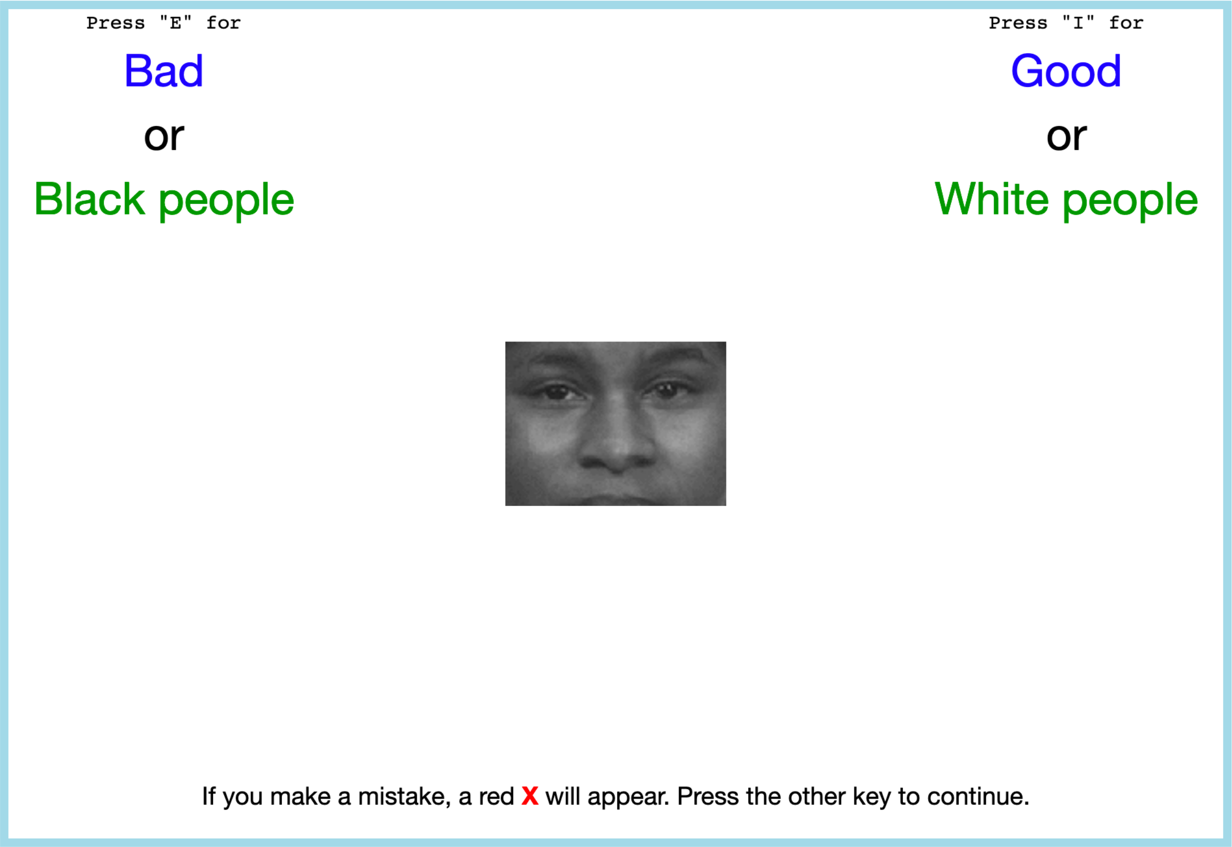
Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. http://doi.org/10.1111/j.1467-7687.2007.00569.x

Stanley, D. A., Phelps, E. A., & Banaji, M. R. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science*, *17*(2), 164–170. http://doi.org/10.1111/j.1467-8721.2008.00568.x

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*(1), 1–39. http://doi.org/10.1146/annurev.ps.33.020182.000245

Weiskrantz, L., & Warrington, E. K. (1979). Conditioning in amnesic patients. *Neuropsychologia*, *17*(2), 187–194. http://doi.org/10.1016/0028-3932(79)90009-5

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815–827. http://doi.org/10.1037//0022-3514.81.5.815



If you make a mistake, a red X will appear. Press the other key to continue.

Figure 1. Illustration of the Implicit Association Test (IAT) procedure. In the congruent block (shown here), participants use the same response key to sort Black faces and negative words and a different response key to sort White faces and positive words. In the incongruent block, the pairings are reversed: Black goes together with good and White goes together with bad. The IAT effect is calculated as a standardized measure of the difference in response times across the congruent (White/good–Black/bad) and incongruent (White/bad–Black/good) blocks. In the male/female–career/home IAT, the White and Black faces are replaced with male and female names and the positive and negative words with words related to home and career. Otherwise, the procedure remains identical.

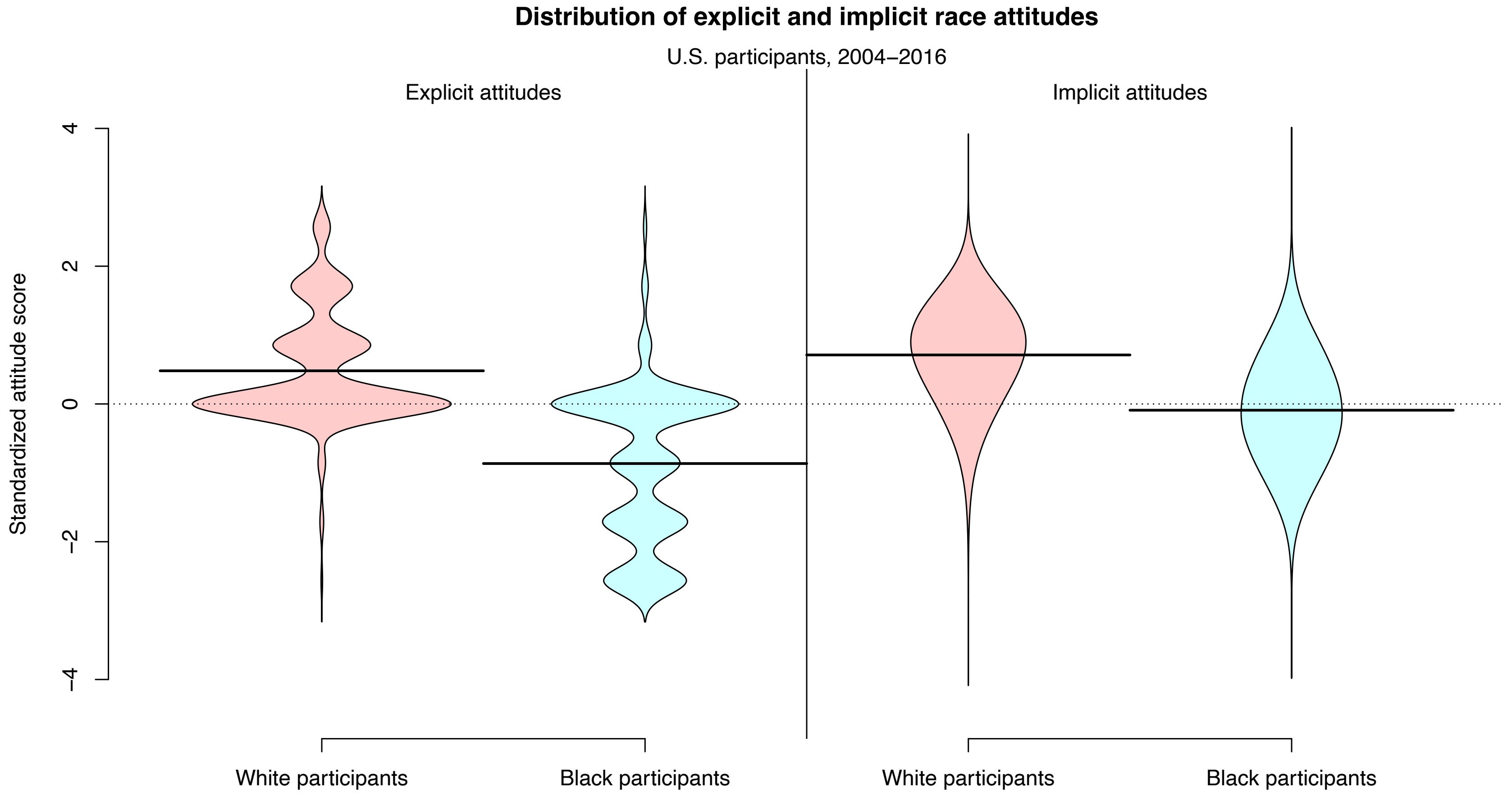


Figure 2. Explicit (self-reported) and implicit (IAT) attitudes toward White and Black Americans obtained from U.S. visitors to the Project Implicit website in the years 2004 through 2016. Positive scores indicate a preference for White over Black Americans. To ensure comparability, the explicit and implicit attitude scores were standardized.

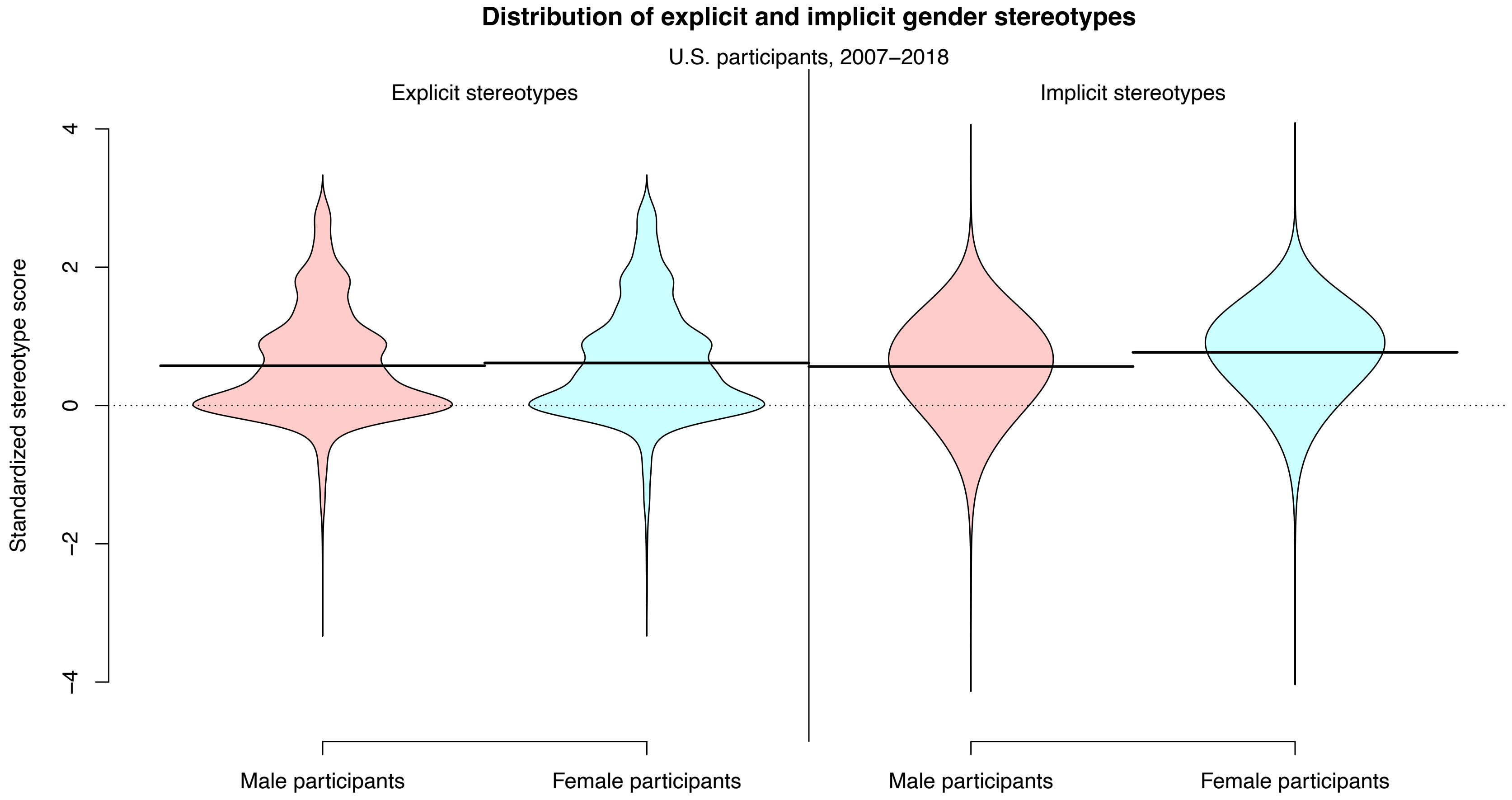


Figure 3. Explicit (self-reported) and implicit (IAT) gender stereotypes obtained from U.S. visitors to the Project Implicit website in the years 2007 through 2018. Positive scores indicate an association of the category "men" with the concept "career" and of the category "women" with the concept "home." To ensure comparability, the explicit and implicit stereotype scores were standardized.