Published in final edited form as:
Science. 2015 May 29; 348(6238): 1013–1015. doi:10.1126/science.aaa3841.

Unlearning Implicit Social Biases During Sleep **

Xiaoqing Hu[1,2], **James W. Antony**[1,3], **Jessica D. Creery**[1], **Iliana M. Vargas**[1], **Galen V. Bodenhausen**[1], and **Ken A. Paller**[1,*]

[1]Department of Psychology, Northwestern University, Evanston, IL

[2]Department of Psychology, University of Texas, Austin, TX

[3]Princeton Neuroscience Institute, Princeton University, Princeton, NJ

Abstract

Although people may endorse egalitarianism and tolerance, social biases can remain operative and drive harmful actions in an unconscious manner. Here we investigated training to reduce implicit racial and gender bias. Forty participants processed counter-stereotype information paired with one sound for each type of bias. Biases were reduced immediately after training. During subsequent slow-wave sleep, one sound was unobtrusively presented to each participant, repeatedly, to reactivate one type of training. Corresponding bias reductions were fortified in comparison to the social bias not externally reactivated during sleep. This advantage remained one week later, the magnitude of which was associated with time in slow-wave and rapid-eye-movement sleep after training. We conclude that memory reactivation during sleep enhances counter-stereotype training, and that maintaining a bias reduction is sleep-dependent.

Social interactions are often fraught with bias. Our preconceptions about other people can influence many types of behavior. For example, documented policing errors have repeatedly shown the potential harm of racial profiling (1). In experiments utilizing a first-person-shooter videogame, both White and Black participants were more likely to shoot Black than White individuals, even when they held a harmless object rather than a gun (2). When hiring potential research assistants, both male and female faculty members were more likely to hire male than equally qualified female candidates (3).

Although the tendency for people to endorse racist or sexist attitudes explicitly has decreased in recent years (4), social biases may still influence people's behavior in an

**This manuscript has been accepted for publication in Science. This version has not undergone final editing. Please refer to the complete version of record at http://www.sciencemag.org/. The manuscript may not be reproduced or used in any manner that does not fall within the fair use provisions of the Copyright Act without the prior, written permission of AAAS.

*Correspondence to: Ken A. Paller, kap@northwestern.edu.

Supplementary Materials
www.sciencemag.org
Materials and Methods
Figures S1-S3
Table S1
Sound Files
References (32-35)

The authors declare no conflict of interests.

implicit or unconscious manner, despite their good intentions and perhaps beyond their conscious control (5). Ample evidence indicates that implicit biases can drive discriminatory behaviors and exacerbate intergroup conflict (5-8). For instance, implicit racial biases decrease investments given to racial out-group members in a trust game (6). At a broader level, the gender gap in science achievement in a nation is correlated with the level of implicit stereotyping of females as not having an aptitude for science (8).

Whereas discriminatory behaviors can be detrimental to individuals and society, implicit social bias can be difficult to correct due to a range of affective, cognitive, motivational, and social factors, as follows (9, 10). First, out-group members can be perceived as threatening, and the fear response to those individuals can resist extinction (11). Second, biases are acquired over many years of exposure to stereotypes and they can efficiently operate without occupying cognitive resources (5, 10). Third, motivation to seek higher status or self-enhancement commonly results in out-group derogation (9, 10). Lastly, perceived social norms can prescribe people’s expression of stereotyping and prejudice (12). Despite such challenges, implicit biases can be reduced via learning about counter-stereotype cases (13). However, benefits of this counter-bias training can be fragile, subject to reversal when the original stereotypes are again reinforced in typical circumstances, such as through the media (14). Longer-term reductions in implicit social biases may necessitate that counter-bias training be followed by further memory consolidation, as is the case for many other types of learning (15).

Recent findings suggest that memory consolidation during sleep may be essential for preserving newly acquired information such as declarative and procedural memories (16-19). During sleep, information recently stored in the brain can be integrated with other information and transformed into stable representations through a process known as systems-level consolidation (15). The mechanisms of this transformation are thought to involve repeated reactivation of information, particularly during sleep, leading to subsequent improvement in post-sleep memory performance (20-24).

Taking into consideration the role of sleep in memory consolidation, we adapted procedures for (a) reducing implicit social biases and (b) reactivating this training during sleep. We were particularly interested in factors that can influence whether such training procedures produce transient versus persistent effects. Because pervasive stereotypes in the media and broader culture could function to regenerate a bias that is momentarily reduced (14), maintaining the benefits of training is crucial for the ultimate usefulness of potential bias-reducing interventions.

We reactivated counter-bias information during sleep using subtle auditory cues that had been associated with counter-bias training. Participants were White males and females from a university community (*N*=40), and were recruited as two subsamples that allowed for a direct replication (see supplementary material, SM). First, biases were quantified using two versions of the implicit association test (IAT, 25). The IAT allows for an assessment of the strength of implicit associations between social groups and attributes (25). One test examined the degree to which female faces were preferentially associated with science versus art words, or the reverse for male faces (gender-bias IAT). The other test examined the degree to which Black faces were preferentially associated with bad versus good words, or the reverse for White faces (racial-bias IAT). Results were quantified using a conventional scoring procedure (26) — zero indicating no bias, larger scores indicating greater bias. Consistent with previous research (7), IAT scores showed that participants held implicit social biases for both gender and race, with both scores significantly greater than zero [mean 0.559 ± 0.044 (SEM); gender $t(39)=9.076$, $P<0.001$; race $t(39)=8.388$, $P<0.001$].

Following this confirmation of baseline levels of implicit bias, participants engaged in training designed to reduce gender and racial bias (13). In both cases, bias reduction was expected because participants intentionally selected counter-stereotype information intermixed with other information. Participants viewed several types of face-word pairing, but were required to attend and respond only to pairings that countered the typical bias (Fig. 1A; SM). Two unusual frequency-modulated sounds were presented during training, one following correct counter-gender-bias responses and the other correct counter-racial-bias responses. To reinforce these associations, we administered another task wherein the same two sounds prompted participants to form a corresponding face-word pairing (SM). Training thus established a strong association between each sound and one type of counter-bias training.

Biases were reduced compared to baseline levels [Fig. 1B, within-subject ANOVA, $F(1,39)=15.453$, $P<0.001$, $\eta_p^2=0.284$]. The mean IAT score was 0.559 at baseline and 0.335 at the prenap test. This bias reduction did not differ as a function of bias type [$F(1,39)=1.840$, $P=0.183$].

Next, participants were invited to take a 90-minute afternoon nap (Fig. 1C; see Table S1 for sleep-stage information). When EEG signals showed clear signs of slow-wave sleep (SWS), we repeatedly played one auditory cue, randomly selected as the counter-gender-bias sound ($n=21$) or the counter-racial-bias sound ($n=19$). Stimulation was discontinued at any sign of arousal from sleep. The number of presentations averaged 258 ± 24 (SEM).

Implicit biases were measured again after waking. Bias change from prenap to postnap varied with cueing condition as predicted [substantiated by a two-way interaction (cued/uncued by prenap/postnap), $F(1,39)=14.612$, $P<0.001$, $\eta_p^2=0.273$]. As shown in Figure 1D, implicit bias was significantly reduced from prenap to postnap when cued [$t(39)=2.698$, $P=0.010$] and unchanged when not cued [$t(39)=-1.478$, $P=0.176$; Fig. S1-2]. This differential bias reduction was not moderated by bias type (SM; Fig. S3). Thus, reactivating counter-bias learning during sleep can selectively reduce implicit racial or gender bias, depending on which form of counter-bias training was cued.

Implicit biases were measured again after 1 week, revealing that the differential bias reduction endured [Fig. 1E; $n=38$; $F(1,37)=4.672$, $P=0.037$, $\eta p^2=0.112$]. Cueing during sleep resulted in sustained counter-bias reduction such that the cued bias did not differ between prenap and delayed testing [$t(37)=-0.774$, $P=0.444$], whereas the uncued bias increased during the delay [$t(37)=-3.078$, $P=0.004$]. When compared with baseline (Fig. 1F), cued biases were weaker after 1 week [$t(37)=2.203$, $P=0.034$] whereas uncued biases were not [$t(37)=0.524$, $P=0.603$], though the interaction was not significant [$F(1,37)=0.471$, $P=0.497$].

Neurophysiological activity during sleep, such as sleep spindles, slow waves, and rapid-eye-movement (REM) duration, can predict later memory performance (16). Accordingly, we explored possible relationships between cueing-specific bias reduction and measures of sleep physiology. We found that only SWS × REM sleep duration consistently predicted cueing-specific bias reduction at 1 week relative to baseline [Fig. 2; $r(38)=0.450$, $P=0.005$; SM].

Past research indicates that by pairing learning episodes with auditory or olfactory stimuli and then presenting these stimuli again during post-learning SWS, learned information can be specifically reactivated and strengthened (19). Benefits of this Targeted Memory Reactivation (TMR) have been documented for declarative, procedural, and emotional learning (19). Such learning typically does not challenge pre-existing knowledge nor compete with daily experiences outside the laboratory. In contrast, we examined learning-induced changes in long-standing social biases. We showed that selectively reactivating counter-bias learning during sleep weakened pre-existing implicit social biases immediately after the nap, and facilitated the retention of this learning going forward. Without TMR during sleep, training effects tended to dissipate and the bias returned to baseline levels. These results thus enlarge our conception of sleep’s role in socially-relevant learning.

Observed relationships between sleep neurophysiology and behavior further reinforced the conclusion that bias reduction is sleep-dependent. Current thinking about consolidation emphasizes sets of cortical networks that can become integrated through interactions with hippocampal networks, possibly via cyclic SWS-REM periods (15, 19, 27). The correlation with SWS × REM duration implicates a benefit from REM-based processing subsequent to SWS-based reactivation, perhaps to integrate learning within associative knowledge networks. These findings support the notion that SWS and REM are both operative in sleep-dependent memory consolidation (15, 16, 27, 28).

Future research is needed to address many outstanding questions in relation to our findings. For example, how much training is needed to make implicit benefits persist for long periods of time and transfer to explicit benefits in interpersonal interactions? To what extent do persistent benefits depend on repeated training, the nature of other waking activities after training, and repeated memory reactivation during sleep? Although IAT measures are imperfect and may sometimes reflect knowledge of cultural stereotypes rather than implicit bias per se (29), prior research has demonstrated consequences for social behavior such that low implicit bias as measured with the IAT may indeed be linked with egalitarianism (6, 7). Given that training to reduce implicit bias can be conceptualized as a type of habit learning (30), perhaps novel sleep manipulations could be adapted to aid people in changing various unwanted or maladaptive habits, such as smoking, unhealthy eating, catastrophizing, or selfishness (31).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support was provided by Northwestern University, by NSF grants BCS-1025697, BCS-1461088, and DGE-0824162, and by NIH grants T32-AG020418 and F31-MH100958. Data are archived and available on the Open Science Framework (https://osf.io/b3k9a/). K.A.P. served as an advisor to Sheepdog Sciences (a company developing educational technologies related to sleep) in exchange for equity compensation.

References and Notes

1. Glaser, J. Suspect Race: Causes and Consequences of Racial Profiling. Oxford Univ. Press; New York: 2014.
2. Correll J, et al. J Pers Soc Psychol. 2007; 92:1006. [PubMed: 17547485]
3. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Proc Natl Acad Sci USA. 2012; 109:16474. [PubMed: 22988126]
4. Bobo L, Zubrinsky CL. Soc Forces. 1996; 74:883.
5. Devine PG. J Pers Soc Psychol. 1989; 56:5.
6. Stanley DA, Sokol-Hessner P, Banaji MR, Phelps EA. Proc Natl Acad Sci USA. 2011; 108:7710. [PubMed: 21518877]
7. Greenwald AG, Poehlman TA, Uhlmann EL, Banaji MR. J Pers Soc Psychol. 2009; 97:17. [PubMed: 19586237]
8. Nosek BA, et al. Proc Natl Acad Sci USA. 2009; 106:10593. [PubMed: 19549876]
9. Allport, GW. The Nature of Prejudice. Addison-Wesley; Reading, MA: 1979.
10. Fiske, S. The Handbook of Social Psychology. 4. Gilbert, DT.; Fiske, ST.; Lindzey, G., editors. Vol. 2. Oxford Univ. Press; New York: 1998. p. 357-411.
11. Olsson A, Ebert JP, Banaji MR, Phelps EA. Science. 2005; 309:785. [PubMed: 16051800]
12. Crandall CS, Eshleman A, O'Brien L. J Pers Soc Psychol. 2002; 82:359. [PubMed: 11902622]
13. Gawronski B, Deutsch R, Mbirkou S, Seibt B, Strack F. J Exp Soc Psychol. 2008; 44:370.
14. Weisbuch M, Pauker K, Ambady N. Science. 2009; 326:1711. [PubMed: 20019288]
15. Rasch B, Born J. Physiol Rev. 2013; 93:681. [PubMed: 23589831]
16. Diekelmann S, Born J. Nat Rev Neurosci. 2010; 11:114. [PubMed: 20046194]
17. Stickgold R, Walker MP. Nat Neurosci. 2013; 16:139. [PubMed: 23354387]
18. Oudiette D, Paller KA. Trends Cogn Sci. 2013; 17:142. [PubMed: 23433937]
19. Paller, KA. Encyclopedia of Neuroscience. Squire, LR., editor. Academic Press; Oxford: 2009. p. 741-749.
20. Rasch B, Buechel C, Gais S, Born J. Science. 2007; 315:1426. [PubMed: 17347444]
21. Antony JW, Gobel EW, O'Hare JK, Reber PJ, Paller KA. Nat Neurosci. 2012; 15:1114. [PubMed: 22751035]
22. Wilson MA, McNaughton BL. Science. 1994; 265:676. [PubMed: 8036517]
23. Peigneux P, et al. Neuron. 2004; 44:535. [PubMed: 15504332]
24. Rudoy JD, Voss JL, Westerberg CE, Paller KA. Science. 2009; 326:1079. [PubMed: 19965421]
25. Greenwald AG, McGhee DE, Schwartz JLK. J Pers Soc Psychol. 1998; 74:1464. [PubMed: 9654756]
26. Greenwald AG, Nosek BA, Banaji MR. J Pers Soc Psychol. 2003; 85:197. [PubMed: 12916565]
27. Walker MP, Stickgold R. Nat Rev Neurosci. 2010; 11
28. Ambrosini MV, Giuditta A. Sleep Med Rev. 2001; 5:477. [PubMed: 12531155]
29. Arkes HR, Tetlock PE. Psychol Inq. 2004; 15:257.

30. Devine PG, Forscher PS, Austin AJ, Cox WTL. J Exp Soc Psychol. 2012; 48:1267. [PubMed: 23524616]

31. Arzi A, et al. J Neurosci. 2014; 34:15382. [PubMed: 25392505]

32. Lundqvist, D.; Flykt, A.; Öhman, A. The Karolinska Directed Emotional Faces— KDEF [CD-ROM]. Department of Clinical Neuroscience, Psychology section; Karolinska Institutet: 1998.

33. Tottenham N, et al. Psychiat Res. 2009; 168:242.

34. Nosek BA, Banaji MR, Greenwald AG. J Pers Soc Psychol. 2002; 83:44. [PubMed: 12088131]

35. Iber, C.; Ancoli-Israel, S.; Chesson, A.; Quan, SF. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. 1. American Academy of Sleep Medicine; Westchester, IL: 2007.

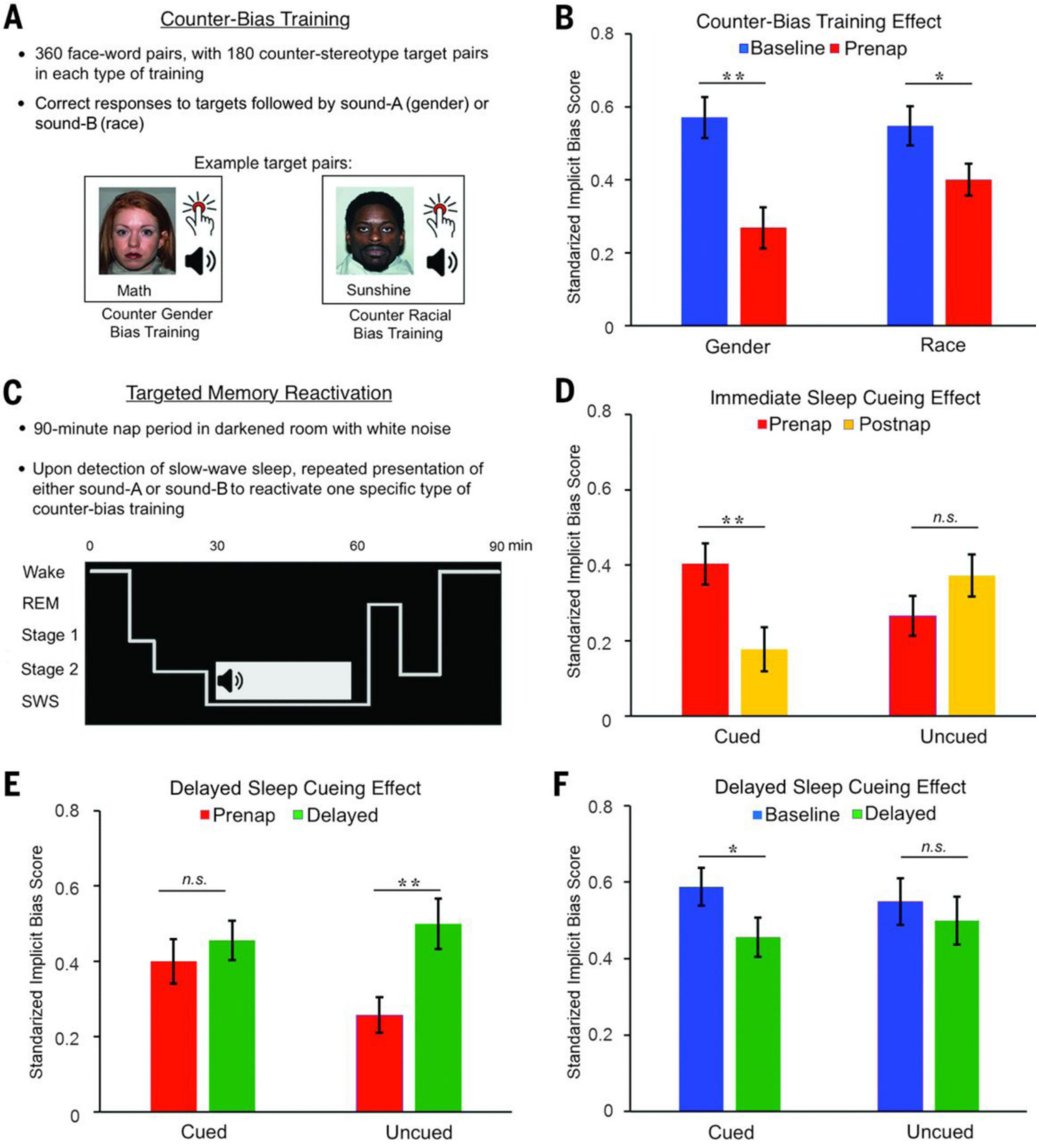

Fig. 1.

Experimental procedures and results. (A) Procedures for counter-bias training with sound cues. (B) Implicit bias reduction was found for both counter-racial-bias training and counter-gender-stereotype training. 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 to participants during SWS, using a low intensity to avoid arousal from sleep. (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. (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. (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. Significant pairwise differences are indicated by * ($P < 0.05$) or ** ($P < 0.01$).

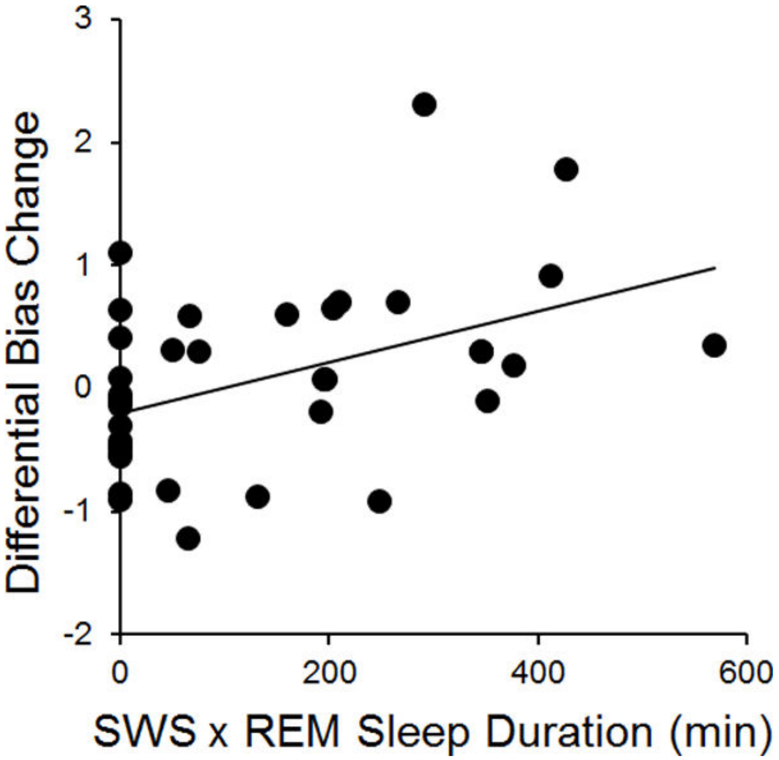

Fig. 2.

The quality of sleep after training, as indexed by the product of minutes in SWS × minutes in REM sleep, predicted differential bias change, quantified as follows. Given that standardized implicit bias scores were preferentially reduced at 1 week for the cued relative to uncued condition overall (Fig. 1F), we computed baseline-minus-delayed difference scores for the two conditions. Differential bias change was taken as the cued difference minus the cued difference, such that higher values indicated larger bias reduction over this interval for the cued compared to the uncued bias.