Invited submission to *Neuroscience*
Forefront Review article

Grappling with implicit social bias: A perspective from memory research

Heather D. Lucas[1], Jessica D. Creery[2], Xiaoqing Hu[3] and Ken A. Paller[2]

[1]Department of Psychology, Louisiana State University, Baton Rouge, LA 70803, USA

[2]Department of Psychology and Cognitive Neuroscience Program, Northwestern University, Evanston, IL 60208, USA

[3]Department of Psychology, The State Key Laboratory of Brain and Cognitive Science, The University of Hong Kong, Pokfulam, Hong Kong

Running head: MEMORY AND IMPLICIT BIAS

Acknowledgements

We are grateful for input from two anonymous reviewers and for support from the National Science Foundation (BCS-1461088, BCS-1461088, BCS-1829414), the Mind and Life Institute (Varela Grant), the Mind Science Foundation, and NIH Training Grant T32-NS047987.

Abstract

There is now widespread consensus that social biases often influence actions independently of the actor’s intention or awareness. The notion that we are sometimes blind to the origins of our thoughts, attitudes, and behaviors also features prominently in research into domain-general human memory systems, which has a long history of distinguishing between implicit and explicit repercussions of past experience. A shared challenge across these fields of study is thus to identify techniques for effectively managing the contents of our memory stores, particularly those aspects into which we have limited metacognitive insight. In the present review, we examine recent developments in the cognitive neuroscience of human memory that speak to this challenge as it applies to the social domain. One area of progress pertains to the role of individuation, the process of attending to and representing in memory unique characteristics of individuals, which can limit the influence of generalizations based on social categories. A second body of work concerns breakthroughs in understanding memory consolidation, which determines the fate of newly encoded memories. We discuss the promise of each of these developments for identifying ways to become better stewards of our social minds. More generally, we suggest that, as with other forms of learning and memory, intentional practice and rehearsal may be critical in learning to minimize unwanted biases.

Introduction

We don't always know why we do what we do — but we do have the ability to gain a better understanding of the hidden causes of our behavior. The past two decades have witnessed a dramatic increase in awareness of how social attitudes and behaviors can be shaped by cognitive processes that operate *implicitly*, outside of conscious awareness and thus beyond direct intentional control. A seminal paper by Greenwald and Banaji (1995) is credited with introducing the term *implicit social cognition* to refer to a broad collection of nonconscious or unintentional influences on behavior, including self-evaluations, social attitudes, and stereotype attribution.

Currently, there is widespread interest in basic and applied research on the topic of *implicit bias*, or automatic expressions of prejudice based on social group memberships. In a 2011 review, Nosek and colleagues catalogued 20 experimental procedures to assess implicit social bias, which together amassed more than 6200 citations. The most widely used method, which accounted for 43.6% of citations, is the Implicit Association Task (IAT, Greenwald, McGhee, & Schwartz, 1998, see also Figure 2). Interpretations of data obtained using this and similar measures, and particularly their relationships to overt discrimination, are the subject of ongoing debate (Gawronski & Bodenhausen, 2017; Greenwald, Banaji, & Nosek, 2015; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013; Payne, Vuletich, & Lundberg, 2017). Nonetheless, the IAT remains a popular and productive research tool, due in part to its good internal consistency (Gawronski, Morrison, Phills, & Galdi, 2017) and temporal stability at the aggregate level (e.g., within a given social environment and demographic group; Jost, 2018; Payne et al., 2017)[1]

The breakthrough insight of Greenwald and Banaji (1995) that led to this proliferation of research was arguably their assertion that explicit assessments of social attitudes and beliefs measure only a subset of the information that people draw upon when navigating the social world. The inadequacy of such self-report measures stems, in part, from concerns about self-preservation that are unique to the social domain. Nevertheless, research on implicit social cognition can be considered a branch of an older body of work on implicit cognitive processing. This area of study was pioneered by research that distinguished between implicit and explicit influences on human memory.

Implicit memory refers to a diverse collection of processes—including cognitive and motor skill learning, habit learning, conditioning, and priming—by which traces of past experience can influence behavior without awareness of this influence. These expressions of memory are quite different from one another, and yet they can all be considered to be implicit. The neural system linked to memory that is characterized by awareness of retrieval (i.e., *declarative* or explicit memory, Eichenbaum & Cohen, 2001; Squire, 2004) operates separately from the implicit memory systems, although in practice, behavior is often determined by interactions among multiple systems.

Implicit memory has been conceptualized as a form of pervasive plasticity by which virtually all of our neural systems continuously adjust to reflect statistical regularities or co-occurrences in our environments (Reber, 2013). In other words, implicit memory improves processing efficiency

[1] It should be noted that the temporal stability the IAT and similar measures at the individual level is rather modest, with an average 1-2 month test-retest reliability of .54 according to a recent series of studies (Gawronski et al., 2017). Later in this review, we will briefly discuss the implications of this lower test-retest reliability for designing and interpreting interventions.

by re-shaping the brain to reflect the environment in which processing takes place. The obligatory and continuous nature of this re-shaping includes socially constructed co-occurrences, such as stereotypic associations of social groups with certain traits, attitudes, or concepts. As such, our position is that research on domain-general principles of human memory systems provides a useful framework through which to further our understanding of the causes, consequences, and potential solutions to negative repercussions of implicit social bias. This research can also help to inform policies that influence the extent to which implicit bias influences outcomes (Payne & Vuletich, 2018), such as the adoption by many orchestras of "blind" auditions, which has been credited with a reduction in gender bias in hiring (Goldin & Rouse, 2000).

In this spirit, we offer a review of recent and emerging points of contact between research on implicit social bias and human memory systems. We begin by discussing parallels that have been drawn between specific forms of implicit memory and specific social phenomena that can be considered aspects of implicit bias. We then discuss two lines of research related to the cognitive neuroscience of memory with applications to the social domain. The first line of research connects with the cognitive strategy of *individuation*, which involves directing attention and processing resources beyond social category memberships to focus on the characteristic features of an individual person. We describe findings on the so-called *other-race effect* in face memory, by which other-race faces are poorly remembered compared to same-race faces. Analyses of neural recordings during initial exposure to novel faces reveal ways in which very early indicators of categorization (as opposed to individuation) can determine the fate of the memory for an other-race face, thus limiting opportunities for individuation. The second line of research will concern the potential utility of *counter-stereotype training* through repeated exposure to members of stereotyped groups in association with counter-stereotypical concepts. Previous attempts to counter implicit social biases using this type of training have found success, but only in the very short-term. Recent developments in the study of memory consolidation, and new insights about the role of sleep in consolidation, have spawned novel implications with respect to enhancing training efforts.

More generally, the hidden causes of our behavior need not remain in the dark. Long-standing habits are not easily changed, particularly in the face of cultural, cognitive, and structural barriers to egalitarianism. Still, we suggest that by achieving a better understanding of the relevant neurocognitive mechanisms, we can ultimately be more proactive in aligning our thoughts and behaviors with our values.

Implicit Social Bias as an Expression of Multiple Memory Systems

The empirical study of implicit memory can be tied to an extended human history of thinking along these lines (reviewed in Schacter, 1987). In the mid-20th century, systematic neuroscientific explorations began to pursue the notion that these implicit phenomena reflect the workings of cognitive and neural systems that are separate from those responsible for explicit memory. A watershed moment was the discovery of memory dissociations in patient H.M. (Scoville & Milner, 1957), whose bilateral hippocampal resection left him profoundly amnesic, while traces of past experience still registered in other ways that were consistent with the operation of dissociable implicit memory systems.

Here we provide a brief review of how priming, one manifestation of implicit memory, can contribute to the acquisition and expression of social biases (see also Amodio & Ratner, 2011, for additional parallels between implicit memory systems and social phenomena). *Priming* in this context refers to altered processing due to prior experience, regardless of whether or not that experience is consciously remembered. In typical priming experiments that were used to show preserved implicit memory in patients with circumscribed amnesia, a specific prior experience is shown to increase the fluency of processing for subsequently encountered information with overlapping features. Often, this fluency leads to processing that is faster and more accurate for primed relative to unprimed information. For example, the amnesic patients described by Paller and colleagues (1991) were better able to identify briefly flashed words when those words had been shown previously in the experiment, even though their conscious memory for having seen the words was severely impaired (**Figure 1**). Additional results showed that these patients were able to attribute their increased fluency to another factor, duration of word presentation. The critical take-home message is that priming was able to influence behavior even though conscious retrieval of the specific prior experiences was severely reduced.

By extension, the behavior of healthy individuals would similarly be subject to influences from priming even when those influences remained hidden from our conscious experience or intentions. The types of fluency that give rise to priming are multi-faceted and can include both perceptual fluency (facilitated processing of percepts such as word forms or visual features), and fluency with meaning-based or conceptual information. Implicit biases as expressed on an IAT can be seen, at least in part, as a form of conceptual priming, in which learned stereotypes of group members are activated and facilitate processing of stereotype-congruent concepts (**Figure 2**). Indeed, in the extensive literature on implicit racial bias there are numerous reports of bi-directional priming between concepts strongly associated with particular social groups and individual members of these social groups, sometimes with serious downstream consequences. For example, Eberhardt, Goff, Purdie, and Davies (2004) found that exposure to Black faces reduced the perceptual threshold of clarity needed to recognize crime-relevant objects embedded in noise, suggesting that crime-related information was processed more fluently due to stereotypic links between Black individuals and crime. Importantly, this phenomenon has also been linked to more frequently mistaking harmless objects for weapons in the presence of Black individuals (e.g., Payne, 2006), which has been implicated in the unnecessary escalation of conflicts between police officers and Black individuals.

The priming of stereotypes can also color interpersonal judgments and more subtle types of interpersonal interactions. Banaji, Hardin, and Rothman (1993) found that female, but not male targets were judged as more dependent after exposure to prime words relevant to dependence, a stereotypically female trait. By contrast, male, but not female targets were judged as more aggressive after exposure to prime words related to aggression, a stereotypically male trait. Greenwald and Banaji (1995) argued that this pattern reflects the well-documented tendency of individuals to rely on processing fluency as a metacognitive cue when making a range of judgments and decisions (Alter & Oppenheimer, 2009; Hertwig, Herzog, Schooler, & Reimer, 2008; Whittlesea, Jacoby, & Girard, 1990). In other words, people tend to mistakenly attribute fluent processing to some characteristic of the target, rather than appreciating that the source of the fluent processing came from the prime. This phenomenon parallels Jacoby and colleagues' pioneering work on the use of fluency heuristics during recognition memory decisions for words (Jacoby & Whitehouse, 1989). In these studies, test cues were preceded by a prime word that was either the same as or different from the word to be tested, thus creating an increased fluency for words that were primed, and leading to a higher level of "old" endorsements for those words. The key extension of this finding from Banaji and colleagues (1993) was that conceptual fluency is more likely to be misattributed to a social target when the concept is stereotype-congruent.

In both of the above examples, bidirectional priming between stereotypic concepts and social group members illustrate that social biases are self-perpetuating (e.g., they can continue and even strengthen in an individual's mind in the absence of external reinforcement). There is also evidence that the relationship between attitudes and *behaviors* is bidirectional, meaning that engaging in actions that signal warmth or positivity toward a target can induce behavior-congruent attitudes. For example, Cacioppo, Priester, and Berntson (1993) found that participants reported liking Chinese idiographs more when instructed to assume bodily positions associated with approach (arm flexion) rather than avoidance (arm extension). These results are consistent with the engagement of implicit memory in conjunction with motor processes in attitude formation. Likewise, Van den Bergh, Vrana, and Eelen (1990) found that typists, but not nontypists, evaluated images of easy-to-type letter combinations more favorably than difficult-to-type combinations, despite having no awareness of the basis for their preference.

This link between motor processing and affect also applies to social preferences, as evidenced by facilitated approach- and avoidance-like movements in response to images of ingroup and outgroup members, respectively (Paladino & Castelli, 2008). Accordingly, an intriguing recent line of research has focused on changing implicit social attitudes by repeatedly pairing simple "approach" motor behaviors with images of racial outgroup members. In Kawakami, Phills, Steele, and Dovidio (2007), White participants showed reduced anti-Black bias on an IAT after a training session that involved pulling a joystick toward oneself (indicating approach) in response to images of Black faces, and away from oneself (indicating avoidance) in response to images of White faces. Strikingly, positive effects of this training were also apparent in certain non-verbal behaviors (decreased proximal distance and more partner-focused body orientation) displayed when participants interacted with a Black confederate. This pattern implicates stimulus-response contingencies within sensorimotor systems as another potential point of entry into social bias. More generally, findings such as these illustrate that the utility of deliberate cross-comparisons of the research on implicit social bias and implicit memory can go beyond explanations for social phenomena to reveal potential solutions.

Limiting Implicit Bias Through Individuation

While some research focuses on changing implicitly-held stereotypes and attitudes, a complementary endeavor has been to determine whether stereotype activation can be willfully suppressed during intergroup interactions. The literature paints a somewhat complicated picture. Like other forms of thought suppression (e.g., as in the "white bear" problem; Wegner, Schneider, Carter, & White, 1987), attempts to suppress stereotype-based thoughts can lead to rebound effects, in which the stereotypic information becomes more rather than less accessible (Macrae, Bodenhausen, Milne, & Jetten, 1994; Macrae, Bodenhausen, Milne, & Wheeler, 1996; but see Monteith, Lybarger, & Woodcock, 2009, for a discussion of moderators of these rebound effects).

Here we will focus on the allied strategy of *individuation*, or deliberately focusing on the unique qualities of each individual, as a means to limit the extent to which stereotype information colors person perception and memory. Individuation is generally positioned on a continuum with categorization, which occurs when an individual is perceived primarily as a member of a specific social group. However, individuation as a strategy can be seen as distinct from stereotype- or category-suppression techniques, in that the focus of individuation is on desirable rather than unwanted outcomes of person processing (e.g., as in "concentration versus suppression" in thought control; Wenzlaff & Bates, 2000).

The well-known other-race effect (ORE) in face memory is considered to be a behavioral manifestation of failures to individuate OR faces. The empirical study of the ORE has a long history rife with debates over its underlying causes (Meissner & Brigham, 2001). Traditionally, a distinction has been drawn between: 1) *perceptual expertise* accounts, which suggest that a lack of interracial contact, particularly in racial majority members, results in perceptual systems that are poorly tuned to the physical dimensions along which faces tend to differ from one another (Chiroro & Valentine, 1995; Walker & Hewstone, 2006), and 2) *social-cognitive accounts*, which point to factors such as a lower motivation to individuate OR faces and a tendency to instead focus on race-specifying information (Hugenberg & Sacco, 2008; MacLin & Malpass, 2003). By contrast, many contemporary accounts of the ORE are “hybrid” models that posit *perceptual-social linkages*, or interactions between perceptual and social elements of face processing (Anzures, Quinn, Pascalis, Slater, & Lee, 2013; Hugenberg, Young, Bernstein, & Sacco, 2010; Wan, Crookes, Reynolds, Irons, & McKone, 2015; Young & Hugenberg, 2012). Such models count perceptual expertise as one of many factors that impact the extent to which a face is initially encoded in a way that is distinct from others of the same social category, which, in turn, may impact the activation of social stereotypes. Similarly, because face recognition is a skill that facilitates social interactions, poor OR face recognition may reduce the number of positive, meaningful interactions that occur for OR individuals, further handicapping learning within both perceptual and social systems.

As such, the ORE should not be viewed as a perceptual phenomenon that is epiphenomenal to social biases. Instead, the ORE (and antecedent failures of individuation) constitutes yet another way in which such biases are self-perpetuating. An implication of this line of reasoning is that research aimed at understanding the mechanisms of the ORE—and factors that help to reduce it—can be included as part of comprehensive efforts to contend with social bias.

In theory, at any point along a social interaction, or a series of social interactions, there is the potential for processing to shift away from individuation toward categorization, or vice versa. However, research using event-related potentials (ERPs) has revealed that social-category membership permeates the very earliest stages of face processing, affecting how attention is directed to physiognomic information within a few hundred milliseconds of when a face is first encountered (see Ito & Senholzi, 2013, for additional review). Understanding these very early signatures of individuation may be particularly important, because this knowledge can point us toward effective intervention strategies that are likely to influence sufficiently early perceptual processes, rather than later stages by which time OR face encoding may have already failed.

In many ERP studies of face perception, neural processing first diverges for SR and OR faces within the time window of the face-sensitive N170. The N170 is a negativity that peaks at occipitotemporal electrodes between ~140-180 ms and is more pronounced over right- than left-hemisphere scalp sites. Activity manifested by the N170 is thought to reflect very early stages of structural encoding of faces, in which the emphasis is on face detection (e.g., extraction of features that categorically distinguish faces from non-faces) rather than on individuation (Bentin & Deouell, 2000; Eimer, 2000). Several studies (Caharel et al., 2011; He, Johnson, Dovidio, & McCarthy, 2009; Herrmann et al., 2007; Stahl, Wiese, & Schweinberger, 2008; Walker, Silvert, Hewstone, & Nobre, 2008; Wiese, Kaufmann, & Schweinberger, 2014) have reported larger N170 amplitudes for OR relative to SR faces, potentially indicating less efficient structural encoding for OR faces. This interpretation is complicated, however, by the fact that N170 modulations by race are not always found. Other studies have report either null findings with respect to SR/OR status (Caldara et al., 2003; Caldara, Rossion, Bovet, & Hauert, 2004; Ito, Thompson, & Cacioppo, 2004), influences on N170 latency but not amplitude (Gajewski, Schlegel, & Stoerig, 2008), or even larger N170 amplitudes for SR faces (Ito & Urland, 2005). Specifying the boundaries and functional significance of race effects on N170 potentials therefore awaits further research. Nonetheless, there is reason to believe that this component is not a strong source of information about the amount of perceptual individuation afforded to specific faces (Scott, Tanaka, Sheinberg, & Curran, 2006).

By contrast, two brain potentials in particular—an occiptotemporal N250 and a frontocentral N200[2]—have been identified as candidate markers of the preferential early individuation of faces belonging to ingroup members. The N250 is a bilateral negative component that peaks ~250 ms after the onset of a visual stimulus, such as a face or object, and is particularly pronounced for stimuli for which the perceiver has expertise differentiating at the subordinate level (e.g., familiar faces, or birds for expert birdwatcher, Scott et al., 2006). Similarly, the magnitude of the N250 to faces of a given race appears to be sensitive to the amount of perceptual experience that the perceiver has distinguishing among faces belonging to the target race. For example, Tanaka and Pierce (2009) tested the effectiveness of a training procedure designed to provide practice differentiating other-race faces at the individual level. The training was successful in improving OR face recognition, and the magnitude of this improvement correlated across subjects with the magnitude of increases in N250 amplitudes for OR faces.

The frontocentral N200 (**Figure 3**) is a midline negative component that is also larger (more negative) for SR than for OR faces (Dickter & Bartholow, 2007; Kubota & Ito, 2007; Lucas, Chiao, & Paller, 2011). The N200 appears to be functionally dissociable from the N250, in that the former is sensitive to race-based attention biases that cannot be directly attributed to perceptual expertise. For example, Willadsen-Jensen and Ito (2008) examined N200 amplitudes during face viewing in Asian-American participants with extensive exposure to both White and Asian faces. Interestingly, N200 amplitudes in these participants were greater for Asian compared to White faces during blocks that were majority-Asian, and greater for White faces in blocks that were majority-White. These findings suggest that Asian-American participants were able to flexibly modulate their attention to individuating information according to the demands of the context, with consequences for brain activity that occurred only 200 ms after the onset of a face. Similarly, both Black and White faces yielded greater N200 amplitudes in White participants when those faces were preceded by descriptions of stereotype-incongruent versus stereotype-congruent behaviors, the former of which may have encouraged individuation (Dickter & Gyurovski, 2012).

Lucas and colleagues (2011) directly related N200 attentional biases in White participants to the ORE by using the *subsequent-memory technique* to shed light on how encoding failures for OR faces differed qualitatively from those for SR faces. The subsequent-memory technique involves comparing brain activity during initial encoding for stimuli that are subsequently remembered versus stimuli that are subsequently forgotten (Paller & Wagner, 2002).These comparisons reveal differential neural activity based on memory (*Dm effects*), which index encoding operations that are pivotal in determining whether a specific stimulus will be retained in long-term memory. For SR faces—as for other types of stimuli such as words and meaningful pictures—Dm takes the form of a widespread positivity starting around 400 ms after stimulus onset, in which ERPs are more positive for subsequently remembered than for subsequently forgotten faces (**Figure 3c**).

This phenomenon is thought to result from elaborative encoding, by which information extracted from the current stimulus becomes meaningfully integrated with other knowledge (Wagner,

[2] It is important to note that, while investigations of N170 and N250 ERPs most often use average references, N200 effects are generally observed in studies that use mastoid references. When data are processed using an average reference, N200 effects may partially or entirely be reflected in an occipitotemporal P2 component that also tends to be sensitive to SR/OR status (see Lucas et al., 2011, for additional discussion).

Koutstaal, & Schacter, 1999), consistent with the well-known beneficial effects of "deep" processing on memory. Interestingly, behavioral evidence (Rhodes, Locke, Ewing, & Evangelista, 2009; Stahl, Wiese, & Schweinberger, 2010) suggests that depth-of-processing at encoding is less effective for OR and SR faces, perhaps because it targets encoding processes that are too far along to "correct" for deficiencies in perceptual individuation. Consistent with this interpretation, while the late positive Dm effects in Lucas et al. (2011) were also found for OR faces, they emerged approximately 400 ms later relative to that for SR faces and were reduced in magnitude[3].

Another key finding from Lucas et al. (2011) was that for OR faces, but not for SR faces, an earlier subsequent memory effect occurred on N200 potentials. Specifically, as shown in Figure 3c, N200 potentials were more negative for OR faces that were remembered relative to those that were forgotten, suggesting that the fate of OR face memory hinged on a very early processing stage associated with individuation. Moreover, analyses based on subsequently collected norms from the faces indicated that, within OR faces, N200 potentials were greater for faces judged to have a race-atypical appearance versus a race-typical appearance. These findings suggest that, not only do early stages of perceptual individuation fail more often for OR faces, but such failures preclude the effective use of downstream memorization strategies that may be related to conceptual or social individuation, thus shifting processing toward categorization and away from individuation.

One important implication of studies such as these is that interventions that focus on developing perceptual expertise with OR faces should be taken seriously as a potential avenue for reducing social biases. Thus far, face individuation as an area of study has proceeded largely separately from studies that target social bias reduction, so there is little direct evidence of how training that targets one of these phenomena can influence the other. In accordance with aforementioned perceptual-social linkage theories, there are now a handful of studies in which such training has led to concomitant reductions in implicit biases. The first of these studies (Lebrecht, Pierce, Tarr, Tanaka, & Ochsner, 2009) used White American adults as participants, and employed a five-session training protocol that focused on individuating Black faces (or, in a control condition, categorizing Black faces by race). The magnitude of the ORE decreased from pre- to post-training for the individuation group only. Most importantly, the magnitude of change in the ORE condition showed a significant relationship across subjects with the magnitude of reduction in implicit anti-Black bias. Similar results have been obtained in pre-school aged children (Qian et al., 2017b; Xiao et al., 2015), with one study finding evidence that two training sessions 1 week apart led to a reduction in bias that lasted at least 70 days (Qian et al., 2017a).

Other researchers have raised the question of whether perceptual training per se is necessary in order to reduce the ORE, pointing to findings in which simply taking steps to re-define the ingroup along dimensions orthogonal to race can influence face memory. For example, Hehman, Mania, and Gaertner (2010) found that when Black and White faces were grouped by university affiliation (University of Delaware faces in one part of the screen and James Madison University faces in another, regardless of race), participants had superior memory for own-

[3] It bears mention that a second investigation of subsequent memory effects for SR and OR faces reported an opposite pattern, in which late positive Dm effects were larger for OR than SR faces (Herzmann, Willenbockel, Tanaka, & Curran, 2011). However, this finding was specific to encoding processes that predicted subsequent recollection as opposed to accurate recognition in general. Moreover, these data seem somewhat atypical in that ERPs for SR and OR faces did not differ in any of the early components discussed here (N170, N200, N250), making it difficult to situate these results within the larger literature.

university faces of either race (see also Van Bavel & Cunningham, 2012). Future research is necessary to determine whether this “in-group reconfiguration” permeates very early stages of face individuation, as reflected by frontocentral N200 brain potentials.

Finally, it is important to note that face perception and recognition are far from the only domains in which relationships between increased individuation and reduced social biases have been found. For example, a recent study by Rubinstein, Jussim, & Stevens (2018) examined how individuating details about non-depicted college applicants whose names sounded either White or Black influenced judgments of competence and intelligence. If no individuating information was provided—or if that information was not particularly diagnostic of competence or intelligence—applicants with Black-sounding names were judged more negatively on explicit and implicit measures than applicants with White-sounding names. However, this stereotype bias was eliminated when highly diagnostic individuating information was provided. These and related findings (Cao & Banaji, 2016) underscore the fact that face processing is only one of many potential avenues by which to decrease implicit bias during tasks that require social perception and judgments.

Thus far we discussed ways in which implicit biases are formed, reinforced, and expressed through the workings of multiple memory systems. However, the extent to which information is retained in memory depends on more than processing during initial encoding. Indeed, new information is likely to be forgotten without some further processing that is tied to a stage of memory termed consolidation.

Countering Implicit Bias by Hacking into Memory During Sleep

Consolidation refers to the process that counters forgetting by gradually transforming newly encoded memories to become long-lasting and integrated with other memories (Squire, Genzel, Wixted, & Morris, 2015). Consolidation of declarative memories at a neural-systems level is thought to involve hippocampal-neocortical interaction. In particular, repeated reactivation of the same distributed cortical circuits that contributed to initial encoding seems to drive the consolidation process by allowing links to be formed and strengthened. Although consolidation is most often discussed in relation to declarative memory, systems-level changes can also influence learning within implicit memory systems.

The history of research implicating sleep in memory processing is extensive (Oudiette & Paller, 2013; Rasch & Born, 2013). The observation that hippocampal place cells replay learning-related firing patterns during sleep (e.g., Wilson & McNaughton, 1994) prompted the hypothesis that memory traces are reactivated during sleep. Consequently, an expansion of this idea advanced further in recent studies is that such reactivation contributes fundamentally to consolidation (Paller, Antony, Mayes, & Norman, in press). Whereas it was long assumed that REM sleep and concomitant dreaming would be at the heart of memory change during sleep, mounting empirical evidence now points instead to slow-wave sleep (SWS). Slow oscillations during sleep may be critical due to their role in time-locking thalamo-cortical spindles, which in turn coordinate the timing of hippocampal sharp wave ripples (Staresina et al., 2015). These nested oscillations conceivably provide a vehicle for the informational interactions across brain regions that are necessary for memory consolidation at a system level.

The study of memory reactivation during SWS has thus helped forge substantial inroads into questions about the mechanisms of memory consolidation, which information is favored for reactivation, and how control can be exerted over the operation of offline memory consolidation. A watershed moment was the discovery by Rasch and colleagues (2007) that sensory reactivation could be used as a tool in this regard. These researchers developed a procedure that involved presenting an odor during explicit learning of a set of object locations, thus establishing an association through which the odor could serve as a reactivation cue for the learning context. In some participants, this odor was then re-presented during overnight periods of SWS. Performance on a subsequent memory test revealed improved object-location memory for participants who received the odor during sleep relative to participants who did not, suggesting that the odor-driven reactivation was indeed sufficient to promote the consolidation of associated spatial memories.

This discovery by Rasch and colleagues ushered in the modern era of manipulating memory processing during sleep, inviting numerous investigations into the nature and boundaries of the phenomenon. Importantly, research by our group and others suggests that sensory reactivation can be used in a highly targeted manner, effectively “singling out” specific memories for consolidation over others even when the preferred and nonpreferred memories occurred within the same spatiotemporal context. For example, in the study reported by Rudoy, Voss, Westerberg, and Paller (2009), we modified Rasch’s paradigm such that, rather than pairing an odor with an entire learning context for object-location associations, we paired a unique sound with *each* learned association (**Figure 4**). By playing only a subset of these sounds during a post-encoding period of SWS, we were able to selectively improve subsequent memory for cued relative to uncued object-location associations.

Subsequent studies have confirmed the reliability of this phenomenon, which we have termed *targeted memory reactivation* or TMR (Cellini & Capuozzo, 2018; Creery, Oudiette, Antony, & Paller, 2015; Schouten, Pereira, Tops, & Louzada, 2017; Vargas, Schechtman, & Paller, 2018). Of particular relevance, there is evidence that the potential utility of TMR extends beyond spatial memories, and may encompass types of implicit learning that have been implicated in aspects of social bias (Antony, Gobel, O’Hare, Reber, & Paller, 2012; Batterink, Oudiette, Reber, & Paller, 2014; Batterink & Paller, 2017; Honma et al., 2016; Hu et al., 2015).

One example comes from the findings of Antony and colleagues (2012) in which auditory cues during SWS enhanced the learning of a specific motor skill, producing a melody by pressing keys in time with repeating sequences of moving circles. This form of learning reflects plasticity within sensorimotor systems, by which conjunctions of sensory cues become linked with specific movements. Interestingly, sensorimotor learning has also been implicated in the form of “approach-based” counter-bias training employed by Kawakami and colleages (2007, see also Amodio & Ratner, 2011; Cacioppo et al., 1993, for related dicussion). While such sensorimotor practice may differ mechanistically in some ways from the skill learning task used by Antony et al. (2012), the potential for TMR to amplify counter-bias training presents an intriguing possibility for future research.

Indeed, in a recent study (Hu et al., 2015) we obtained direct evidence that manipulating memory processing during sleep can facilitate a type of counter-bias training that relies on exposure to counter-stereotypical information **(Figure 5**). In this type of training, participants practice selectively responding to stereotype-incongruent face-word pairings (e.g., female faces presented with science-related words, or Black faces presented with positive words), while withholding responses to stereotype-congruent pairings (male faces presented with science-related words; White faces with positive words). Similar training procedures have been shown to be effective in reducing implicit social biases when measured immediately after training, likely because the training reinforces counter-stereotypical associations and renders them more accessible (Gawronski & LeBel, 2008; Olson & Fazio, 2006). However, the results of this and similar bias-reduction methods tend to fade disappointingly quickly. Indeed, the findings of a recent large-scale, multisite study (Lai et al., 2016) suggests that even strong immediate effects tend to dissipate within 1-2 days.

Lai and colleagues (2016) put forth several explanations for the fleeting nature of these training effects, ranging from rapid post-experiment re-enculturation to inherent limits in the malleability of implicit preferences. Given that the long-term viability of these training effects likely depends on systems consolidation, an intriguing additional consideration is that the durability of counter-bias training effects can be enhanced by taking steps to promote consolidation.

To investigate this question, Hu et al. (2015) required participants to complete baseline IATs for both the Black-White racial prejudice IAT and a gender stereotype IAT that tests preferential associations of males with science and women with art, respectively, followed by counter-training procedures targeting both of these biases. To reinforce the associations between counter-bias training and the sound cues, participants also completed a second training task, in which sound cues were presented to prompt participants to create counter-bias pairings by dragging an appropriate exemplar face to the word that corresponded to the designated counter-stereotype association. As with other studies of TMR, a key design feature was the pairing of a unique sound with each of the two counter-bias training contexts in order to establish reactivation cues for use during SWS in a post-training nap. Participants were randomly assigned to receive only one of the two sounds (e.g., either the sound associated with the gender bias counter-training or the sound associated with the race bias counter-training). In this way, we were able to separate contributions of TMR to training effects from those of the nap per se, given that sleep after training might be expected to facilitate *all* learning from the training session.

As in previous studies, counter-bias training led to a reduction in both social biases immediately following training. Of most interest was the comparison of the targeted and untargeted bias levels on IATs administered following sleep with TMR. Bias levels for the targeted bias were further reduced relative to before the nap, whereas bias levels for the non-targeted bias were relatively unaffected. These results are consistent with the notion that sound-induced reactivation of the learning context in which counter-training occurred served to amplify the effects of that training. Perhaps most strikingly, a follow-up visit one week later revealed that TMR also increased the longevity of the counter-bias training effects.[4] The effects at the 1-week delay were clearly weaker than the effects immediately after sleep, suggesting that one session of training followed by TMR may still be insufficient if longer-lasting effects are desired.

Considerable additional work is needed to explore the possibilities brought up by these initial findings, as well as to probe the boundaries of this phenomenon. For example, it remains to be

[4] This effect was found when comparing pre- versus post-nap levels of differential bias. The comparison of post-nap to baseline levels (e.g., prior to initial counter-bias training) was not significant. This nonsignificant finding was seized upon by Aczel, Palfi, Szaszi, Szollosi, & Dienes (2015), but that criticism is obviated by the finding of a significant interaction in the comparison with the pre-nap scores. Furthermore, Aczel et al., (2015) attempted to argue that the Hu et al. (2015) results were inconclusive based on this issue with respect to the findings after 1 week, but the results obtained shortly after waking confirm that training results were indeed influenced by memory reactivation during sleep. Whereas the sample size used by Hu et al. (2015) was larger than in most TMR studies in the literature, data from additional experimentation would still be helpful for deepening our understanding of the many relevant factors that may be operative in such circumstances. There is much yet to be learned about the relevant neural mechanisms and about potential applications of these methods.

determined whether repeated training or regular TMR sessions can further extend the effects of this counter-training, whether effects transfer to relevant behavior outside the laboratory, or whether other prosocial interventions are amenable to TMR. Presumably, this training technique and others like it cannot be used to reduce bias in individuals who do not want to change. If a participant maintains explicit biases while going through the training, for example, those biases rather than the training associations could be tied to the cue and strengthened during sleep. Accordingly, memory processing during sleep may be best conceptualized as a means of reinforcing prior training, such that the foremost requirement is for a program of counter-bias training that is robustly effective.

A general conclusion from this research is that the consolidation phase of learning warrants greater attention when designing interventions aimed at long-term bias reduction. The study by Hu and colleagues (2015) underscores the dependence of bias formation and bias reduction on principles of memory processing, which manifest both during wake and sleep. Importantly, memory consolidation can be influenced by several factors other than sensory input. For example, information that is believed to hold future relevance tends to receive preferential consolidation over information that is not (Bennion, Payne, & Kessinger, 2016; Wilhelm et al., 2011). The well-established benefits of engaging in spaced rather than massed practice on long-term learning has also been linked to spaced practice being more favorable to consolidation (Savion-Lemieux & Penhune, 2005). Keeping these principles in mind may help to optimize counter-bias training, particularly given previously reported discrepancies between immediate and long-term effectiveness.

Where do we go from here?

Engaging thoughts, making decisions, and committing actions — all three of these can entail changes in the storage of information in the brain. The guiding mission of memory research is to understand neurocognitive processing that underlies this pervasive neuroplasticity, as well as how learning drives subsequent thoughts and behaviors. As theories of learning and memory have advanced, they have both informed and benefitted from parallel work in social psychology and particularly social cognition.

In keeping with this tradition, we have here examined two areas of recent progress with applicability across these fields of study. One area focuses on the ways in which social categorization permeates very early stages of face processing, thereby detracting from the extent to which individuating information is processed and stored in memory. Whereas the other-race effect as a phenomenon is hardly a new discovery, recent years have witnessed a shift away from isolated investigations of hypothesized perceptual and/or social causes and toward the study of perceptual-social linkages—ways in which experience-based perceptual processing advantages for same-race faces can interact in a bidirectional manner with tendencies to categorize outgroup faces by race (Lee, Quinn, & Pascalis, 2017). These developments call for interventions to be tested that consider both perceptual and social aspects of person perception. Moreover, high temporal-resolution methods such as ERPs can be used to provide insights into which stages of face processing and memory are relevant.

We have also reviewed evidence that processing that transpires offline—and particularly during sleep— may influence long-term outcomes of bias reduction interventions. The large-scale synchrony across brain regions during slow-wave sleep provides a mechanism by which areas of the brain communicate to strengthen newly-learned associations, rendering them more stable and integrated into long-term memory stores. The fact that TMR was able to strengthen and stabilize counter-bias training effects further underscores that these biases can be learned and unlearned through fundamental memory processing. Moreover, these findings point to memory consolidation as a possible source of discrepancies between the magnitudes of immediate and delayed effects found in counter-bias training studies.

A critical issue for future work concerns the extent to which the learning that occurs during interventions may be tied to aspects of the context in which the intervention took place. As previously mentioned, a limitation of implicit association tests in general is their relatively low temporal stability (e.g., Gawronski et al., 2017). Interestingly, this within-person instability is found even on measures for which internal consistency is quite high, which is suggestive of systematic rather than random variance. On the basis of this and other findings (e.g., Gschwendner, Hofmann, & Schmitt, 2008), it has been proposed that implicit bias may be best conceptualized as the interactive influence of a person-related component—which reflects "chronic" concept accessibility due to each individual's idiosyncratic learning history—and a situational component, which reflects the concepts that are activated by a given set of inputs (Bargh, Bond, Lombardi, & Tota, 1986; Gawronski & Bodenhausen, 2017). This interactionist view can also help to explain weak relationships that have been found between implicit bias and discriminatory behavior, insofar as the bias and the behavior are measured in different contexts (see also Friese, Hofmann, & Schmitt, 2008, for other moderators of implict bias-behavior relationships).

Little is known about the extent to which the positive effects of the interventions discussed here generalize across situations with different concept accessibility. It seems reasonable to suggest that perceptual expertise training operates largely on the person component, in that the increased sensitivity to individuating features of other-race faces may reduce the chronic accessibility of social-categorical concepts during interracial interactions. Regarding counter-bias training, it is possible that bias reduction is bounded to specific contexts where the training takes place, and thus reduction may not generalize to new contexts (i.e., contextualized attitude change; Gawronski et al., 2018). On the other hand, there is mounting evidence that sleep-based memory processing promotes integration of new information into existing memory stores (Lewis, Knoblich, & Poe, 2018), which may also contribute to the generalization of counter-bias training effects. However, these speculations await additional empirical evaluation.

There are limits to how far we can take our claim that social biases can be countered by appealing to basic principles of human memory. Altering a social bias is a far more complex undertaking than, for example, learning a new motor skill or correcting a factual misconception. For one, we seem to be evolutionarily pre-disposed to making ingroup/outgroup distinctions (Kurzban, Tooby, & Cosmides, 2001). We are constantly bombarded with socially constructed affirmations of biases we may wish to unlearn. Moreover, the sheer number and diversity of memory systems that appear to house aspects of implicit bias—which include those responsible for priming, motor habits, plasticity within visual processing streams, and others (e.g., Amodio & Ratner, 2011)—could alone be reasonably expected to complicate corrective efforts.

On the other hand, principles of learning and memory are arguably rendered even more valuable in light of these complexities. Perhaps most important is the principle that we retain what we practice. Mastering any complex skill—whether in the domain of music, athletics, or the cultivation of compassion and pro-social attitudes—requires regular, intentional practice. Our brains are remarkably plastic, and reducing discrepancies between our values and our implicit knowledge requires a sustained and proactive approach to harnessing and managing this plasticity.

Finally, it bears mention that grappling with implicit social bias must involve more than challenging individual biases. Although some social biases can be helpful in navigating the world, negative stereotypes result in systemic psychological, physical, and financial harm. An important benefit of understanding these biases is to aid in the development of policies and interventions that acknowledge this reality.

References

Aczel, B., Palfi, B., Szaszi, B., Szollosi, A., & Dienes, Z. (2015). Commentary: Unlearning implicit social biases during sleep. *Frontiers in Psychology*, *6*, 1428.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*, 219–235.

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*, 143–148.

Antony, J. W., Gobel, E. W., O'Hare, J. K., Reber, P. J., & Paller, K. A. (2012). Cued memory reactivation during sleep influences skill learning. *Nature Neuroscience*, *15*, 1114–6.

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., & Lee, K. (2013). Development of own-race biases. *Visual Cognition*, *21*, 1165–1182.

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*, 272–281.

Bargh, J. A., Bond, R. N., Lombardi, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology*, *50*, 869–878.

Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia*, *65*, 169–79.

Batterink, L. J., & Paller, K. A. (2017). Sleep-based memory processing facilitates grammatical generalization: Evidence from targeted memory reactivation. *Brain and Language*, *167*, 83–93.

Bentin, S., & Deouell, L. Y. (2000). Structural encoding and identification in face processing: ERP evidence for separate mechanisms. *Cognitive Neuropsychology*, *17*, 35–55.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology*, *65*, 5–17.

Caharel, S., Montalan, B., Fromager, E., Bernard, C., Lalonde, R., & Mohamed, R. (2011). Other-race and inversion effects during the structural encoding stage of face processing in a race categorization task: An event-related brain potential study. *International Journal of Psychophysiology*, *79*, 266–271.

Caldara, R., Rossion, B., Bovet, P., & Hauert, C.-A. (2004). Event-related potentials and time course of the "other-race" face classification advantage. *Neuroreport*, *15*, 905–10.

Caldara, R., Thut, G., Servoir, P., Michel, C. M., Bovet, P., & Renault, B. (2003). Face versus non-face object perception and the "other-race" effect: A spatio-temporal event-related potential study. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *114*, 515–28.

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 7475–80.

Cellini, N., & Capuozzo, A. (2018). Shaping memory consolidation via targeted memory reactivation during sleep. *Annals of the New York Academy of Sciences*. doi:10.1111/nyas.13855

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *48*, 879–894.

Creery, J. D., Florczak, S. M., Zaheed, A. B., Antony, J. W., & Paller, K. A. (2014). An Implicit Measure of Pro-Social Attitudes: Can a Liberal Feel Equanimity toward a Conservative? In *Poster presented at the Cognitive Neuroscience Society annual meeting, Boston, Massachusetts.*

Creery, J. D., Oudiette, D., Antony, J. W., & Paller, K. A. (2015). Targeted memory reactivation during sleep depends on prior learning. *Sleep*, *38*, 755–63.

Dickter, C. L., & Bartholow, B. D. (2007). Racial ingroup and outgroup attention biases revealed by event-related brain potentials. *Social Cognitive and Affective Neuroscience*, *2*, 189–98.

Dickter, C. L., & Gyurovski, I. (2012). The effects of expectancy violations on early attention to race in an impression-formation paradigm. *Social Neuroscience*, *7*, 240–251.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing Black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*, 876–893.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain. Oxford psychology series; no. 35.*

Eimer, M. (2000). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology*, *111*, 694–705.

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, *19*, 285–338.

Gajewski, P. D., Schlegel, K., & Stoerig, P. (2008). Effects of human race and face inversion on the N170. *Journal of Psychophysiology*, *22*, 157–165.

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry*, *28*, 268–272.

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, *44*, 1355–1361.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures. *Personality and Social Psychology Bulletin*, *43*, 300–312.

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. *Advances in Experimental Social Psychology*, *57*, 1–52.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. *American Economic Review*, *90*, 715–741.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

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

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*, 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Gschwendner, T., Hofmann, W., & Schmitt, M. (2008). Differential stability: The effects of acute and chronic construct accessibility on the temporal stability of the Implicit Association Test. *Journal of Individual Differences*, *29*, 70–79.

He, Y., Johnson, M. K., Dovidio, J. F., & McCarthy, G. (2009). The relation between race-related implicit associations and scalp-recorded neural activity evoked by faces from different races. *Social Neuroscience*, *4*, 426–42.

Hehman, E., Mania, E. W., & Gaertner, S. L. (2010). Where the division lies: Common ingroup identity moderates the cross-race facial-recognition effect. *Journal of Experimental Social Psychology*, *46*, 445–448.

Herrmann, M. J., Schreppel, T., Jäger, D., Koehler, S., Ehlis, A.-C., & Fallgatter, A. J. (2007). The other-race effect for face perception: an event-related potential study. *Journal of Neural Transmission*, *114*, 951–957.

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1191–1206.

Herzmann, G., Willenbockel, V., Tanaka, J. W., & Curran, T. (2011). The neural correlates of memory encoding and recognition for own-race and other-race faces. *Neuropsychologia*, *49*, 3103–3115.

Honma, M., Plass, J., Brang, D., Florczak, S. M., Grabowecky, M., & Paller, K. A. (2016). Sleeping on the rubber-hand illusion: Memory reactivation during sleep facilitates multisensory recalibration. *Neuroscience of Consciousness*, *2016*, niw020.

Hu, X., Antony, J. W., Creery, J. D., Vargas, I. M., Bodenhausen, G. V., & Paller, K. A. (2015). Unlearning implicit social biases during sleep. *Science*, *348*.

Hugenberg, K., & Sacco, D. F. (2008). Social categorization and stereotyping: How social categorization biases person perception and face memory. *Social and Personality Psychology Compass*, 2, 1052–1072.

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, *117*, 1168–1187.

Ito, T. A., & Senholzi, K. B. (2013). Us versus them: Understanding the process of race perception with event-related brain potentials. *Visual Cognition*, *21*, 1096–1120.

Ito, T. A., Thompson, E., & Cacioppo, J. T. (2004). Tracking the timecourse of social perception: the effects of racial cues on event-related brain potentials. *Personality & Social Psychology Bulletin*, *30*, 1267–80.

Ito, T. A., & Urland, G. R. (2005). The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective, & Behavioral Neuroscience*, *5*, 21–36.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, *118*, 126–135.

Jost, J. T. (2018). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, 096372141879730.

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*, 957–971.

Kubota, J. T., & Ito, T. A. (2007). Multiple cues in social perception: The time course of processing race and facial expression. *Journal of Experimental Social Psychology*, *43*, 738–752.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*, 15387–15392.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016.

Lebrecht, S., Pierce, L. J., Tarr, M. J., Tanaka, J. W., & Ochsner, K. (2009). Perceptual other-race training reduces implicit racial bias. *PLoS ONE*, *4*, e4215.

Lee, K., Quinn, P. C., & Pascalis, O. (2017). Face race processing and racial bias in early development: A perceptual-social linkage. *Current Directions in Psychological Science*, *26*, 256–262.

Lewis, P. A., Knoblich, G., & Poe, G. (2018). How memory replay in sleep boosts creative problem-solving. *Trends in Cognitive Sciences*, *22*, 491–503.

Lucas, H. D., Chiao, J. Y., & Paller, K. A. (2011). Why some faces won’t be remembered: Brain potentials illuminate successful versus unsuccessful encoding for same-race and other-race faces. *Frontiers in Human Neuroscience*. doi:10.3389/fnhum.2011.00020

MacLin, O. H., & Malpass, R. S. (2003). The ambiguous-race face illusion. *Perception*, *32*, 249–252.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, *67*, 808–817.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Wheeler, V. (1996). On resisting the temptation for simplification: Counterintentional effects of stereotype suppression on social memory. *Social Cognition*, *14*, 1–20.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3–35.

Monteith, M. J., Lybarger, J. E., & Woodcock, A. (2009). Schooling the cognitive monster: The role of motivation in the regulation and control of prejudice. *Social and Personality Psychology Compass*, *3*, 211–226.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences*, *15*, 152–159.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*, 421–433.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192.

Oudiette, D., & Paller, K. A. (2013). Upgrading the sleeping brain with targeted memory reactivation. *Trends in Cognitive Sciences*, *17*, 142–9.

Paladino, M.-P., & Castelli, L. (2008). On the immediate consequences of intergroup categorization: Activation of approach and avoidance motor behavior toward ingroup and outgroup members. *Personality and Social Psychology Bulletin*, *34*, 755–768.

Paller, K. A., Antony, J. W., Mayes, A., & Norman, K. A. (n.d.). Replay-based consolidation governs enduring memory storage. In M. S. Gazzaniga, G. R. Mangun, & D. Poeppel (Eds.), *The Cognitive Neurosciences* (6th ed.). MIT Press.

Paller, K. A., Mayes, A. R., McDermott, M., Pickering, A. D., & Meudell, P. R. (1991). Indirect measures of memory in a duration-judgement task are normal in amnesic patients. *Neuropsychologia*, *29*, 1007–18.

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, *6*, 93–102.

Payne, B. K. (2006). Weapon Bias. *Current Directions in Psychological Science*, *15*, 287–291.

Payne, B. K., & Vuletich, H. A. (2018). Policy insights from advances in implicit bias research. *Policy Insights from the Behavioral and Brain Sciences*, *5*, 49–56.

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*, 233–248.

Qian, M. K., Quinn, P. C., Heyman, G. D., Pascalis, O., Fu, G., & Lee, K. (2017a). A long-term effect of perceptual individuation training on reducing implicit racial bias in preschool children. *Child Development*. doi:10.1111/cdev.12971

Qian, M. K., Quinn, P. C., Heyman, G. D., Pascalis, O., Fu, G., & Lee, K. (2017b). Perceptual individuation training (but not mere exposure) reduces implicit racial bias in preschool children. *Developmental Psychology*, *53*, 845–859.

Rasch, B., & Born, J. (2013). About sleep’s role in memory. *Physiological Reviews*, *93*, 681–766.

Rasch, B., Buchel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*, *315*, 1426–1429.

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*, 2026–2042.

Rhodes, G., Locke, V., Ewing, L., & Evangelista, E. (2009). Race coding and the other-race effect in face recognition. *Perception*, *38*, 232–241.

Rubinstein, R. S., Jussim, L., & Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology*, *75*, 54–70.

Rudoy, J. D., Voss, J. L., Westerberg, C. E., & Paller, K. A. (2009). Strengthening individual memories by reactivating them during sleep. *Science (New York, N.Y.)*, *326*, 1079.

Savion-Lemieux, T., & Penhune, V. B. (2005). The effects of practice and delay on motor skill learning and retention. *Experimental Brain Research*, *161*, 423–431.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 501–518.

Schouten, D. I., Pereira, S. I. R., Tops, M., & Louzada, F. M. (2017). State of the art on targeted memory reactivation: Sleep your way to enhanced cognition. *Sleep Medicine Reviews*, *32*, 123–131.

Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2006). A reevaluation of the electrophysiological correlates of expert object processing. *Journal of Cognitive Neuroscience*, *18*, 1453–1465.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *12*, 103–13.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*, 171–177.

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor Perspectives in Biology*, *7*, a021766.

Stahl, J., Wiese, H., & Schweinberger, S. R. (2008). Expertise and own-race bias in face processing: an event-related potential study. *NeuroReport*, *19*, 583–587.

Stahl, J., Wiese, H., & Schweinberger, S. R. (2010). Learning task affects ERP-correlates of the own-race bias, but not recognition memory performance. *Neuropsychologia*, *48*, 2027–2040.

Staresina, B. P., Bergmann, T. O., Bonnefond, M., van der Meij, R., Jensen, O., Deuker, L., … Fell, J. (2015). Hierarchical nesting of slow oscillations, spindles and ripples in the human hippocampus during sleep. *Nature Neuroscience*, *18*, 1679–1686.

Tanaka, J. W., & Pierce, L. J. (2009). The neural plasticity of other-race face recognition. *Cognitive, Affective, & Behavioral Neuroscience*, *9*, 122–131.

Van Bavel, J. J., & Cunningham, W. A. (2012). A social identity approach to person memory. *Personality and Social Psychology Bulletin*, *38*, 1566–1578.

Van den Bergh, O., Vrana, S., & Eelen, P. (1990). Letters from the heart: Affective categorization of letter combinations in typists and nontypists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1153–1161.

Vargas, I. M., Schechtman, E., & Paller, K. A. (n.d.). Targeted memory reactivation during sleep to strengthen memory for arbitrary pairings.

Wagner, A. D., Koutstaal, W., & Schacter, D. L. (1999). When encoding yields remembering:

insights from event-related neuroimaging. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *354*, 1307–24.

Walker, P. M., & Hewstone, M. (2006). A perceptual discrimination investigation of the own-race effect and intergroup experience. *Applied Cognitive Psychology*, *20*, 461–475.

Walker, P. M., Silvert, L., Hewstone, M., & Nobre, A. C. (2008). Social contact and other-race face processing in the human brain. *Social Cognitive and Affective Neuroscience*, *3*, 16–25.

Wan, L., Crookes, K., Reynolds, K. J., Irons, J. L., & McKone, E. (2015). A cultural setting where the other-race effect on face recognition has no social–motivational component and derives entirely from lifetime perceptual experience. *Cognition*, *144*, 91–115.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*, 5–13.

Wenzlaff, R. M., & Bates, D. E. (2000). The relative efficacy of concentration and suppression strategies of mental control. *Personality and Social Psychology Bulletin*, *26*, 1200–1212.

Whittlesea, B. W. ., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language*, *29*, 716–732.

Wiese, H., Kaufmann, J. M., & Schweinberger, S. R. (2014). The neural signature of the own-race bias: Evidence from event-related potentials. *Cerebral Cortex*, *24*, 826–835.

Wilhelm, I., Diekelmann, S., Molzow, I., Ayoub, A., Molle, M., & Born, J. (2011). Sleep selectively enhances memory expected to be of future relevance. *Journal of Neuroscience*, *31*, 1563–1569.

Willadsen-Jensen, E. C., & Ito, T. A. (2008). A foot in both worlds: Asian Americans' perceptions of Asian, White, and racially ambiguous faces. *Group Processes & Intergroup Relations*, *11*, 182–200.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science (New York, N.Y.)*, *265*, 676–9.

Xiao, W. S., Fu, G., Quinn, P. C., Qin, J., Tanaka, J. W., Pascalis, O., & Lee, K. (2015). Individuation training with other-race faces reduces preschoolers' implicit racial bias: a link between perceptual and social representation of faces in children. *Developmental Science*, *18*, 655–663.

Young, S. G., & Hugenberg, K. (2012). Individuation motivation and face experience can operate jointly to produce the own-race bias. *Social Psychological and Personality Science*, *3*, 80–87.

Figure 1. Preserved priming in amnesia (Paller et al., 1991). Memory tests were administered to a group of people with relatively circumscribed amnesia (*n* = 9) and a group of healthy individuals matched on age, intelligence, and socio-economic background (*n* = 9). In each group, the mean age was 40 years old. (**A**) In the study phase of the experiment, participants were exposed to 60 words which they read aloud. (**B**) In the priming test, participants viewed 80 words, half of which had been presented in the study phase (i.e., 40 primed words and 40 unprimed words). Each word was preceded and followed by a 500-ms mask and the word duration was adjusted to be near the threshold for reading for each individual (the mean duration was 248 ms in the amnesic group and 147 ms in the control group). Participants attempted to read each word and then rate the duration of exposure on a four-point scale based on word durations they experienced in a training phase (all durations were very brief: 1 was the briefest and 4 was the longest). (**C**) Upper panel: Word identification performance was superior for primed words compared to unprimed words, and the magnitude of this priming effect did not differ between the two groups. Lower panel: Duration estimates were longer for primed words compared to unprimed words, and the magnitude of this priming effect also did not differ between the two groups. Accordingly, both measures demonstrated preserved priming. (**D**) After the priming test was completed, each primed word appeared together with two new words in a three-alternative forced-choice recognition test. As expected, amnesics were impaired in their ability to recognize words from the study phase. The memory impairment in these patients thus compromised declarative memory, but implicit memory in these two priming tests showed no indication of impairment and likely does not require a contribution from the brain areas damaged in amnesia. These results and many other demonstrations of preserved priming in amnesia thus support the idea that different memory systems are operative in ways that vary depending on the way memory is assessed.

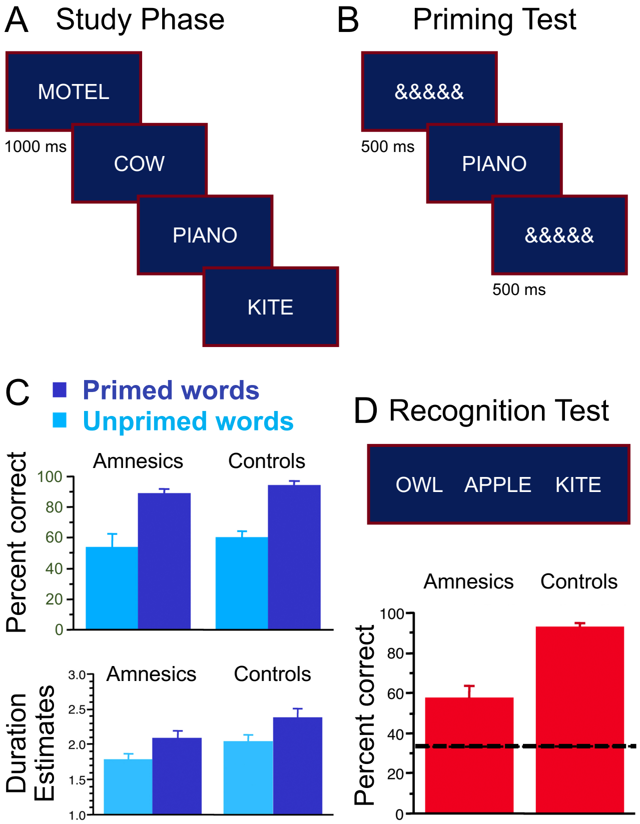
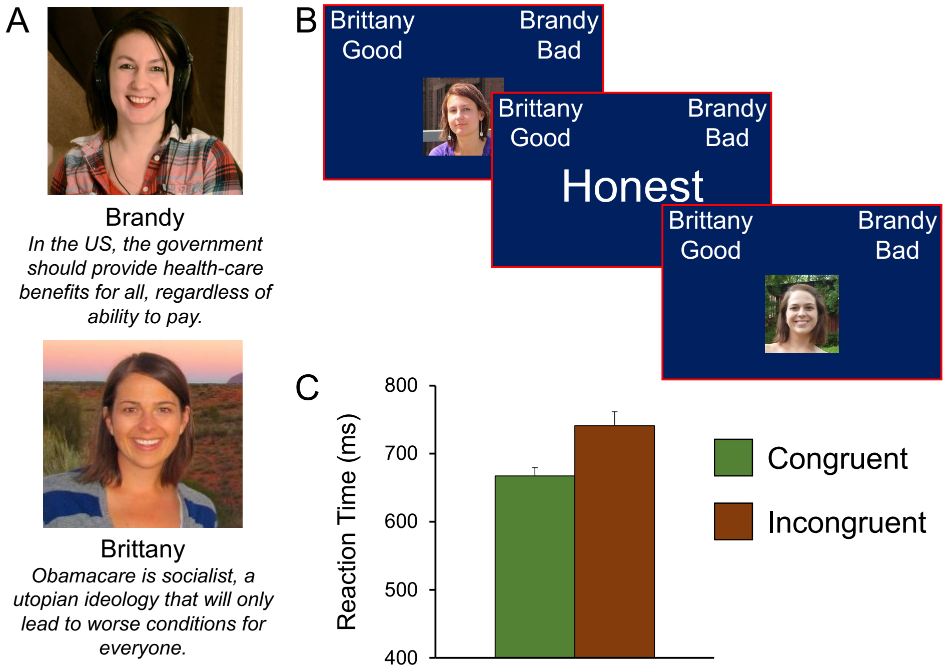


Figure 2. A novel example of an Implicit Association Test (IAT) from Creery, Florczak, Zaheed, Antony, & Paller (2014). We hypothesized that a person with strong political ideals might show implicit bias in favor of someone else with the same versus opposite political views. To assess this type of bias, we administered a variant of the IAT. (**A**) Participants (*N* = 32) first selected two political views that were personally very important (e.g., gay rights, healthcare). For each of these issues, a view expressing the same position was ascribed to one person and a sharply opposing view was ascribed to another similar-looking person. In the learning phase of the experiment, the faces of these two people were selectively linked with the corresponding political views. (**B**) These two faces were then used in the IAT along with words corresponding to positive and negative personal qualities (e.g., intelligent, rude). Button assignments differed across blocks of trials (left/right hands, counterbalanced, and shown in the upper left and right portion of the screen). In congruent blocks, participants responded using the same button for positive words and views of the individual with the same political view. In incongruent blocks, the same button was used for positive words and views of the individual with the opposing political view. The example depicted here would correspond to the congruent condition for participants who favor Brittany's view or the incongruent condition for participants who favor Brandy's view. (**C**) Reaction-time differences were found as a function of whether the response assignments were congruent or incongruent ($t_{(31)}$ = 5.96, p < .001), demonstrating that judgments were influenced by whether the political views of the person shown matched those strongly held by the participant. This IAT thus produced a pattern of behavior indicative of implicit social bias.



In the US, the government should provide health-care benefits for all, regardless of ability to pay.

Obamacare is socialist, a utopian ideology that will only lead to worse conditions for everyone.

Figure 3. Illustration of the experimental methods and results from Lucas et al. (2011). (**A**) Examples of same-race (SR/White) and other-race (OR/Black) faces presented to White participants. Participants were asked to attempt to commit all faces to memory in anticipation of a subsequent memory test. (**B**) Larger N200 potentials were elicited by SR relative to OR faces. N200 differences by race have elsewhere been linked to greater individuation of SR faces (see text). (**C**) Subsequent-memory effects (termed *Dm*) are shown for SR and OR faces. For OR faces only, N200 potentials were larger for faces that were later remembered relative to faces that were later forgotten. These results suggest that the fate of OR faces in memory hinges on a very early stage of individuation that is robust and reliable for SR faces. Cza and Pzs refer to EEG recording locations over frontocentral and centroparietal scalp, respectively.

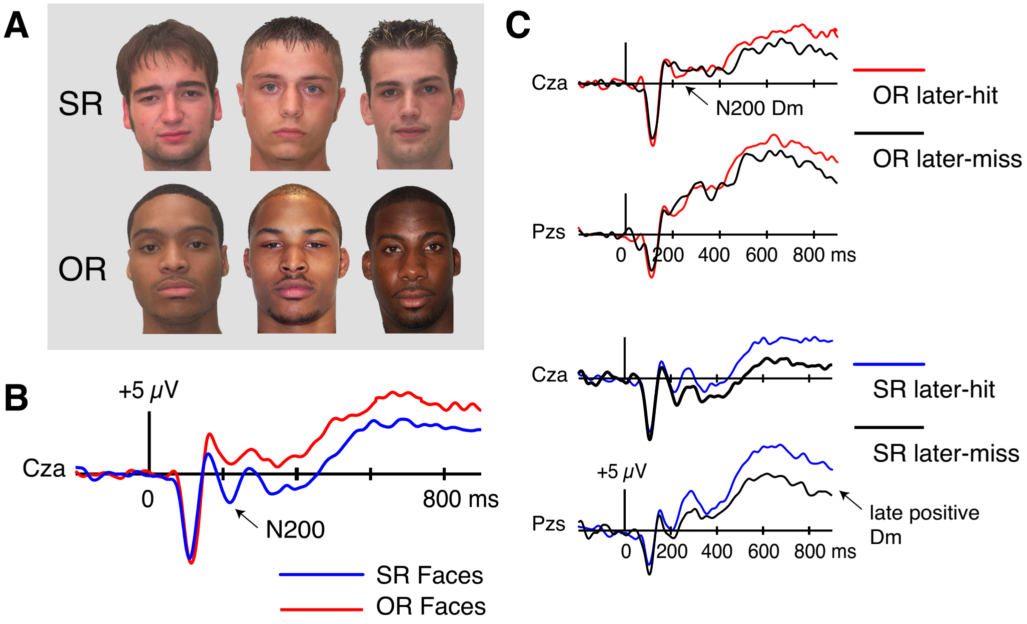


Figure 4. Targeted memory reactivation (TMR). (**A**) Participants in the study (Rudoy et al., 2009) first learned 50 object-location associations. Each object was presented with its characteristic sound. Following an interactive learning procedure, location recall was tested. Half of the objects were assigned to be cued during sleep such that recall accuracy was matched for cued and uncued objects. (**B**) Next, EEG was recorded while participants napped. When a participant reached slow-wave sleep, 25 of the sounds were presented sequentially at an intensity that was low enough such that sleep was not disrupted. (**C**) Recall of locations was tested again after the nap. Subjects moved each object from the center of the screen to the location where they believed it to have originally appeared (arrows). Recall was more accurate for cued versus uncued objects. Mean EEG responses from 400-800 ms following the onset of each sound presented during sleep were found to be more positive for those objects with less decline in recall ("Less forgetting") compared to the remaining objects or to baseline sounds. These responses resembled typical event-related potentials predictive of later memory (Dm effects; Paller & Wagner, 2002). The results thus suggest that memory reactivation occurred as a consequence of cue presentation, and that spatial recall was improved as a result of this reactivation. Figure reprinted from Rudoy et al. (2009).

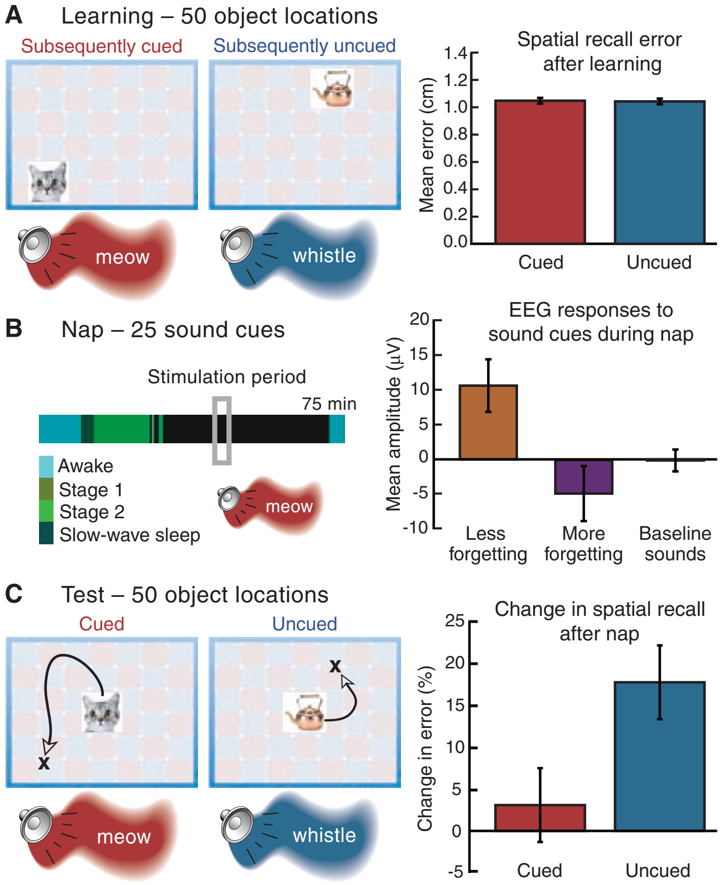
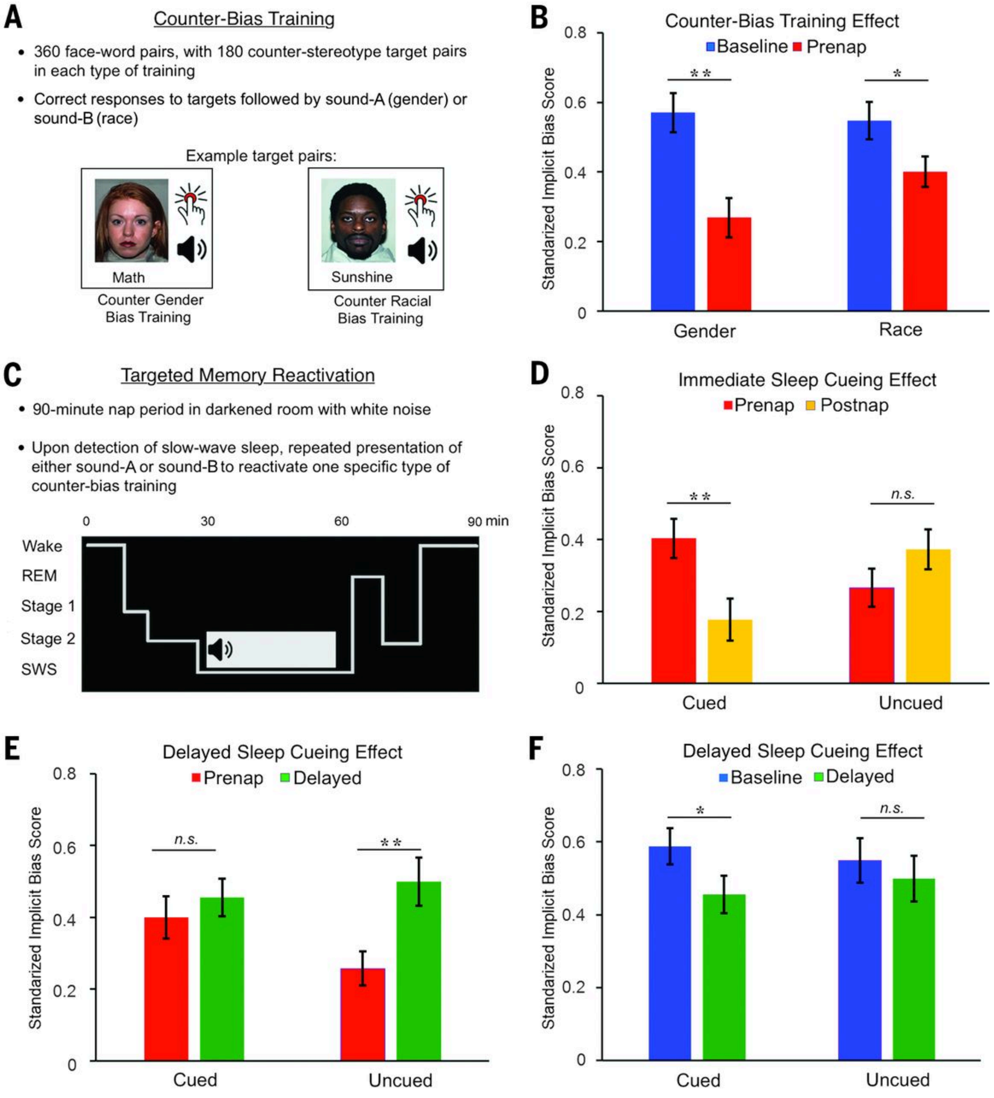


Figure 5. Targeted memory reactivation used with training to reduce implicit social bias. (**A**) Procedure for counter-bias training. (**B**) Reductions in implicit bias were found for both racial and gender conditions. Bias was measured using the Implicit Association Test before training (baseline) and after training (prenap). Error bars indicate ±1 SEM adjusted for within-subject comparisons. (**C**) Procedures for the nap phase of the experiment, when one sound was repeatedly played during SWS. (**D**) The change in implicit bias from prenap to postnap diverged as a function of cueing condition, showing a further reduction only for the cued social bias, and a significant interaction. (**E**) The change in implicit bias from prenap to the 1-week delay diverged as a function of cueing condition, showing a significant increase only for the uncued social bias, and a significant interaction. (**F**) The change in implicit bias from baseline to the 1-week delay diverged as a function of cueing condition, showing a significant reduction only for the cued social bias, and a nonsignificant interaction. Significant pairwise differences are indicated by * ($p < 0.05$) or ** ($p < 0.01$). Figure reprinted from Hu et al (2015).


A
Counter-Bias Training
• 360 face-word pairs, with 180 counter-stereotype target pairs in each type of training
• Correct responses to targets followed by sound-A (gender) or sound-B (race)
Example target pairs:
Math
Counter Gender Bias Training
Sunshine
Counter Racial Bias Training
B
Counter-Bias Training Effect
Baseline
Prenap
Standarized Implicit Bias Score
**
*
Gender
Race
C
Targeted Memory Reactivation
• 90-minute nap period in darkened room with white noise
• Upon detection of slow-wave sleep, repeated presentation of either sound-A or sound-B to reactivate one specific type of counter-bias training
90 min
Wake
REM
Stage 1
Stage 2
SWS
D
Immediate Sleep Cueing Effect
Prenap
Postnap
Standarized Implicit Bias Score
**
n.s.
Cued
Uncued
E
Delayed Sleep Cueing Effect
Prenap
Delayed
Standarized Implicit Bias Score
n.s.
**
Cued
Uncued
F
Delayed Sleep Cueing Effect
Baseline
Delayed
Standarized Implicit Bias Score
*
n.s.
Cued
Uncued