Neural Mechanisms of Attitude Change Toward Stigmatized Individuals: Temporoparietal Junction Activity Predicts Bias Reduction

Yoona Kang* and Emily B. Falk

Annenberg School for Communication, University of Pennsylvania

Author Note

This study uses a subset of a larger dataset from a study that examined an orthogonal question of health behavior change (Reported in Kang et al., 2018). No prior report examined the implicit attitudes toward stigmatized individuals that are the focus of the current manuscript. We thank Nicole Cooper, Christin Scholz, Prateekshit Pandey, Alison Elliott, and Elizabeth Beard for research assistance, Matthew Brook O'Donnell for assistance in data analysis, Bruce Doré and Emile Bruneau for helpful feedback, and the staff of the University of Pennsylvania fMRI Center for support and assistance. This research was supported by NIH New Innovator Award 1DP2DA03515601 (PI, E.B.F), NIH/National Cancer Institute Grant 1R01CA180015-01 (PI, E.B.F) and HopeLab (PIs, E.B.F, Y.K).

*Correspondence author

Annenberg School for Communication, University of Pennsylvania, 3620 Walnut street Philadelphia, PA 19104. Email: yoona.kang@asc.upenn.edu

Neural mechanisms of attitude change toward stigmatized individuals: Temporoparietal junction activity predicts bias reduction

Abstract

Objectives: Psychological and neural evidence suggests that negative attitudes toward stigmatized individuals arise in part from failures to perceive them as social targets. Here, we tested whether experimentally up-regulating neural regions involved in social cognition would predict subsequent decreases in bias toward stigmatized individuals (i.e., people who use substances). **Methods:** Participants underwent fMRI while completing either a lovingkindness intervention task or a control task, and each task was reinforced via daily text messages for a month following the one-time fMRI scan. Changes in implicit bias against stigmatized individuals were measured by Implicit Association Tests. **Results:** The lovingkindness intervention task, compared to a control task, elicited greater baseline activity in right temporoparietal junction (RTPJ), implicated in mentalizing, or the process of making inferences about others' mental states. The lovingkindness task compared to the control task also produced marginal decreases in bias over the month of the intervention. Individual differences in initial RTPJ activity at baseline during the fMRI intervention tasks further predicted improved implicit attitudes toward stigmatized individuals a month later. **Conclusions**: The current study suggests that individual differences in people's tendency to engage brain regions that support taking others' perspectives are associated with greater changes in bias reduction over time. It is possible that strategies that up-regulate mentalizing activity, such as lovingkindness training and other strategies that increase social-cognitive processing, may be effective in shifting people's biases against stigmatized individuals.

Neural mechanisms of attitude change toward stigmatized individuals: Temporoparietal junction activity predicts bias reduction

Highly stigmatized people such as those living with substance use disorders (Cuddy et al., 2008) are often subjected to unfair treatment in public, health care systems, law enforcement systems, and job decisions (Lloyd, 2013), which may deter treatment-seeking and recovery (Ahern et al., 2007; Luoma et al., 2007; Simmonds & Coomber, 2009; van Olphen et al., 2009). Neural data suggest that biases against stigmatized individuals are often characterized by a lack of social cognition such as mentalizing or empathy, such that the usual activation of social cognitive networks in the brain in response to social targets becomes absent or reduced in response to stigmatized others. For example, depictions of stigmatized individuals failed to elicit neural activity within the medial prefrontal cortex (MPFC), previously associated with social cognition such as understanding others' thoughts and feelings, and broader person perception (Van Overwalle, 2009), to the extent that other non-stigmatized individuals do (Harris & Fiske, 2011). Similarly, depictions of sexualized and objectified females elicited diminished activity within neural regions associated with mental state attribution, including MPFC, posterior cingulate, and temporal poles (Cikara et al., 2011). Conversely, observing or interacting with ingroup (vs. outgroup) members tends to elicit greater activation within regions implicated in social processing. For example, viewing the same (vs. other) race group members in pain recruited greater activity within key brain regions implicated in person perception, mentalizing, and empathy, including the MPFC (Mathur et al., 2010) and temporoparietal junction (TPJ) (Cheon et al., 2011). Further, interacting with an ingroup (vs. outgroup) member in prisoner's

dilemma games produced greater activity in the mentalizing network including right TPJ (RTPJ) and dorsomedial PFC (Rilling et al., 2008).

If reduced social cognition is a key component of negative person perception, then activating social cognition may be an effective strategy to improve attitudes toward stigmatized individuals. Consistent with this view, mentalizing, a particular form of social cognition that involves making inferences about others' mental states (Frith & Frith, 2006; Galinsky et al., 2008), can decrease prejudice (Galinsky & Ku, 2004; Todd et al., 2011; Vescio et al., 2003). Mentalizing might also be a key mechanism that underlies other bias reduction programs. For example, taking other people's perspectives mediated the effects of imagined contact interventions (Husnu & Crisp, 2015) in which individuals imagined having positive interactions with members of stigmatized groups (Crisp et al., 2009).

What types of interventions might boost mentalizing activity? Evidence suggests that lovingkindness practice is potentially one such intervention that can also alter bias against stigmatized individuals (Kang et al., 2014). Lovingkindness practice involves making positive well-wishes for others by considering what would alleviate suffering and bring happiness to them (Kang, 2018). As such, mentalizing and positive other-directed affect are intrinsic components of lovingkindness practice. Specifically, considering the needs and desires of others from the targets' perspectives is one of the main goals of the practice, and as might be expected, lovingkindness practice increased self-reported levels of mentalizing activity (Wallmark et al., 2013).

Further, lovingkindness practice can improve attitudes toward stigmatized others. A brief lovingkindness induction increased positive attitudes toward homeless people (Parks et al., 2014), and reduced implicit bias against members of a different racial group (Stell & Farsides,

2016). In a longitudinal trial, participants who were randomly assigned to complete a 6-week lovingkindness training, but not those who discussed ideas about lovingkindness and compassion for 6 weeks without practicing (active control) or wait-listed controls, showed significant reduction in implicit bias against homeless people over time (Kang et al., 2014).

Despite the collective evidence, however, the mechanisms by which lovingkindness interventions may reduce bias remain largely unclear. Linking initial intervention neural responses during lovingkindness practice at baseline to later attitude change can provide additional evidence to explain potential mechanisms of attitude change without relying solely on self-reports. Kang et al. (2018) found that compared to a control task, considering others' needs and desires through lovingkindness practice increased activity within RTPJ (see SI2 for results from the current sample). The TPJ is active in neuroimaging studies of mentalizing across tasks and laboratories with remarkable reliability (Frith & Frith, 2003). In particular, RTPJ is robustly involved in reasoning about the contents of others' minds (Dufour et al., 2013; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005; Scholz et al., 2009). Increased activity in RTPJ during lovingkindness practice is consistent with the idea that mentalizing activity might be one main component of lovingkindness practice. If mentalizing activity improves person perceptions and reduces bias, then increased RTPJ activity, associated with mentalizing, during initial lovingkindness practice at baseline may also lead to subsequent bias reduction over time. That is, participants' social understanding of others may improve through the lovingkindness intervention such that they are better able to take the views of a broad range of people, including stigmatized individuals whom they were previously biased against.

In the Kang et al. (2018) study, lovingkindness training also elicited activity in ventral striatum (VS), previously associated with positive valuation and reward (Bartra et al., 2013).

Lovingkindness and compassion practices are often described to be intrinsically positive (Singer & Klimecki, 2014), and increased positive other-directed processing across several studies (Galante et al., 2014). Further, compassion interventions recruited neural networks associated with social reward and prosocial behavior in response to social stimuli of suffering, including VS (Klimecki et al., 2014) and nucleus accumbens which predicted subsequent altruistic behavior (Weng et al., 2013). Therefore, VS activity during lovingkindness practice could also index processes relevant to bias, such as anticipated social reward. Therefore, in addition to considering others' mental states, positive value signals in the brain during early lovingkindness intervention at baseline might also be associated with later diminishing negative attitudes toward stigmatized individuals over time.

The current study examined mentalizing and positive valuation processing as possible neural mechanisms that support attitude change toward stigmatized individuals. Participants underwent fMRI completing either a lovingkindness intervention task or a control task, and their pre- to post-intervention changes in implicit attitudes toward culturally stigmatized individuals (i.e., people who use substances) were assessed using an implicit association task (IAT). Behaviorally, we tested the effect of initial lovingkindness intervention on subsequent changes in implicit attitudes. Using fMRI, we tested whether increases in RTPJ and VS activity, implicated in mentalizing and positive valuation, respectively, would be associated with subsequent bias reduction. Finally, we tested whether the effect of lovingkindness practice on bias reduction was mediated by differences in neural responses.

Method

Participants

Participants (*n*=132; mean age=34.27 years, *SD*=11.92; 87 females; 58 Black, 53 White, 9 Asians, 4 Hispanic, 8 Other; Table 1; See supplemental information [SI]1 for demographics of participants with usable behavioral and neural data) responded to an online advertisement for a study on "daily activities." Eligibility criteria, based on self-reports collected via online survey, included: 1) no current use of illicit drugs or psychotropic medications (in order to avoid self bias related to substance use and because drugs can alter brain function), 2) no history of serious psychiatric/medical conditions including substance use, and 3) standard fMRI scanning criteria (no metal in body, not claustrophobic, not pregnant/nursing, right-handed). Participants also reported whether they had prior experience with lovingkindness or compassion training, which was used as an exclusion criterion in later analyses. Eligibility criteria unrelated to the current report, but were included as part of the larger study (Kang et al., 2018), were engagement in less than 200 minutes of weekly physical activity and a body mass index (BMI) over 25. Research assistants contacted eligible participants via phone to reconfirm their eligibility and scheduled study visits.

Given that this work was conducted as part of a larger study related to health behavior change, the sample size for the control condition was determined by power analyses based on effect sizes found in prior work and related to the larger study on physical activity (Falk et al., 2015), and the sample size for the lovingkindness condition was determined by funding availability from an additional pilot grant. Due to these constraints, the lovingkindness condition had half as many participants (*n*=44) as the control (*n*=88) conditions. Despite the unequal sample sizes, we do not observe significant heteroscedasticity (Breusch-Pagan test *p*s>.05).

Participants were excluded from the neural data collection, neural outcome analyses, and/or behavioral outcome analyses for the following reasons: Failure to complete the fMRI

study appointment (*n*=12), frontal distortion (*n*=2), excessive motion (*n*=5; 10 or more 1mm spikes and/or 4mm or higher total displacement per run), technical difficulties in scanning (*n*=1), or ineligibilities discovered after the baseline visit, including metal in body (*n*=2) or brain abnormalities (*n*=6). Participants who did not complete the end-point study appointment (*n*=2) or are missing one or more IAT data points across three timepoints (*n*=11) were excluded from the relevant behavioral outcome analyses. A participant who reported of having had previous training in lovingkindness or compassion meditation was excluded (*n*=1). The rate of total data loss (*n*s=18, 24 for lovingkindness and control conditions, respectively) was equivalent across conditions (χ^2=1.925, *p*=.165).

Participants in the lovingkindness and control conditions did not significantly differ with respect to age, gender, ethnicity, or education (*p*s>.10) in analyses of behavioral data (*n*=110; mean age=34.68 years, *SD*=12.12; 72 females; 47 Black, 42 White, 9 Asians, 4 Hispanic, 8 Other). In analysis linking neural to behavioral data (*n*=95; mean age=33.76 years, *SD*=11.91; 65 females; 40 Black, 38 White, 6 Asians, 3 Hispanic, 8 Other), age was associated with condition (*p*=.01; Table SI1). When we controlled for age in all analyses linking neural to behavioral data, the main results remained parallel.

Procedure

Participants visited the laboratory for the pre-intervention appointment (T1), an fMRI intervention appointment (T2) approximately 10 days later, and a post-intervention appointment (T3) approximately 1 month after the fMRI intervention visit. At T1 baseline, all participants provided informed consent and completed value ranking (used for the control task at the fMRI appointment; see *control task* descriptions below) and a T1 baseline IAT. During the T2 fMRI intervention appointment, participants were randomly assigned to complete either a

lovingkindness or control task (there was also a third self-affirmation condition as part of a larger study that is not the focus of the current report; see Kang et al,., 2018) in an MRI scanner. The intervention tasks (lovingkindness, control) performed during the fMRI session served to assess participants' baseline response. Participants then completed an IAT task for the second time outside the scanner. During the T3 post-intervention appointment, participants completed an IAT task for the last time, were debriefed, paid, and thanked for their participation. All IAT and scanner tasks were embedded among other surveys and tasks as part of a larger investigation of health behavior change. Scanner tasks were presented using PsychoPy2 (Peirce, 2007). In-scanner responses were collected using a four-button response device attached to participant's right wrist.

Measures

Implicit attitudes. A modified version of an open-source IAT task was used (https://github.com/winteram/IAT) that followed standard IAT task procedures (Greenwald et al., 2003) to assess implicit attitudes toward individuals who use substances (vs. individuals who do not use substances). To create the substance use IAT, eight images that depict people using drugs and eight images that depict non-drug-using controls in a professional office environment were drawn from the IAPS image database (Lang et al., 1997), matched by gender (four males, four females), age (young to middle-aged), race (white), and neutral facial expressions. The images were paired with eight positive (beauty, good, great, happy, joy, laugh, love, peace) and eight negative (agony, awful, bad, evil, fail, gross, hurt, nasty) words, and repeated across 3 practice (19 trials each) and 4 main (39 trials each) blocks. In the main blocks, each participant completed 2 blocks in which the photos of substance users were paired with the positive words and photos of controls were paired with the negative words (incongruent blocks), and 2 blocks in which the

photos of the substance users were paired with negative words and the photos of controls were paired with positive words (congruent blocks). The logic of the IAT is that the extent of implicit bias is reflected in the difference in reaction times when the target category is paired with positive vs. negative words. For example, if people hold negative bias against people who use substances, then they will be slower to respond when photos of people who use substances are paired with the positive words, compared to when they are paired with the negative words.

IAT *d* scores were calculated following the standard guidelines (Greenwald et al., 2003): 1) delete trials longer than 10,000ms, 2) delete subjects who had more than 10% of trials have latency shorter than 300ms, 3) compute inclusive standard deviation for two congruent (bad/addict, good/professional) and for two incongruent (good/addict, bad/professional) blocks, 4) compute mean latencies for responses in each of the two congruent and two incongruent blocks, 5) compute the two mean latency differences by subtracting mean latency scores of each congruent block from mean latency scores of each incongruent block, 6) divide each difference score by its associated inclusive standard deviation, and 7) compute IAT *d* scores as the equal weight averages of the two resulting ratios, such that higher *d* scores indicate more bias against people who use substances versus those who are not addicted to substances, and therefore a decrease in *d* scores from T1 to T3 represents less bias overtime.

Demographics. At the end of the T1 pre-intervention appointment, participants reported their age, gender, ethnicity, and years of education. Participants in the lovingkindness and control conditions did not significantly differ with respect to age, gender, ethnicity, or education (*p*s>.10; Table 1).

fMRI scanner intervention tasks

The stimuli and detailed instructions for the fMRI scanner tasks are available at

https://github.com/cnlab/IAT/.

Lovingkindness intervention task. A short-term lovingkindness intervention task consisted of 20 lovingkindness trials and 20 everyday activity trials, presented across two runs (20 trials in each run) in a randomized order. In a lovingkindness trial, participants were instructed to make positive wishes for three target groups that varied in psychological closeness, including close others, acquaintances, and everyone in the world. Participants were first presented with an initial onset wish phrase (2s; "May you be at ease"). Next, they were presented with a target group to direct positive wishes to (10s; "Someone close to me"), followed by an importance rating (4s). Participants were instructed to think of what it would mean for the target group to have this wish come true (e.g., What would make your close friend to be at ease?) and imagine situations in which these wishes come true in the future as vividly as they could (e.g., A close friend relaxing on the beach). As within-subjects contrasts to facilitate fMRI analysis, control trials focused on everyday activities to allow comparisons of neural activity during interpersonal versus non-interpersonal wishing processes (e.g., "May it be done easily: Doing the laundry to clean clothes"). Trials were separated by fixation rest periods (3s); every fifth trial contained a longer (10s) period of rest (Figure 1).

Control task. The format of control task paralleled the main experimental manipulation of lovingkindness intervention task, and was adapted from self-affirmation literature that showed reflecting on unimportant values did not change participants' usual (without any manipulations) responses to emotional stimuli (Cohen & Sherman, 2014; Sherman & Cohen, 2006). This design also allowed us to match low-level properties of the lovingkindness task in terms of the brain's response to text, images, and psychological processing unrelated to our main study question (e.g., vividly imagining future situations). Specifically, at T1 all participants were presented with

six value types (compassion and kindness, family and friends, spirituality, wealth, power, fame), and ranked them in order of importance. At T2, participants in the control condition were guided through a control task in the fMRI scanner to reflect on their lowest ranked values determined at T1. Participants were instructed to think about situations, as vividly as they could, that pertain to a value that they ranked as the lowest of importance (e.g., if fame was their lowest value, one prompt might ask to imagine a situation when they would: "Become a local celebrity"). The majority of participants (91%) in the control condition ranked self-enhancing (i.e., wealth, power, status) to be the least important, and reflected on these values. However, even wealth, power and status can be used to benefit others (e.g., using your position as a local celebrity to help people), suggesting that this control represents a particularly conservative test of our hypothesis.

Forty trials (20 value trials, 20 everyday activity trials) were presented across two runs (20 trials in each run) in a randomized order. Each trial consisted of an initial onset trial type (low value/everyday activity; 2s), followed by the situation description (10s) and importance rating (4s). Trials were separated by fixation rest periods (3s); every fifth trial contained a longer (10s) period of rest. Within-subjects control trials included imagining value-neutral everyday activities, using the same activities that formed the within-subjects control trials in the lovingkindness intervention task (e.g., imagine a situation when they would: "do the laundry to clean clothes").

Previous work showed that state-like changes in attitudes are possible through a brief lovingkindness induction (Hutcherson et al., 2008). However, more stable changes may require repeated practice over time. Therefore, we reinforced both lovingkindness and control tasks with daily mobile text messages between the fMRI scan at T2 and endpoint T3 appointments, during

which we expected to observe a more stable shift in attitude.

The lovingkindness and control messages participants viewed in the scanner were reinforced via daily mobile text messages for 30 days following the T2 visit. Each day at a morning time participants chose, participants received either a lovingkindness or control message, depending on their condition assignment, encouraging them to make wishes for others (lovingkindness condition) or reflect on a life value they rated as unimportant (control condition). Participants were instructed to reply to the text messages by indicating how important it is for this wish to come true (lovingkindness condition) or how important the situation pertaining to an unimportant value is (control condition) on a scale of 1=not important at all to 5=very important. The mean response rate was 90.6 percent (*SD*=17.80). The lovingkindness and control text messages were drawn from the corresponding fMRI tasks (https://github.com/cnlab/IAT/).

Data Analysis

A series of models were computed to test the hypothesized relationships between the neural activity during the intervention tasks (lovingkindness, control) and subsequent changes in implicit biases against stigmatized individuals. In addition, a regression model tested the effect of intervention with the IAT *d* score at T3 as an outcome controlling for earlier IAT scores (T1, T2). For analyses linking neural data to IAT scores, we focused on regions of interest implicated in social cognition (mentalizing; RTPJ) and positive valuation processing (VS) that were activated by the lovingkindness intervention task (Kang et al., 2018), and relevant based on prior literature; additional regions associated with the intervention task are reported in SI3. Subsequent whole-brain analysis identified additional regions associated with implicit attitude change (SI4). The coefficient of determination (R^2, $R^2_{adjusted}$), beta coefficients (β), and 95% confidence

intervals (CI) are reported. All reported *p* values are two-tailed. All analyses were performed in R (v3.0.1, www.r-project.org) using the R-studio interface (v1.0.136).

fMRI data acquisition, preprocessing, and modeling. The imaging data were acquired on a 3 Tesla Siemens Trio scanner equipped with a 32 or 64 channel head coil. The head coil type was not associated with any of the main neural and behavioral outcome measures (*p*s>.20). Participants were self-guided through two runs of either the lovingkindness or control task (294 volumes each; 588 volumes total), embedded among four other tasks not reported here.

We collected high-resolution T1-weighted structural images using an MPRAGE sequence (TI=1,100ms, 160 slices, slice thickness=1mm, voxel size=0.9 × 0.9 × 1), and recorded T2*-weighted functional images (repetition time=1,500ms, echo time=25ms, flip angle=70°, -30° tilt relative to AC-PC line, 54 slices, field of view=200mm, slice thickness=3mm, multiband acceleration factor=2, voxel size=3.0 × 3.0 × 3.0 mm).

The anatomical and functional data were acquired and preprocessed using a standard processing stream in Statistical Parametric Mapping (SPM8; Wellcome Department of Cognitive Neurology, Institute of Neurology, London, UK) for all stages except for the initial despiking, which was carried out using the 3dDespike program as implemented in the AFNI toolbox. We corrected differences in time of acquisition using a sinc interpolation algorithm with the first slice as reference. Next, data were realigned spatially to the first slice of each volume, and co-registered to functional and structural images using two six-parameter affine stages. The average image across all blood oxygen level-dependent (BOLD) functional images was registered to high-resolution T1 images (total of 12 parameter affine).

Following co-registration, we segmented the high-resolution T1 images into gray matter, white matter and cerebrospinal fluid to create a brain mask used to determine voxels to be

included in first and second-level models. We then normalized structural and functional images to the skull-stripped MNI template ("MNI152_T1_1mm_brain.nii") provided by the FMRIB Software Library (FSL). In the final preprocessing step, we smoothed the functional images using a Gaussian kernel (8-mm FWHM). To allow for the stabilization of the BOLD signal, we discarded the first five volumes (7.5s) of each run before analysis. Movement parameters (a total of six rigid-body parameters, three for translation and three for rotation) derived from spatial realignment were included as nuisance regressors in all first-level models. Data were high pass filtered with a cutoff of 128s.

Fixed-effects models of the lovingkindness and control tasks were constructed using a boxcar function for each trial, specifying two trial types (lovingkindness/value trials, everyday activity trials). For the lovingkindness intervention task, a contrast between trials in which people made wishes for people vs. wishes for everyday activities was used. For the control task, a contrast between trials in which people reflected on their lowest values vs. thinking about everyday activities was used. Second-level random-effects models were constructed by averaging across participants and were subjected to further region of interest (ROI) analysis described below.

ROI selection and neural activity predicting bias reduction. We monitored ROIs implicated in mentalizing (RTPJ) and positive valuation (VS) during scanner tasks. We hypothesized that positive valuation processing should be relevant to lovingkindness practice, but the identification of the mentalizing ROI as key to lovingkindness practice was motivated by the whole-brain analysis that showed increased activity in RTPJ and VS during a lovingkindness intervention task (Kang et al., 2018). Thus, we formed our hypotheses for the current study after this information was known, but before any analyses were conducted linking brain data to

changes in attitudes or bias against stigmatized individuals. To monitor activity associated with mentalizing, an entire functional map of RTPJ was taken from a study (n=462) that examined neural responses during a task that required mental state inferences, compared to a task that does not (http://saxelab.mit.edu/use-our-theory-mind-group-maps) (Dufour et al., 2013). In addition, to monitor activity implicated in positive valuation, an entire map of VS was taken from a meta-analysis of 206 studies that reported valuation/reward-related neural signals (Bartra et al., 2013). Parameter estimates of activity during the 20 lovingkindness trials (lovingkindness intervention task) or 20 lowest value trials (control task) were compared with 20 everyday activity trials for each person within subjects, using MarsBaR (Brett et al., 2002), and converted to percent signal change. Then the activity scores within each ROI were used to predict changes in implicit bias using R, as described below.

Separate regression analyses tested the links between lovingkindness practice, neural activity, and changes in implicit bias. First, the effect of lovingkindness intervention on pre-to-post (T1-T3) changes in implicit bias against people who use substances were tested. Second, neural activity during the lovingkindness/control tasks was used to predict the subsequent changes in implicit bias. Third, the indirect effect of lovingkindness intervention on bias reduction through neural responses was tested.

Whole-brain analyses. Exploratory whole-brain searches were done for regions associated with intervention tasks (lovingkindness, control; SI3), as well as regions associated with pre- to post-intervention changes in implicit attitudes (SI4). Whole-brain analyses were corrected for multiple comparisons using a family-wise false discovery rate (FDR) with corrected p value of .05 and cluster-corrected at k=10. For the neural regions associated with later attitude change, no clusters survived FDR correction and results are reported at p<.005,

k=10.

Results

Effects of Lovingkindness Practice on Bias against Stigmatized Individuals

First, we examined the impact of a short-term lovingkindness intervention task on implicit attitudes toward people who use substances. At T1 baseline, there was no difference across the conditions on the substance use IAT scores (p=.88; indicating successful randomization), and participants across conditions tended to have negative implicit attitudes toward those who use substances (M=0.61, SD=0.39, $t(127)$=17.687, p<.001, comparing the mean to 0). We predicted that those in the lovingkindness condition would show a greater decline in implicit prejudice by the time of T3 post-intervention, compared to those in the control condition. To test this, regression analyses compared changes in the IAT *d* scores from pre- to post-intervention across conditions. After the intervention, at T3, those in the lovingkindness condition showed marginally less bias (M=0.54, SD=0.29) compared to controls (M=0.63, SD=0.33), R^2=.337, $R^2_{adjusted}$=.318, β=-.140, $t(106)$=-1.77, p=.080, 95% CI [-0.206, 0.018], controlling for the T1 and T2 IAT scores.

Neural responses predicting attitude change

Next, we tested whether initial neural responses during the intervention tasks at baseline (lovingkindness, control) predicted subsequent decreases in bias. Initial individual differences in brain activity during the intervention tasks within the ROIs that showed a main effect of the task, and were previously associated with mentalizing (RTPJ) and positive valuation/reward (VS), separately, were used as predictor variables (Figure 2). The pre (T1) to post (T3) change in IAT *d* scores was used as an outcome variable, with lower scores indicating greater decreases in bias,

to assess the combined effect of a brief lovingkindness fMRI intervention task as well as a longer-term exposure to lovingkindness intervention via text messages over a month period.

The degree of RTPJ activity, associated with mentalizing, during the intervention tasks predicted greater decreases in implicit bias, R^2=.111, β=-.333 , t(93)=-3.405, p<.001, 95% CI [-2.878, -0.758]. By contrast, activity in the VS, associated with valuation processing, did not predict later changes in implicit bias scores, R^2=.022, β=-.148, t(93)=-1.438, p=.154. Condition did not interact with RTPJ, R^2=.115, β=.085, t(91)=0.645, p=.521, or VS activity, R^2=.030, β=-.023, t(91)=-0.181, p=.857, in predicting changes in IAT scores (Figure 2).

Finally, we tested the indirect relationship between condition (lovingkindness vs control) and changes in pre- to post-intervention implicit attitudes through neural activity in RTPJ. When RTPJ activity and condition were considered simultaneously as predictors of implicit attitude change from T1 to T3, the condition effect on bias reduction reduced from marginal to non-significant, β=-.016, t(92)=-0.160, p=.873, while RTPJ activity significantly predicted changes in bias, β=-.329, t(92)=-3.235, p=.002, with increased RTPJ activity predicting decreased bias (Figure 3). The bootstrapped estimates of the indirect effect was significant, such that those in the lovingkindness condition, compared to controls, showed greater activity in RTPJ; in turn those who showed greater activity in RTPJ also showed greater reductions in biasover time (average causal mediation effect [ACME]; β=-.039, CI.(-.081, -.001), p=.01).

Discussion

Lovingkindness practice, compared to a control activity, was associated with greater increases in activity in a neural region associated with mentalizing (RTPJ). Differences in the degree of this initial activation at baseline were in turn associated with later decreases in implicit bias against people who use substances. A significant indirect effect of lovingkindness practice

on reduced bias, through brain activity, suggests that the mentalizing activity induced by lovingkindness practice, or the act of considering others' needs and desires, might be one key mechanism of how people change their views about stigmatized individuals.

Prior work showed that thinking about derogated outgroups elicited less brain activity in regions that support social cognition (Harris & Fiske, 2011). The current finding suggests that the reverse might also be true, such that increasing responsiveness in brain regions that support thinking about other people's minds and perspectives may improve implicit attitudes. This neurocognitive account for attitude change suggests that interventions designed to increase social perception, and mentalizing specifically, may decrease implicit bias. Further, mentalizing-based bias control interventions may complement previous models that primarily focused on reducing intergroup anxiety as a mechanism of change (Islam & Hewstone, 1993; Paolini et al., 2004; Voci & Hewstone, 2003). When shifting affective responses toward stigmatized individuals is difficult, upregulating social cognitive processes by mentalizing might be an effective alternative way to reduce bias.

RTPJ activity was most strongly associated with the lovingkindness intervention and implicit attitude change, whereas other regions commonly implicated in mentalizing were weakly or not significantly related in the current study. For example, a network of neural regions including dorsomedial PFC, right superior temporal sulcus, or posterior cingulate also tend to show greater activity during mentalizing, in addition to RTPJ (Dufour et al., 2013). Of all the regions associated with mentalizing, however, RTPJ has been identified as a unique neural substrate of mentalizing activity, such that RTPJ activity was specific to inferring mental states of others whereas other regions of mentalizing network was less so (Saxe & Wexler, 2005). Our data suggest that RTPJ activity is also most specifically responsive to lovingkindness practice

among all the regions in the mentalizing network, and that psychological processing most selectively associated with RTPJ might be the active ingredient of lovingkindness interventions that alters person perception. However, the current data cannot directly speak to this possibility, and we encourage future researchers to examine the unique role of RTPJ while people engage in lovingkindness practice and in the context of interpersonal attitude change.

By contrast, increased VS activity alone did not predict bias reduction, although activity within substantia nigra/ventral tegmental area (SN/VTA), a region that is also implicated in reward processing, was associated with changes in implicit bias, identified by using a liberal threshold for a whole brain search (p=.005, k=10; SI4). As an exploratory follow-up, we conducted an ROI analysis testing whether the activity in the SN/VTA cluster predicted changes in bias. We found that greater activity within the functional cluster of SN/VTA also marginally predicted greater pre to post (T1-T3) bias reduction, R^2=.033, β=-.182, t(93)=-1.782, p=.078. In previous work, effects of positive affect/reward processing on interpersonal bias are largely mixed. On one hand, positive affect can promote inclusivity and decrease bias (Dunn & Schweitzer, 2005; Dovidio et al., 1995; Johnson & Fredrickson, 2005; Waugh & Fredrickson, 2006). On the other, it can also increase the use of stereotypic heuristics (Bodenhausen et al., 1994; Isbell, 2004). In one study, positive reward actually decreased mentalizing, such that sexist males showed diminished activity in regions associated with mental state attribution while viewing sexualized females and putatively feeling positive toward them (Cikara et al., 2011). The current findings suggest that brain activity associated with positive valuation during lovingkindness training, independent of activity implicated in considering others' minds, may not be sufficient to alter previously established biases.

Participants' initial neural responses to intervention tasks at baseline predicted reduction in bias a month later regardless of the condition assignment, suggesting that individual differences during early intervention tasks might predict who may potentially benefit from a longer one-month intervention (e.g., Mascaro et al., 2013). It is possible, for example, that individuals who show greater sensitivity in RTPJ when they make positive wishes for others (lovingkindness condition), or when they reflect on certain values that are potentially interpersonal, even when these may not be their top values, might benefit more in terms of bias reduction a month later. That is, even in the control condition, daily practice reflecting on values could have benefits that parallel lovingkindness intervention for individuals who are naturally inclined to taking other's perspectives.

Exploratory whole-brain analysis (see SI4) did not identify extensive activations beyond our primary regions of interest. However, analyses with a liberal threshold (p<.005, k=10) suggested some regions that may further be associated with decreases in implicit bias, which included areas implicated in mentalizing (bilateral TPJ), further supporting the role of these processes in attitude change, as well as bilateral fusiform gyri near fusiform face area (FFA), which responds to face-related cues (Kanwisher et al., 1997; McCarthy et al., 1997) and biological motion (Grossman et al., 2004). The FFA activity in the absence of explicit biological stimuli might reflect visualization of target people during the lovingkindness intervention task, or it might suggest that the role of FFA potentially extends beyond detecting biological information (Schultz et al., 2003) and might involve broader social processing.

Limitations and Future Research Directions

Although the current study had several strengths, including longitudinal measurement of implicit attitudes combined with neuroimaging, the results should be interpreted in the context of

limitations inherent in the study design. First, while IAT is a widely used measure of implicit attitudes that does not rely on self-reports, IAT scores are sometimes weakly correlated with actual behavior toward others (Connor et al., 2019; Oswald et al., 2013), and might be subject to desirability issues (Boysen et al., 2006). Social neuroscience has made progress on linking brain activity to behavioral outcomes, such as physical activity (Falk et al., 2015), smoking (Chua et al., 2011; Falk et al., 2011; Wang et al., 2013), and sunscreen use (Burns et al., 2018; Falk et al., 2010, 2011; Vezich et al., 2017); likewise, more research that links brain response to real world social behaviors and particularly behaviors toward stigmatized groups (Richeson et al., 2003) will be of great use moving forward.

Second, although the indirect effect of lovingkindness practice on attitude change was significant, the direct effect was only marginal, and the effect size of the intervention was small (Cohen's d=.20), especially compared to more intensive course-based lovingkindness intervention effects on implicit bias (e.g., mean Cohen's d=0.61 in Kang et al., 2014). It is possible that more intensive training would have produced stronger effects, or that our control condition creates a very conservative test of our hypothesis. Broader data on prosocial effects of lovingkindness intervention is increasingly promising (Luberto et al., 2018), yet inconclusive; a recent meta-analysis reported that social outcomes of various meditation interventions are specific to types of prosociality and methodological quality (Kreplin et al., 2018). The varying magnitudes of lovingkindness intervention effects highlight the need for a more precise delineation of boundary conditions and clarifications of specific mechanisms through which lovingkindness meditation practices may enhance social outcomes. For example, it is possible that longer-term interventions, beyond brief laboratory induction followed by daily text messaging, might produce more stable changes in RTPJ that may be detected in resting state.

Therefore, we encourage future studies to identify effects of longer-term interventions on more stable changes in neural activity, and multiple components and mechanisms through which lovingkindness practice may enhance social outcomes to more precisely test effects of lovingkindness interventions.

Third, psychological processes inferred from neuroimaging data should be interpreted with caution. For example, TPJ has been associated with attentional processes in addition to social cognition (Corbetta & Shulman, 2002), and the VS is associated with functions beyond reward processing and coordinates multiple aspects of cognition, including motivation (Cardinal et al., 2002) and decision-making (Balleine et al., 2007). However, taken together with the behavioral results on subsequent bias reduction, we offer the interpretation related to social cognitive processes as a theoretically motivated connection to the broader literature.

The current study provides neural evidence that lovingkindness training increased activity in RTPJ, previously associated with mentalizing, which in turn predicted implicit attitude change toward stigmatized individuals. Bias reduction interventions may consider employing strategies that boost mentalizing signals, such as lovingkindness practice used here, in order for people to more effectively change their views about stigmatized others.

Compliance with Ethical Standards

Conflict of Interest

The authors declare that they have no conflicts of interest.

Ethics Statement

The study was approved by the University of Pennsylvania Institutional Review Board.

Informed Consent

All participants provided informed consent..

Author Contributions

YK: designed research, performed research, analyzed data, and wrote the paper. EBF: designed research, wrote the paper, and oversaw the project.

References

Ahern, J., Stuber, J., & Galea, S. (2007). Stigma, discrimination and the health of illicit drug users. *Drug and Alcohol Dependence*, *88*(2-3), 188–196.

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(31), 8161–8165.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427.

Bodenhausen, G. V., Kramer, G. P., & Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, *66*(4), 621.

Boysen, G. A., Vogel, D. L., & Madon, S. (2006). A public versus private administration of the implicit association test. *European Journal of Social Psychology*, *36*(6), 845–856.

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *NeuroImage*, *16*(2), S497.

Burns, S. M., Barnes, L. N., Katzman, P. L., Ames, D. L., Falk, E. B., & Lieberman, M. D. (2018). A functional near infrared spectroscopy (fNIRS) replication of the sunscreen persuasion paradigm. *Social Cognitive and Affective Neuroscience*, *13*(6), 628–636.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews*, *26*(3), 321–352.

Cheon, B. K., Im, D. M., Harada, T., Kim, J. S., Mathur, V. A., Scimeca, J. M., Parrish, T. B., Park, H. W., & Chiao, J. Y. (2011). Cultural influences on neural basis of intergroup empathy. *NeuroImage*, *57*(2), 642–650.

Chua, H. F., Ho, S. S., Jasinska, A. J., Polk, T. A., Welsh, R. C., Liberzon, I., & Strecher, V. J. (2011).

Self-related neural response to tailored smoking-cessation messages predicts quitting. *Nature Neuroscience*, *14*(4), 426–427.

Cikara, M., Eberhardt, J. L., & Fiske, S. T. (2011). From agents to objects: Sexist attitudes and neural responses to sexualized targets. *Journal of Cognitive Neuroscience*, *23*(3), 540–551.

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology*, *65*, 333–371.

Connor, P., Sarafidis, V., Zyphur, M. J., Keltner, D., & Chen, S. (2019). Income inequality and white-on-black racial bias in the United States: Evidence from Project Implicit and Google Trends. *Psychological Science*, *30*(2), 205–222.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215.

Crisp, R. J., Stathi, S., Turner, R. N., & Husnu, S. (2009). Imagined intergroup contact: Theory, paradigm and practice. *Social and Personality Psychology Compass*, *3*(1), 1–18.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). Academic Press.

Dovidio, J. F., Gaertner, S. L., Isen, A. M. & Lowrance, R. (1995). Group representations and intergroup bias: Positive affect, similarity, and group size. *Personality & Social Psychology Bulletin*, *21*(8), 856–865.

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D. E., & Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PloS One*, *8*(9), e75468.

Dunn, J., & Schweitzer, M. E. (2005). Why good employees make unethical decisions: The role of reward systems, organizational culture, and managerial oversight. *Managing Organizational Deviance*, 39–68.

Falk, E. B., Berkman, E. T., Mann, T., Harrison, B., & Lieberman, M. D. (2010). Predicting persuasion-

induced behavior change from the brain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(25), 8421–8424.

Falk, E. B., Berkman, E. T., Whalen, D., & Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *30*(2), 177–185.

Falk, E. B., O'Donnell, M. B., Cascio, C. N., Tinney, F., Kang, Y., Lieberman, M. D., Taylor, S. E., An, L., Resnicow, K., & Strecher, V. J. (2015). Self-affirmation alters the brain's response to health messages and subsequent behavior change. *Proceedings of the National Academy of Sciences*, *112*(7), 1977–1982.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1431), 459–473.

Galante, J., Galante, I., Bekkers, M. J., & Gallacher, J. (2014). Effect of kindness-based meditation on health and well-being: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *82*(6), 1101–1114.

Galinsky, A. D., & Ku, G. (2004). The effects of perspective-taking on prejudice: The moderating role of self-evaluation. *Personality & Social Psychology Bulletin*, *30*(5), 594–604.

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science*, *19*(4), 378–384.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216.

Grossman, E. D., Blake, R., & Kim, C. Y. (2004). Learning to see biological motion: Brain activity parallels behavior. *Journal of Cognitive Neuroscience*, *16*(9), 1669–1679.

Harris, L. T., & Fiske, S. T. (2011). Perceiving humanity or not: A social neuroscience approach to

dehumanized perception. *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*, 123–134.

Husnu, S., & Crisp, R. J. (2015). Perspective-taking mediates the imagined contact effect. *International Journal of Intercultural Relations: IJIR*, *44*, 29–34.

Hutcherson, C. A., Seppala, E. M., & Gross, J. J. (2008). Loving-kindness meditation increases social connectedness. *Emotion* , *8*(5), 720–724.

Isbell, L. M. (2004). Not all happy people are lazy or stupid: Evidence of systematic processing in happy moods. *Journal of Experimental Social Psychology*, *40*(3), 341–349.

Islam, M. R., & Hewstone, M. (1993). Dimensions of contact as predictors of intergroup anxiety, perceived out-group variability, and out-group attitude: An integrative model. *Personality & Social Psychology Bulletin*, *19*(6), 700–710.

Johnson, K. J. & Fredrickson, B. L. (2005). "We all look the same to me": Positive emotions eliminate the own-race bias in face recognition. *Psychological Science*, *16*(11), 875–881.

Kang, Y. (2018). Examining interpersonal self-transcendence as a potential mechanism linking meditation and social outcomes. *Current Opinion in Psychology*, *28*, 115–119.

Kang, Y., Cooper, N., Pandey, P., Scholz, C., O'Donnell, M. B., Lieberman, M. D., Taylor, S. E., Strecher, V. J., Dal Cin, S., Konrath, S., Polk, T. A., Resnicow, K., An, L., & Falk, E. B. (2018). Effects of self-transcendence on neural responses to persuasive messages and health behavior change. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(40), 9974–9979.

Kang, Y., Gray, J. R., & Dovidio, J. F. (2014). The nondiscriminating heart: Lovingkindness meditation training decreases implicit intergroup bias. *Journal of Experimental Psychology. General*, *143*(3), 1306–1313.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: aAmodule in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(11), 4302–4311.

Klimecki, O. M., Leiberg, S., Ricard, M., & Singer, T. (2014). Differential pattern of functional brain plasticity after compassion and empathy training. *Social Cognitive and Affective Neuroscience*, *9*(6), 873–879.

Kreplin, U., Farias, M., & Brazil, I. A. (2018). The limited prosocial effects of meditation: A systematic review and meta-analysis. *Scientific Reports*, *8*(1), 2403.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 39–58.

Lloyd, C. (2013). The stigmatization of problem drug users: A narrative literature review. *Drugs: Education, Prevention and Policy*, *20*(2), 85–95.

Luberto, C. M., Shinday, N., Song, R., Philpotts, L. L., Park, E. R., Fricchione, G. L., & Yeh, G. Y. (2018). A systematic review and meta-analysis of the effects of meditation on empathy, compassion, and prosocial behaviors. *Mindfulness*, *9*(3), 708–724.

Luoma, J. B., Twohig, M. P., Waltz, T., Hayes, S. C., Roget, N., Padilla, M., & Fisher, G. (2007). An investigation of stigma in individuals receiving treatment for substance abuse. *Addictive Behaviors*, *32*(7), 1331–1346.

Mascaro, J. S., Rilling, J. K., Negi, L. T., & Raison, C. L. (2013). Pre-existing brain function predicts subsequent practice of mindfulness and compassion meditation. *NeuroImage*, *69*, 35–42.

Mathur, V. A., Harada, T., Lipke, T., & Chiao, J. Y. (2010). Neural basis of extraordinary empathy and altruistic motivation. *NeuroImage*, *51*(4), 1468–1475.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, *9*(5), 605–610.

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*(2), 171–192.

Paolini, S., Hewstone, M., Cairns, E., & Voci, A. (2004). Effects of direct and indirect cross-group friendships on judgments of Catholics and Protestants in Northern Ireland: The mediating role of an

anxiety-reduction mechanism. *Personality & Social Psychology Bulletin*, *30*(6), 770–786.

Parks, S., Birtel, M. D., & Crisp, R. J. (2014). Evidence that a brief meditation exercise can reduce prejudice toward homeless people. *Social Psychology*, *45*(6), 458–465.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, *6*(12), 1323–1328.

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *NeuroImage*, *41*(4), 1447–1461.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–1399.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PloS One*, *4*(3), e4869.

Schultz, R. T., Grelotti, D. J., Klin, A., Kleinman, J., Van der Gaag, C., Marois, R., & Skudlarski, P. (2003). The role of the fusiform face area in social cognition: Implications for the pathobiology of autism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1430), 415–427.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In *Advances in Experimental Social Psychology,* (Vol. 38, pp. 183–242). Academic Press.

Simmonds, L., & Coomber, R. (2009). Injecting drug users: A stigmatised and stigmatising population. *The International Journal on Drug Policy*, *20*(2), 121–130.

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current Biology: CB*, *24*(18), R875–

R878.

Stell, A. J., & Farsides, T. (2016). Brief loving-kindness meditation reduces racial bias, mediated by positive other-regarding emotions. *Motivation and Emotion*, *40*(1), 140–147.

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, *100*(6), 1027–1042.

van Olphen, J., Eliason, M. J., Freudenberg, N., & Barnes, M. (2009). Nowhere to go: How stigma limits the options of female drug users after release from jail. *Substance Abuse Treatment, Prevention, and Policy*, *4*, 10.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858.

Vescio, T. K., Sechrist, G. B., & Paolucci, M. P. (2003). Perspective taking and prejudice reduction: The mediational role of empathy arousal and situational attributions. *European Journal of Social Psychology*, *33*(4), 455–472.

Vezich, I. S., Katzman, P. L., Ames, D. L., Falk, E. B., & Lieberman, M. D. (2017). Modulating the neural bases of persuasion: Why/how, gain/loss, and users/non-users. *Social Cognitive and Affective Neuroscience*, *12*(2), 283–297.

Voci, A., & Hewstone, M. (2003). Intergroup contact and prejudice toward immigrants in Italy: The mediational role of anxiety and the moderational role of group salience. *Group Processes & Intergroup Relations: GPIR*, *6*(1), 37–54.

Wallmark, E., Safarzadeh, K., Daukantaitė, D., & Maddux, R. E. (2013). Promoting altruism through meditation: An 8-week randomized controlled pilot study. *Mindfulness*, *4*(3), 223–234.

Wang, A. L., Ruparel, K., Loughead, J. W., Strasser, A. A., Blady, S. J., Lynch, K. G., Romer, D., Cappella, J. N., Lerman, C., & Langleben, D. D. (2013). Content matters: Neuroimaging investigation of brain and behavioral impact of televised anti-tobacco public service announcements. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(17), 7420–

7427.

Waugh, C. E., & Fredrickson, B. L. (2006). Nice to know you: Positive emotions, self–other overlap, and complex understanding in the formation of a new relationship. *The Journal of Positive Psychology*, *1*(2), 93–106.

Weng, H. Y., Fox, A. S., Shackman, A. J., Stodola, D. E., Caldwell, J. Z. K., Olson, M. C., Rogers, G. M., & Davidson, R. J. (2013). Compassion training alters altruism and neural responses to suffering. *Psychological Science*, *24*(7), 1171–1180.

Table 1.
Baseline demographic characteristics by condition

	Loving-kindness (*n*=44)	**Control** (*n*=88)	**Statistic** (*p*)
Age (yrs)	31.91 (11.20)	35.45 (12.16)	$F = 2.63$ (.11)
Female	27 (61.4%)	60 (68.2%)	$\chi^2 = 0.34$ (.56)
Black	21 (47.7%)	37 (42.1%)	$\chi^2 = 0.19$ (.66)
Education (yrs)	15.80 (2.83)	15.81 (3.05)	$F = 0.001$ (.98)
Implicit bias (*d*)	0.60 (0.38)	0.62 (0.40)	$F = 0.06$ (.80)

Note: Mean values and sample sizes are displayed with standard deviations and percentages, respectively, in parentheses where applicable. See SI1 for demographics of participants with usable behavioral and neural data.

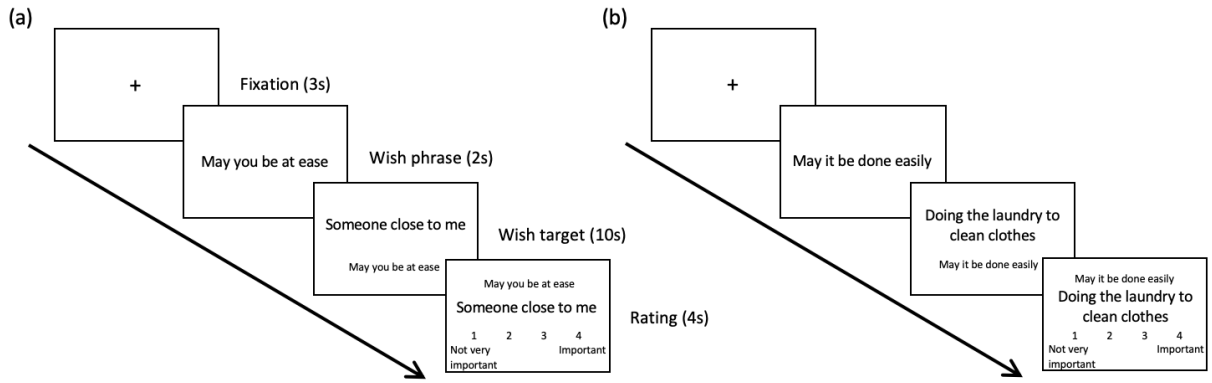

Figure 1. A short-term lovingkindness intervention fMRI task trial types; **(a)** A lovingkindness trial; **(b)** An everyday activity trial.

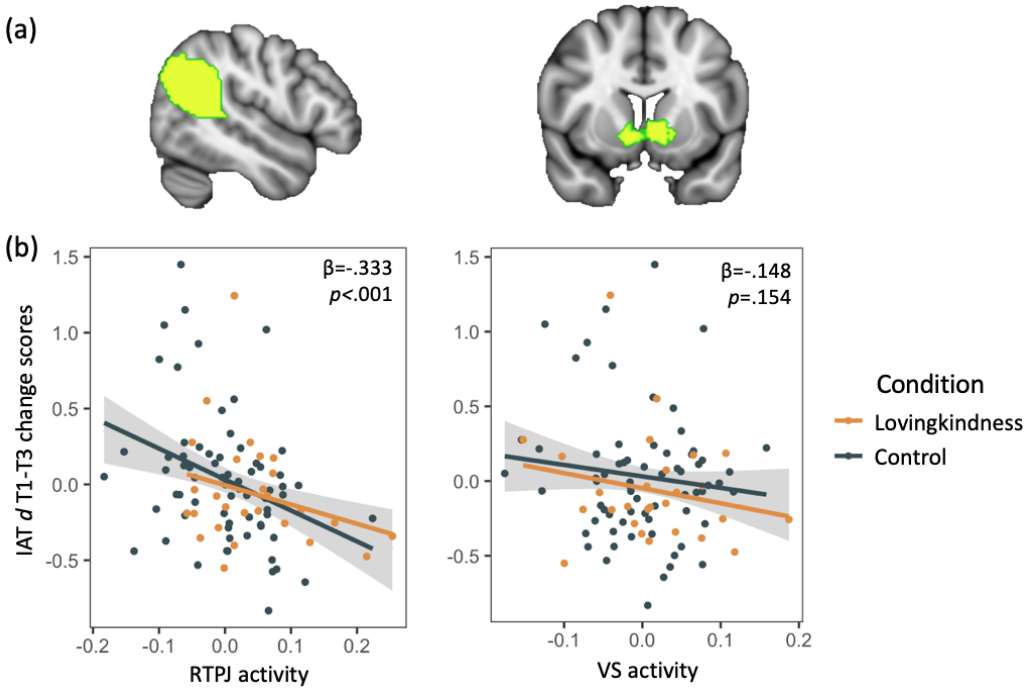

Figure 2. Neural activity during initial intervention tasks at baseline predicting subsequent changes in IAT scores. **(a)** RTPJ and VS regions of interest chosen for their role in the task and previously associated with mentalizing (Dufour et al., 2013) and positive valuation (Bartra et al., 2013), respectively. **(b)** Neural activity within the RTPJ ROI predicted decreases in implicit bias against individuals who use substances, whereas VS activity was not associated with changes in implicit bias scores. Condition did not interact with RTPJ in predicting changes in IAT scores. Notes: RTPJ = right temporoparietal junction; VS = ventral striatum

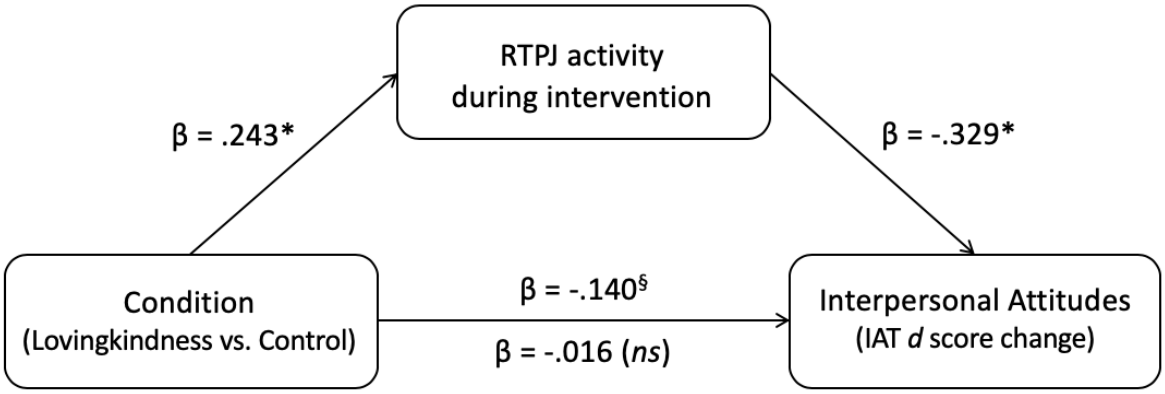

Figure 3. The condition indirectly predicted changes in T1 to T3 IAT scores via RTPJ activity. §<.01, *<.05