COGNITIVE NEUROSCIENCE

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 Ken A. Paller¹*

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 counterstereotype 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 with the social bias not externally reactivated during sleep. This advantage remained 1 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 counterstereotype training and that maintaining a bias reduction is sleep-dependent.

ocial 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 that used a first-person-shooter videogame, both White and Black participants were more likely to shoot Black

¹Department of Psychology, Northwestern University, Evanston, IL 60208, USA. ²Department of Psychology, University of Texas at Austin, Austin, TX 78712, USA. ³Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.

*Corresponding author. E-mail: kap@northwestern.edu

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 nevertheless influence people's behavior in an implicit or unconscious manner, regardless of their intentions or efforts to avoid bias (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 because of 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 (II). 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). Last, perceived social norms can prescribe people's expression of stereotyping and prejudice (12). Despite such challenges, implicit biases can be reduced by learning about counterstereotype cases (13). However, benefits of this counterbias 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 counterbias 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 (*15–19*). 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 postsleep memory performance (*19–24*).

Taking into consideration the role of sleep in memory consolidation, we adapted procedures for (i) reducing implicit social biases and (ii) reactivating this training during sleep. We were particularly interested in factors that can influence whether such training procedures produce transient or 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 counterbias information during sleep using subtle auditory cues that had been associated with counterbias 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 (25). First, biases were quantified using two versions of the implicit association test (IAT) (26). The IAT allows for an assessment of the strength of implicit associations between social groups and attributes (26). One test examined the degree to which female faces were preferentially associated with art versus science 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 by using a conventional scoring procedure (27), in which zero indicates no bias and larger scores indicate greater bias. Consistent with previous research (7), IAT scores showed that participants held implicit so-cial biases for both gender and race, with both scores significantly greater than zero [mean ± SEM, 0.559 ± 0.044; gender t(39) = 9.076, P < 0.001; race t(39) = 8.388, P < 0.001].

After 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 counterstereotype 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) (25). Two unusual frequency-modulated sounds were presented during training, one after correct counter-gender bias responses and the other after 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 (25). Training thus established a strong association between each sound and one type of counterbias training.





Biases were reduced compared with baseline levels (Fig. 1B) [within-subject analysis of variance (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-min afternoon nap (Fig. 1C; see table S1 for sleep-stage information). When electroencephalographic 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 or uncued by prenap or postnap), $F_{1,29} = 14.612$, P < 0.001, $\eta_p^2 = 0.273$]. As shown in Fig. 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) = -1.378, P = 0.176] (fig. S1 and S2). This differential bias reduction was not moderated by bias type (fig. S3). Thus, reactivating counterbias learning during sleep can selectively reduce implicit racial or gender bias, depending on which form of counterbias 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 counterbias reduction, such that the cued bias did not differ between prenap and delayed testing



Fig. 2. Relation between long-term bias reduction and sleep physiology. 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 for the cued relative to uncued condition overall (Fig. 1F), we computed the reduction separately for cued and uncued conditions [(baseline score) – (delayed score) in both cases]. Differential bias change was taken as the cued reduction minus the uncued reduction, such that higher values indicated larger bias reduction over this interval for the cued compared to the uncued bias.

[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], although 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-eyemovement (REM) duration—can predict later memory performance (17). Accordingly, we explored possible relations 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] (25).

Past research indicates that by pairing learning episodes with auditory or olfactory stimuli and then presenting these stimuli again during postlearning 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 memories (19). Such learning typically does not challenge preexisting 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 counterbias learning during sleep weakened preexisting 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 relations 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 by means of cyclic SWS-REM periods (15–17, 28). 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 both SWS and REM are operative in sleep-dependent memory consolidation (16, 17, 28, 29).

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 (*30*), 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 (31), 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 (32).

REFERENCES AND NOTES

- . J. Glaser, Suspect Race: Causes and Consequences of Racial Profiling (Oxford Univ. Press, New York, 2014).
- 2. J. Correll et al., J. Pers. Soc. Psychol. 92, 1006–1023 (2007).
- C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, J. Handelsman, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16474–16479 (2012).
- 4. L. Bobo, C. L. Zubrinsky, Soc. Forces 74, 883-909 (1996).
- 5. P. G. Devine, J. Pers. Soc. Psychol. 56, 5–18 (1989).
- D. A. Stanley, P. Sokol-Hessner, M. R. Banaji, E. A. Phelps, Proc. Natl. Acad. Sci. U.S.A. 108, 7710–7715 (2011).
- A. G. Greenwald, T. A. Poehlman, E. L. Uhlmann, M. R. Banaji, J. Pers. Soc. Psychol. 97, 17–41 (2009).
- B. A. Nosek et al., Proc. Natl. Acad. Sci. U.S.A. 106, 10593–10597 (2009).
- 9. G. W. Allport, *The Nature of Prejudice* (Addison-Wesley, Reading, MA, 1979).
- S. Fiske, in *The Handbook of Social Psychology*, D.T. Gilbert, S.T. Fiske, G. Lindzey, Eds. (Oxford Univ. Press, New York, ed. 4, 1998), vol. 2, pp. 357–411.
- A. Olsson, J. P. Ebert, M. R. Banaji, E. A. Phelps, *Science* **309**, 785–787 (2005).
- C. S. Crandall, A. Eshleman, L. O'Brien, J. Pers. Soc. Psychol. 82, 359–378 (2002).
- B. Gawronski, R. Deutsch, S. Mbirkou, B. Seibt, F. Strack, J. Exp. Soc. Psychol. 44, 370–377 (2008).
- M. Weisbuch, K. Pauker, N. Ambady, Science 326, 1711–1714 (2009).
 K. A. Paller, in *Encyclopedia of Neuroscience*, L. R. Squire, Ed.
- (Academic Press, Oxford, 2009), pp. 741-749.
- 16. B. Rasch, J. Born, Physiol. Rev. 93, 681-766 (2013)
- S. Diekelmann, J. Born, *Nat. Rev. Neurosci.* 11, 114–126 (2010).
 R. Stickgold, M. P. Walker, *Nat. Neurosci.* 16, 139–145 (2013).
- 18. R. Stickgold, M. P. Walker, Nat. Neurosci. 16, 139–145 (2013)
- 19. D. Oudiette, K. A. Paller, Trends Cogn. Sci. 17, 142–149 (2013)
- B. Rasch, C. Büchel, S. Gais, J. Born, *Science* **315**, 1426–1429 (2007).
 J. W. Antony, E. W. Gobel, J. K. O'Hare, P. J. Reber, K. A. Paller,
- Nat. Neurosci. 15, 1114–1116 (2012). 22. M. A. Wilson, B. L. McNaughton, Science 265, 676–679 (1994).
- 23. P. Peigneux *et al.*, *Neuron* **44**, 535–545 (2004).
- 24. J. D. Rudoy, J. L. Voss, C. E. Westerberg, K. A. Paller, Science
- 326, 1079 (2009).
 25. Materials and methods are available as supplementary
- 25. Materials and methods are available as supplementary materials on *Science* online.
- A. G. Greenwald, D. E. McGhee, J. L. K. Schwartz, J. Pers. Soc Psychol. 74, 1464–1480 (1998).
- A. G. Greenwald, B. A. Nosek, M. R. Banaji, J. Pers. Soc. Psychol. 85, 197–216 (2003).
- R. Stickgold, D. Whidbee, B. Schirmer, V. Patel, J. A. Hobson, J. Cogn. Neurosci. 12, 246–254 (2000).
- 29. M. V. Ambrosini, A. Giuditta, Sleep Med. Rev. 5, 477-490 (2001).
- 30. H. R. Arkes, P. E. Tetlock, Psychol. Ing. 15, 257-278 (2004).
- 31. P. G. Devine, P. S. Forscher, A. J. Austin, W. T. L. Cox, J. Exp.
- Soc. Psychol. **48**, 1267–1278 (2012). 32. A. Arzi et al., J. Neurosci. **34**, 15382–15393 (2014).

ACKNOWLEDGMENTS

Support was provided by Northwestern University and 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://osfio/b3k9a/). KA.P. served as an advisor to Sheepdog Sciences (a company developing educational technologies related to sleep) in exchange for equity compensation. The authors declare no conflict of interests.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6238/1013/suppl/DC1 Materials and Methods Supplementary Text Figs. S1 to S3 Table S1 Sound Files References (*33–36*) 2 December 2014; accepted 20 April 2015

10.1126/science.aaa3841



Supplementary Materials for

Unlearning implicit social biases during sleep

Xiaoqing Hu, James W. Antony, Jessica D. Creery, Iliana M. Vargas, Galen V. Bodenhausen, Ken A. Paller*

*Corresponding author. E-mail: kap@northwestern.edu

Published 29 May 2015, *Science* **348**, 1013 (2015) DOI: 10.1126/science.aaa3841

This PDF file includes

Materials and Methods Supplementary Text Figs. S1 to S3 Table S1 References

Other Supplementary Material for this manuscript includes the following: (available at www.sciencemag.org/content/348/6238/1013/suppl/DC1)

Sound files S1 and S2

Materials and Methods

Participants: Forty White participants (20 females, age: $M \pm SD$, 21.75 \pm 3.60 years old, range: 19–32 years old) were recruited from the local university community. The study was approved by the Northwestern University Institutional Review Board and all participants gave informed consent. Data from 17 additional participants were excluded from further analyses because they did not enter slow-wave sleep (n = 12), reported hearing sound cues during sleep (n = 4), or an EEG data file was corrupted (n = 1). Data were collected from two independent samples, allowing for a direct replication. Data from sample-1 (n = 21) were collected during June–July 2014; data from sample-2 (n = 19) were collected during February–March 2015. On the day before the experiment, all participants were reminded via email that they should wake up at least 2 hours earlier than usual and refrain from caffeine the next day.

Stimuli: Neutral face stimuli (Black and White, both genders) were chosen from the Eberhardt Lab Face Database, the NimStim Face Stimulus Set, and the Karolinska Directed Emotional Faces (*32, 33*). Word stimuli (categorized as good, bad, science, or art) were chosen from the literature on implicit racial bias and gender stereotypes (*8, 25, 34*). Different stimulus sets were used in the baseline administration of Implicit Attitude Tests (IATs), the counterbias training, the sound cue–retrieval task, and the subsequent IATs. Words used in both the test and the training sessions are provided below:

Good/bad words used in the racial-bias IATs: Paradise, Freedom, Harmony, Smile, Cheer, Sunrise, Disaster, Vomit, Bomb, Cancer, Rotten, Coma

Good/bad words used in the counter-racial bias training: *Heaven, Pleasure, Sunshine, Miracle, Honor, Evil, Murder, Grief, Poison, Virus,*

Science/art words used in the gender-bias IATs: Geology, Physics, Einstein, Experiment, NASA, Biology, Poetry, Literature, Shakespeare, Symphony, English, Humanities

Science/art words used in the counter–gender bias training: Science, Chemistry, Math, Geometry, Engineering, Dance, Theater, Drama, History, Music

Procedure: All participants completed the following procedures: (i) baseline IATs, ~10 min; (ii) counterbias training sessions with sound cues, ~30 min; (iii) EEG set-up, ~20 min; (iv) prenap IATs, ~10 min; (v) sound cue–retrieval task, ~5 min; (vi) 90-min nap plus a 10-min break; (vii) postnap IATs, ~10 min; and (viii) 1-week delayed IATs, ~10 min. For the delayed tests, 38 out of 40 participants returned. Among these, 31 participants returned after 7 days, and the remaining returned either 6 (n = 3), 8 (n = 2), 9 (n = 1), or 10 days later (n = 1). All tests were conducted between 10 a.m. and 6 p.m., with the nap starting between 12:30 p.m. and 4:30 p.m. Experimenters were blind to the cueing conditions during IAT administrations. Speed and accuracy were emphasized equally during all tasks.

Baseline Implicit Bias Assessments. Participants completed two IATs, a racial-bias IAT and a gender-stereotype IAT. Both IATs were structured in the seven-block version [see reference (26) for task structure details]. Twenty-one participants completed the racial IAT first and 19 completed the gender IAT first. In the racial IAT, participants performed a categorization task that included good and bad words randomly mixed with White and Black male faces. In the

gender-stereotype IAT, participants performed a similar task but with science and art words randomly mixed with White female and male faces. On each trial, a single stimulus was presented centrally until a correct response was registered (onset 150 ms after prior response). If an incorrect response was registered, an error feedback "X" was presented on the screen until participants gave the correct response. Participants were instructed to make a categorization response as quickly and accurately as possible. There were always two response buttons (buttons "E" and "I" on a keyboard), but button assignments alternated across assessment blocks (each with 60 trials). Assignments were either consistent or inconsistent with the typical bias. In the former case, one button was used for White+good (or male+science) and the other for Black+bad (or female+art). In the latter case, one button was used for Black+good (or female+science) and the other for White+bad (or male+art). Implicit racial or gender bias was inferred when categorization responses were quicker and more accurate in blocks with button assignments that were consistent with the typical bias compared with blocks with button assignments that were inconsistent with the typical bias.

To quantify IAT results, a D_{600} score was calculated as the dependent variable following a conventional algorithm (26). First, response times (RTs) shorter than 300 ms or longer than 3 s were deleted (<1%). Second, RTs for correct responses were averaged separately for blocks with button assignments consistent with the typical bias and for blocks with button assignments inconsistent with the typical bias. Third, we calculated the standard deviation of the RT distributions from correct trials of both consistent and inconsistent blocks combined. Fourth, any incorrect responses were replaced with the mean RT associated with that particular block plus a 600-ms penalty (26). Fifth, the means of consistent and inconsistent blocks were calculated separately including RTs of incorrect responses with the error penalties. Sixth, the RT differences between the consistent and the inconsistent blocks (RT_{inconsistent}) from step five were divided by the inclusive standard deviation obtained from step three. The result of step six was the D_{600} score (26). A larger D_{600} score indicates a stronger implicit social bias. The measure does not indicate whether there is also an explicit social bias, or whether there are interactions between processing related to implicit and explicit measures of social bias.

Counterbias Training with Sound Cues: Participants completed a counter-racial bias training session and a counter-gender bias training session, and the order of the two sessions was counterbalanced (counter-racial bias training first for 20 participants and counter-gender bias training first for 20 participants). During each training session, participants were presented with 360 trials (divided into 3 blocks each followed by a short break, and with trials separated by an interstimulus interval of 1 s). Half of the trials were counterbias face+word pairs: Black+good or female+science. On each of these trials, the participant was required to press the spacebar. Critically, participants were informed that if they made a quick and accurate response to a counterbias pair, correct feedback would be given in the form of a 1-s sound from a speaker (for this purpose, quick responses were those made within 800 ms, though participants were not directly informed of this specific response deadline). One of the two training sounds, sound-a or sound-b, (Sounds S1 and S2, respectively) was used for this feedback in each counterbias training session (counterbalanced across subjects; sound-a was used in counter-racial bias training for half of the participants, and in counter-gender bias training for the remaining participants; sounds and other stimuli are available from the first author upon request). The remaining trials were fillers that required no response. Filler trials included 60 trials in each of three categories: Black+bad; White+good; and White+bad (counter-racial bias training) or

female+art; male+science; and male+art (counter-gender bias training). Filler trials were included so that participants actively discriminated between counterbias and filler trials, with an emphasis on selecting counterbias trials.

EEG Setup. After counterbias training, an EEG cap with 21 electrodes was applied. EOG and EMG electrodes were also attached to monitor eye movement and muscle activity. Continuous EEG was recorded from International 10–20 locations Fpz, Fz, Cz, Pz, Oz, Fp1/2, F3/4, F7/8, C3/4, P3/4, T3/4, T5/6, and O1/2. Two electrodes were placed to monitor eye movements, one below the left eye and the other next to the right eye for recording the vertical and horizontal EOG; one electrode was placed on the chin to record EMG. Online EEG recordings were referenced to the average of left and right mastoid, filtered at 0.1–200 Hz with a 60 Hz notch filter, and sampled at 500 Hz.

Prenap Implicit Bias Assessments: Participants were tested with the two IATs again to measure training effects. The order of the two tests was the same as in the baseline measurements, but the word and face stimuli used in each IAT were unique.

Sound-Cue-Retrieval Task: This task included 120 trials that were divided into six blocks, with trials separated by an interstimulus interval of 1 s. Each trial began with presentation of one of the two sounds that were used in the counterbias training tasks. Concurrently with the onset of the sound, either a White female face or a Black male face was presented on the left side of the monitor, while a science word and a good word were presented on the upper-right or lower-right side of the monitor (location of words randomized across trials). Participants were required to use a mouse to drag the face to its corresponding word to create a counterbias pair in accordance with the sound that was presented. They were told that if they heard the sound from countergender bias training, they were to link the female with the science word, and if they heard the sound from counter-racial bias training, they were to link the Black face with the good word. This task thus provided further training in associating each sound with its corresponding counterbias associations, as well as active training in counterbias face-word associations. It thus ensured that the sound-to-bias-type associations were well learned. It is also notable that the sounds were presented as a signal to retrieve counterbias information in this task, whereas during counterbias training the sounds were presented only after correct responses. We cannot determine which aspects of these procedures were most relevant, but the combination was apparently effective in strongly linking each sound to a specific type of counterbias training such that presenting sound cues during sleep could reactivate the corresponding memory.

Nap. Participants were prepared for a 90-min nap session in the same room, with a futon pillow, blanket, and sheets supplied. Participants were told that white noise would be played over a speaker to block out any environmental noise. When the experimenter observed that the participant had entered slow-wave sleep on the basis of standard EEG criteria, a sound cue was repeatedly presented; 19 participants received the sound associated with counter–racial bias training and 21 received the sound associated with counter–gender-stereotype training. Each sound cue lasted for 1 s, with an interstimulus interval of 4 s. The intensity of the cue was similar to the background white noise [38–40 db SPL]. Sound cues were halted whenever arousal was evidenced in physiological recordings. This sleep phase ended after 90 min unless the participant was still in slow-wave sleep (n = 3), in which case additional time was allowed for the slow-wave sleep stage to end naturally.

Sleep Physiological Analysis. Experimenters who were blind to participants' results conducted all sleep physiological analyses. Continuous EEG traces were down-sampled to 125 Hz and filtered at 0.5–60 Hz using a two-way least squares finite impulse response ("eeg_filt" function) offline. Sleep stages were formally identified offline using standard sleep-scoring methods (*35*).

Postnap Implicit Bias Assessments. This session began 10 min after the participant woke up from the nap. Participants finished two IATs, the order of which was opposite from baseline measurements. After the postnap bias assessments, participants were asked in an open manner whether they had heard any sound during sleep other than the white noise. Participants who reported that they heard a sound associated with counterbias training were excluded (n = 4).

Delayed Implicit Bias Assessments. Participants returned to the lab for the two IATs after 1 week. Implicit biases were assessed using the same methods as used in the postnap implicit bias assessment.

Supplementary Text

Physiology-behavior correlations

We calculated cueing advantage effects in several ways. For cued and uncued biases separately, we focused on three comparisons: (i) prenap bias minus postnap bias to reflect the immediate bias change, (ii) prenap bias minus delayed bias to reflect the persistent bias change and (iii) baseline bias minus delayed bias to reflect the persistent bias change. We then subtracted the uncued bias change from the cued bias change for each of the three comparisons, which yielded a measure for Differential Bias Change. A positive value indicates that cued biases were reduced more than uncued biases (i.e., sleep-dependent cueing advantage). Results showed that only the baseline-versus-delayed Differential Bias Change values were positively correlated with the product of SWS and REM sleep durations [r(38) = 0.450, 95% confidence interval (CI): 0.212, 0.658; Fig. 2]. This correlation was highly consistent across sample-1 and sample-2 [r(19) = 0.443, 95% CI: -0.001, 0.775, and r(19) = 0.568, 95% CI: 0.220, 0.792, respectively]. The variability in scores across participants can also be viewed in a scatterplot of IAT results from each test separately (Fig. S1).

Subsamples

In both sample-1 and sample-2, we found an impact of memory reactivation during sleep on bias reduction. The primary comparison was for prenap versus postnap IAT scores as a function of sleep cueing condition (cued/uncued), as shown in Fig. S2. Indeed, cueing one type of counterbias training during sleep revealed a selective reduction in implicit bias from prenap to postnap, both in sample-1 [F(1,20) = 7.027, P = 0.015, $\eta_p^2 = 0.260$] and in sample-2 [F(1,18) = 7.382, P = 0.014, $\eta_p^2 = 0.291$].

Although prenap biases were not closely matched for cued and uncued conditions in sample-2, procedures for these two conditions were identical at this point. The failure to have a close match can thus be attributed to the randomization process for sample-2. However, we can be confident that the central sleep cueing interaction (Fig. S2; cued/uncued by prenap/postnap) is not dependent on prenap differences, because the same interaction was found in sample-1 when cued and uncued conditions were closely matched (Fig. S2A). Moreover, prenap differences in sample-2 were driven by lower uncued gender bias scores, as shown in Fig. S3.

Effects of bias type

To examine whether bias type (gender vs. race) modulated immediate or delayed sleep cueing effects, we conducted two mixed ANOVAs on IAT scores (Fig. S3). Both ANOVAs included bias type (gender/race) as a between-subject variable, with cueing (cued/uncued) and time (prenap/postnap in the first ANOVA, prenap/delayed in the second ANOVA) as within-subject variables. Results showed that the three-way interactions were not significant, suggesting that bias type did not modulate either the prenap/postnap sleep cueing effect [F(1,38) = 0.212, P = 0.648, $\eta_p^2 = 0.006$], or the prenap/delayed sleep cueing effect [F(1,36) = 0.006, P = 0.939, $\eta_p^2 = 0.001$].



Fig. S1. Individual participant data for each sub-sample for baseline, prenap, postnap and 1-week delayed IAT assessments.



Fig. S2. Replication of the prenap versus postnap sleep cueing effect in sample-1 and sample-2. Cue presentations during sleep led to further bias reduction after a nap in both samples. In contrast, no reduction was observed for uncued biases. Error bars indicate ± 1 SEM adjusted for within-subject comparisons.



Fig. S3. Behavioral data analyzed separated by type of bias. When counter–gender bias training was cued, racial bias was uncued; when counter–racial bias training was cued, gender bias was uncued. (A) Similar patterns of prenap versus postnap sleep cueing effects were found in both

cases. (B) Similar patterns of prenap versus delayed sleep cueing effects were found in both cases. Error bars indicate ± 1 SEM adjusted for within-subject comparisons.

Table S1. Sleep stage results.

Sleep Stage	Time (Mean min \pm SEM)
Wake	21.38 ± 3.11
Stage 1	7.33 ± 0.92
Stage 2	26.03 ± 1.67
Slow-Wave Sleep	27.16 ± 2.40
REM	5.35 ± 0.92

REFERENCES AND NOTE

- 1. J. Glaser, Suspect Race: Causes and Consequences of Racial Profiling (Oxford Univ. Press, New York, 2014).
- J. Correll, B. Park, C. M. Judd, B. Wittenbrink, M. S. Sadler, T. Keesee, Across the thin blue line: Police officers and racial bias in the decision to shoot. *J. Pers. Soc. Psychol.* 92, 1006–1023 (2007). <u>Medline doi:10.1037/0022-3514.92.6.1006</u>
- C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, J. Handelsman, Science faculty's subtle gender biases favor male students. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16474–16479 (2012). <u>Medline</u>
- L. Bobo, C. L. Zubrinsky, Attitudes on residential integration: Perceived status differences, mere in-group preference, or racial prejudice? *Soc. Forces* 74, 883–909 (1996). doi:10.1093/sf/74.3.883
- 5. P. G. Devine, Stereotypes and prejudice: Their automatic and controlled components. J. Pers. Soc. Psychol. 56, 5–18 (1989). doi:10.1037/0022-3514.56.1.5
- D. A. Stanley, P. Sokol-Hessner, M. R. Banaji, E. A. Phelps, Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7710–7715 (2011). <u>Medline doi:10.1073/pnas.1014345108</u>
- A. G. Greenwald, T. A. Poehlman, E. L. Uhlmann, M. R. Banaji, Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *J. Pers. Soc. Psychol.* 97, 17–41 (2009). Medline doi:10.1037/a0015575
- 8. B. A. Nosek, F. L. Smyth, N. Sriram, N. M. Lindner, T. Devos, A. Ayala, Y. Bar-Anan, R. Bergh, H. Cai, K. Gonsalkorale, S. Kesebir, N. Maliszewski, F. Neto, E. Olli, J. Park, K. Schnabel, K. Shiomura, B. T. Tulbure, R. W. Wiers, M. Somogyi, N. Akrami, B. Ekehammar, M. Vianello, M. R. Banaji, A. G. Greenwald, National differences in gender-science stereotypes predict national sex differences in science and math

achievement. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10593–10597 (2009). <u>Medline</u> doi:10.1073/pnas.0809921106

- 9. G. W. Allport, The Nature of Prejudice (Addison-Wesley, Reading, MA, 1979).
- 10. S. Fiske, in *The Handbook of Social Psychology*, D.T. Gilbert, S.T. Fiske, G. Lindzey, Eds. (Oxford Univ. Press, New York, ed. 4, 1998), vol. 2, pp. 357–411.
- 11. A. Olsson, J. P. Ebert, M. R. Banaji, E. A. Phelps, The role of social groups in the persistence of learned fear. *Science* **309**, 785–787 (2005). <u>Medline doi:10.1126/science.1113551</u>
- C. S. Crandall, A. Eshleman, L. O'Brien, Social norms and the expression and suppression of prejudice: The struggle for internalization. *J. Pers. Soc. Psychol.* 82, 359–378 (2002). <u>Medline doi:10.1037/0022-3514.82.3.359</u>
- 13. B. Gawronski, R. Deutsch, S. Mbirkou, B. Seibt, F. Strack, When "Just Say No" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *J. Exp. Soc. Psychol.* **44**, 370–377 (2008). doi:10.1016/j.jesp.2006.12.004
- M. Weisbuch, K. Pauker, N. Ambady, The subtle transmission of race bias via televised nonverbal behavior. *Science* 326, 1711–1714 (2009). <u>Medline</u> <u>doi:10.1126/science.1178358</u>
- 15. K. A. Paller, in *Encyclopedia of Neuroscience*, L. R. Squire, Ed. (Academic Press, Oxford, 2009), pp. 741–749.
- 16. B. Rasch, J. Born, About sleep's role in memory. *Physiol. Rev.* **93**, 681–766 (2013). <u>Medline</u> doi:10.1152/physrev.00032.2012
- S. Diekelmann, J. Born, The memory function of sleep. *Nat. Rev. Neurosci.* 11, 114–126 (2010). <u>Medline</u>
- R. Stickgold, M. P. Walker, Sleep-dependent memory triage: Evolving generalization through selective processing. *Nat. Neurosci.* 16, 139–145 (2013). <u>Medline</u> <u>doi:10.1038/nn.3303</u>
- 19. D. Oudiette, K. A. Paller, Upgrading the sleeping brain with targeted memory reactivation. *Trends Cogn. Sci.* **17**, 142–149 (2013). <u>Medline doi:10.1016/j.tics.2013.01.006</u>
- 20. B. Rasch, C. Büchel, S. Gais, J. Born, Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* **315**, 1426–1429 (2007). <u>Medline</u> <u>doi:10.1126/science.1138581</u>
- 21. J. W. Antony, E. W. Gobel, J. K. O'Hare, P. J. Reber, K. A. Paller, Cued memory reactivation during sleep influences skill learning. *Nat. Neurosci.* 15, 1114–1116 (2012). <u>Medline doi:10.1038/nn.3152</u>
- M. A. Wilson, B. L. McNaughton, Reactivation of hippocampal ensemble memories during sleep. *Science* 265, 676–679 (1994). <u>Medline doi:10.1126/science.8036517</u>
- 23. P. Peigneux, S. Laureys, S. Fuchs, F. Collette, F. Perrin, J. Reggers, C. Phillips, C. Degueldre, G. Del Fiore, J. Aerts, A. Luxen, P. Maquet, Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* 44, 535–545 (2004). <u>Medline doi:10.1016/j.neuron.2004.10.007</u>

- 24. J. D. Rudoy, J. L. Voss, C. E. Westerberg, K. A. Paller, Strengthening individual memories by reactivating them during sleep. *Science* **326**, 1079 (2009). <u>Medline</u> <u>doi:10.1126/science.1179013</u>
- 25. Materials and methods are available as supplementary materials on *Science* online.
- A. G. Greenwald, D. E. McGhee, J. L. K. Schwartz, Measuring individual differences in implicit cognition: The implicit association test. *J. Pers. Soc. Psychol.* 74, 1464–1480 (1998). <u>Medline doi:10.1037/0022-3514.74.6.1464</u>
- A. G. Greenwald, B. A. Nosek, M. R. Banaji, Understanding and using the implicit association test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* 85, 197–216 (2003). <u>Medline doi:10.1037/0022-3514.85.2.197</u>
- R. Stickgold, D. Whidbee, B. Schirmer, V. Patel, J. A. Hobson, Visual discrimination task improvement: A multi-step process occurring during sleep. *J. Cogn. Neurosci.* 12, 246– 254 (2000). <u>Medline doi:10.1162/089892900562075</u>
- 29. M. V. Ambrosini, A. Giuditta, Learning and sleep: The sequential hypothesis. *Sleep Med. Rev.* **5**, 477–490 (2001). <u>Medline doi:10.1053/smrv.2001.0180</u>
- 30. H. R. Arkes, P. E. Tetlock, Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?". *Psychol. Ing.* 15, 257–278 (2004). doi:10.1207/s15327965pli1504_01
- P. G. Devine, P. S. Forscher, A. J. Austin, W. T. L. Cox, Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *J. Exp. Soc. Psychol.* 48, 1267–1278 (2012). <u>Medline doi:10.1016/j.jesp.2012.06.003</u>
- A. Arzi, Y. Holtzman, P. Samnon, N. Eshel, E. Harel, N. Sobel, Olfactory aversive conditioning during sleep reduces cigarette-smoking behavior. *J. Neurosci.* 34, 15382– 15393 (2014). Medline doi:10.1523/JNEUROSCI.2291-14.2014
- D. Lundqvist, A. Flykt, A. Öhman, *The Karolinska Directed Emotional Faces— KDEF* [CD-ROM] (Psychology Section, Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden, 1998).
- 34. N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, C. Nelson, The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Res.* 168, 242–249 (2009). <u>Medline</u> <u>doi:10.1016/j.psychres.2008.05.006</u>
- 35. B. A. Nosek, M. R. Banaji, A. G. Greenwald, Math = male, me = female, therefore math not = me. *J. Pers. Soc. Psychol.* **83**, 44–59 (2002). <u>Medline doi:10.1037/0022-3514.83.1.44</u>
- 36. C. Iber, S. Ancoli-Israel, A. Chesson, S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications* (American Academy of Sleep Medicine, Westchester, IL, ed. 1, 2007).

NEUROSCIENCE

Exploiting sleep to modify bad attitudes

Targeted memory reactivation of training-induced social counterbias during sleep

By Gordon B. Feld and Jan Born

ince the age of enlightenment in the 18th century, liberty and equality have spread across the Western world, leading to a decline in explicit racism and sexism. Nevertheless, the tendency to hold implicit prejudices of race or gender continues to drive discrimination (1, 2). Indeed, recent news has been filled with reports on the rise of nationalistic groups, excessive police violence against minority group members, persisting unequal pay for women, and sexual harassment all across the developed world. On page 1013 in this issue, Hu et al. (3) show how such unwanted attitudes may be persistently changed by a social counterbias training when the fresh memories of this training are systematically reactivated during sleep after training.

Sleep, and specifically deep or slow-wave sleep [non-rapid eye movement (REM) sleep], benefits memory formation by reactivating neuronal traces that were formed during the preceding period of wakefulness. This reactivation of specific memories leads to their strengthening and transformation (4). Such reactivation can be experimentally induced during slow-wave sleep by presenting cues that were present during the prior period of memory acquisition. Initial studies showed that an odor present during learning of object locations enhances these memories when the participant is reexposed to the odor during slow-wave sleep after learning (5). These findings have been confirmed in numerous studies investigating different memory systems and also when auditory instead of olfactory cues are used (6-8). This basic research has firmly established the possibility of influencing sleep to enhance specific newly learned memories by targeted memory reactivation.

The findings by Hu *et al.* now suggest that this method can also be used to influence implicit attitudes that are known to typically manifest themselves early during childhood and remain very stable into adulthood (9). Before a 90-min nap, participants underwent training aimed at countering typical implicit gender and racial biases by learning to associate genders and

races with opposing attributes; that is, to associate female faces with science-related words and black faces with "good" words (see the figure). Critically, presentation of the to-be-learned counterassociations was combined with a sound, which served as a cue to promote the reactivation of the newly learned associations during a subsequent nap while the participant was deep in slow-wave sleep. Only when this sound was re-presented during slow-wave sleep

"...this method can be used to break long-lived, highly pervasive response habits deeply rooted in memory..."

did the posttraining reduction in implicit social bias survive and was even evident 1 week later. These findings are all the more convincing as the authors conducted the reactivation step during a 90-min daytime nap. During normal sleep at night, the effects are expected to be even stronger, owing to the generally deeper and longer periods of slow-wave sleep and REM sleep. Additionally, the accompanying neuroendocrine milieu makes nocturnal sleep even more efficient for memory reinforcement.

Previous studies have shown that such targeted reactivation of memory during sleep can effectively extinguish unwanted behavior such as experimentally induced fear in humans (10, 11). The present study is the first to demonstrate that this method can be used to break long-lived, highly pervasive response habits deeply rooted in memory and thereby influence behavior at an entirely unconscious level. A hallmark of such automatically implemented habit memories is that their retrieval escapes cognitive control exerted over consciously recalled explicit memory by the prefrontal cortical-hippocampal system. In this regard, Hu *et al.*'s data

University of Tuebingen, Institute of Medical Psychology and Behavioral Neurobiology, Otfried-Müller-Straße 25, 72076 Tuebingen, Germany. E-mail: jan.born@uni-tuebingen.de; gordon.feld@uni-tuebingen.de



Sleeping your way out of a bad attitude. Implicit social bias scores could be improved by applying a counterbias intervention comprising two tasks: counterbias training with sound feedback and retrieval of the sound-counterbias association. This training-induced improvement was then stabilized by re-presenting the sound cue during sleep. (**A**) On the counterbias training, participants were shown separate pictures of men and women of different racial groups together with words from the opposing categories "science/art" and "good/bad". When seeing an "incongruous" pair (e.g., face of a woman and the word "math"), participants had to respond by pressing a button. "Correct" and timely responses received a feedback sound. Congruent trials afforded no response. (**B**) On the sound-counterbias retrieval task, participants were instructed to drag and drop a face (e.g., a female face) onto the incongruous word (e.g., "math") whenever they heard the sound that was associated with this specific counterbias during the preceding counterbias training. (**C**) The sound was then used to cue, and thereby reactivate, the memory of the newly learned counterbias association when the participant entered slow-wave sleep during a subsequent 90-min nap. (**D**) A stable reduction of implicit social bias, persisting 1 week later, was only achieved if the counterbias intervention was cued during the nap.

suggest that sleep is a brain state labilizing even strong habit-like implicit memories. This putatively makes them susceptible to modification through treatments targeting, in the first place, the explicit hippocampusdependent memory system.

Hu et al. certainly demonstrate the striking potential of targeted memory reactivation during sleep to modify deeply rooted attitudes. Together with the ongoing debate on the longevity and effectivity of social bias interventions (12), these findings raise a number of questions regarding the underlying neurophysiological and psychological mechanisms. One is why such enduring modifications cannot be achieved by targeted reactivations during wakefulness. Perhaps countering the preexisting racial or gender bias by memory reactivation during sleep leads to an immediate weakening of the original bias memory trace, indicating that during sleep even very old traces can reenter a labile state. Alternatively, the procedure might induce an extinction-like suppression of the original bias by a newly learned counterbias. Such new learning of response suppression implicates the activation of the machinery of synaptic plasticity during sleep. Here, another detail of Hu *et al.*'s findings is of interest, suggesting that the long-term improvement in social bias was additionally linked to posttraining REM sleep, which represents a state of increased synaptic plasticity. This indeed fits a "sequential" view on sleep's role in memory formation (4): Each bout of slowwave sleep and accompanying reactivation transforms the new counterbias traces and primes the participating synapses for lasting synaptic changes that are induced during the subsequent phase of REM sleep (e.g., by activation of immediate early genes and protein synthesis).

However, new learning of an extinctionlike suppression also suggests that effects might depend strongly on the learning context. If so, the risk of spontaneous recovery of the original social bias is increased when the person leaves the behavior modification setting, as well as with increasing time after the treatment. Unless the underlying mechanisms are understood, there is a risk that seemingly marginal changes in the procedures of targeted memory reactivation during sleep might produce precisely opposite results of the desired effects. Indeed, just such reversals have been seen in studies applying this method to counter fear responses (13). The possibility of such an outcome is further suggested by interventions that paradoxically increase social bias instead of reducing it (14).

There is little doubt that the study of Hu *et al.*—with its clear implications for society—will motivate research to resolve these remaining issues. However, Aldous Huxley's description of a dystopian "brave new world" where young children are conditioned to certain values during sleep reminds us that this research also needs to be guided by ethical considerations. Sleep is a state in which the individual is without willful consciousness and therefore vulnerable to suggestion. Beyond that, Hu *et al*.'s findings highlight the breadth of possible applications to permanently modify any unwanted behavior by targeted memory reactivation during sleep.

REFERENCES

- 1. N. Ellemers, M. Barreto, Curr. Opin. Behav. Sci. 3, 142 (2015).
- 2. A. G. Greenwald, M. R. Banaji, Psychol. Rev. 102, 4 (1995).
- 3. X. Hu et al., Science **348**, 1013 (2015).
- 4. S. Diekelmann, J. Born, Nat. Rev. Neurosci. 11, 218 (2010).
- B. Rasch *et al.*, *Science* **315**, 1426 (2007).
 T. Schreiner, B. Rasch, *Cereb. Cortex* 10.1093/cercor/
- Schlener, J. Rasch, Cereb. Contex 10:10007 cerebity bhu139 (2014).
 M. Schönauer et al., J. Cogn. Neurosci. 26, 143 (2014).
- M. Scholader et al., S. Cogn. Neurosci. 20, 145 (2014).
 J. D. Rudoy, J. L. Voss, C. E. Westerberg, K. A. Paller, Science 202, 1270 (2000).
- **326**, 1079 (2009). 9. Y. Dunham *et al.*, *Trends Cogn. Sci.* **12**, 248 (2008).
- K. K. Hauner et al., Nat. Neurosci. 16, 1553 (2003).
- 11. J. He et al., Sleep **38**, 423 (2015).
- E.L.Paluck, D.P. Green, Annu. Rev. Psychol. 60, 339 (2009).
- 13. S. Diekelmann, J. Born, *Sleep* **38**, 337 (2015).
- 14. L. Legault et al., Psychol. Sci. 22, 1472 (2011).

ILLUSTRATION: P. HUEY/SCIENCE