Neural signals in amygdala predict implicit prejudice toward an ethnic outgroup

Abbreviated Title: Decoding ethnic prejudice

Keise Izuma[1,*], Ryuta Aoki[2], Kazuhisa Shibata[3] & Kiyoshi Nakahara[2]

[1] Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.
[2] Research Center for Brain Communication, [4]School of Information, Kochi University of Technology, Kami, Kochi 782-8502, Japan.
[3] Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

* Correspondence should be addressed to Keise Izuma,

Current address

Department of Psychology, University of Southampton,

University Road, Southampton, SO17 1BJ, UK.

Email: K.Izuma@soton.ac.uk

Conflict of Interest

There is no conflict of interest.

Abstract

Racial and ethnic prejudice is one of the most pressing problems in modern societies. Although previous social neuroscience research has suggested the amygdala as a key structure in racial prejudice, it still remains elusive whether the amygdala activity reflects negative attitudes toward an outgroup or other unrelated processes. The present study aims to rigorously test the role of the amygdala in negative prejudice toward an outgroup. Seventy Japanese individuals passively viewed images related to an ethnic outgroup (South Korea) inside a functional magnetic resonance imaging scanner. Using Multi-Voxel Pattern Analysis (MVPA), we found that Japanese individuals' level of implicit (but not explicit) evaluations of South Korea could be predicted from neural signals in the left amygdala. Our result further suggested that the medial and lateral parts of amygdala play different roles in implicit evaluations. In contrast to the MVPA findings, conventional univariate analyses failed to find any reliable relationship between brain activation and both implicit and explicit evaluations. Our findings provide evidence for the amygdala's role in representing an implicit form of prejudice and highlight the utility of the multivariate approach to reveal neural signatures of this complex social phenomenon.

Keywords: fMRI, MVPA, IAT, implicit attitude

Highlights

- Amygdala is considered as a candidate region involved in negative prejudice
- However, past neuroimaging studies on prejudice have generated mixed findings
- We investigated the neural signatures of ethnic prejudice using MVPA
- Our results demonstrated the link between left amygdala and implicit evaluations

Introduction

Historically, prejudice has been of primary interest to social scientists for a long time. Despite the long history, research on intergroup relations remains of high importance as intergroup relations continue to be fluid and volatile in modern societies. In the past two decades, social neuroscientists have investigated neural mechanisms of prejudice by using neuroimaging methods (Amodio, 2014; Chekroud et al., 2014; Kubota et al., 2012). However, this endeavor has turned out to be challenging due to the high complexity of the phenomenon.

Because of its well known role in fear learning (Fendt and Fanselow, 1999; Pape and Pare, 2010), amygdala is considered as a primary candidate as a neural substrate underlying negative prejudice (Phelps et al., 2000). In fact, the amygdala has been most frequently reported in previous neuroimaging research on racial prejudice (Amodio, 2014; Chekroud et al., 2014; Kubota et al., 2012). However, past findings are not necessarily consistent, and the amygdala's involvement in negative prejudice toward an outgroup still remains unclear. For example, while some studies found increased amygdala activation in response to racial outgroup faces compared to ingroup faces (Cunningham et al., 2004; Hart et al., 2000; McCutcheon et al., 2018; Wheeler and Fiske, 2005), a number of other studies failed to find such amygdala activations (Brosch et al., 2013; Cassidy and Krendl, 2016; Gilbert et al., 2012; Golby et al., 2001; Li et al., 2016; Mattan et al., 2018; Phelps et al., 2000; Richeson et al., 2003; Stanley et al., 2012; Terbeck et al., 2015). Furthermore, even among those studies that found amygdala activations, its functional interpretations differ. Whereas some studies have provided evidence that amygdala activation in response to outgroup faces reflects negative emotional reaction to an outgroup (i.e., culturally learned negative association) (Lieberman et al., 2005; Phelps et al., 2000; Telzer et al., 2013b), other studies demonstrated that amygdala activity simply reflects novelty of outgroup faces (Cloutier et al., 2014; Hart et al., 2000; Telzer et al., 2013a). Still other studies showed that amygdala activity is modulated by skin-tone (Ronquillo et al., 2007), gaze direction (Richeson et al., 2008) and status (Mattan et al., 2018). Furthermore, it is known that amygdala activity is sensitive to facial features such as subtle differences in pupil size (Demos et al., 2008), trustworthiness (Said et al., 2009; Winston et al., 2002) and general valence evaluation of faces (Todorov and Engell, 2008). These findings suggest that the amygdala activations in response to outgroup faces found in the previous studies (Cunningham et al., 2004; Hart et al., 2000; McCutcheon et al., 2018; Wheeler and Fiske, 2005) might not be directly related to prejudice toward an outgroup.

There exist three studies (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000) that reported a correlation between amygdala activity and individual differences in implicit attitudes toward an outgroup as measured by implicit measures of attitudes such as an Implicit Association Test (IAT) (Greenwald et al., 1998). However, these across-participant correlations were based on a considerably small sample size (n = 12-13) (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000), and six other studies with larger sample sizes (n = 15-44) failed to find such a link (Cassidy and Krendl, 2016; Cassidy et al., 2016; Gilbert et al., 2012; Li et al., 2016; Richeson et al., 2003; Terbeck et al., 2015), suggesting that these findings need to be interpreted with extreme caution (see Power Analysis below for more discussion).

Thus, although several previous studies have reported amygdala activation in response to outgroup faces, evidence for the amygdala's involvement in prejudice is still weak. Furthermore, evidence for the involvement of other brain regions (such as the anterior insula, striatum and fusiform face area [FFA]) is, if anything, even more mixed (Amodio, 2014; Kubota et al., 2012). Given the complex and multifaceted nature of prejudice, the field requires a more powerful approach to unveil its neural signatures, especially the role of the amygdala in racial and ethnic prejudice.

In the present study, rather than simply contrasting brain activations in response to outgroup vs. ingroup faces where observed activations can be explained by a variety of different factors as discussed above, we directly tested if negative evaluations of an outgroup as measured by the IAT are related to neural activations, especially in the amygdala. The IAT measures implicit attitude as the strength of associations between representations of groups and valenced semantic concepts (Greenwald et al., 1998). Thus, if the amygdala reflects an *automatic* negative evaluation of an outgroup, which has been acquired via direct or observational fear learning processes (i.e., repeated associations with a social group and negatively-valenced information), its activities should be associated with IAT scores.

To test the role of the amygdala in negative prejudice toward an outgroup, we employed a machine learning method (Multi-Voxel Pattern Analysis; MVPA (Haynes, 2015; Norman et al., 2006)). MVPA makes it possible to detect a wider variety of signals and has been proven to be effective for detecting differences in cognitive or affective states that cannot be probed by conventional univariate analysis (e.g., (Izuma et al., 2017; Jimura and Poldrack, 2012; Sapountzis et al., 2010) and thus is more suitable for investigating complex neural representations of prejudice. Seventy Japanese university students passively viewed Japan- and South Korea-related images (Figure 1) inside an fMRI scanner. After the scanning, they completed explicit and implicit measures of attitudes toward each of Japan and South Korea. Using the MVPA, we tested if neural signals especially in the amygdala can predict individuals’ level of implicit and explicit negative evaluations of South Korea.

The present study focuses on the intergroup relation between Japan and South Korea. Despite that prejudice is a world-wide problem (Landis and Albert, 2012; Noor and Montiel, 2009), the vast majority of past neuroimaging studies on prejudice focused on the intergroup relation between White vs. Black Americans, and prejudice in other intergroup contexts has been rarely investigated (see (Bruneau and Saxe, 2010; McCutcheon et al., 2018) for notable exceptions). Fiske (2017) recently argued that stereotypes of race, ethnicity and region are more variable across different cultures compared to stereotypes of gender and age, and thus it is important to formally test whether past social neuroscience findings only apply to the White vs. Black Americans intergroup context or can be generalizable to other contexts. The relationship between the two neighboring countries of Japan and South Korea has deteriorated especially in recent years due to several disputes over political, historical and territorial issues. For example, the recent analysis of comments on a Japanese online news site showed that among 1,000 randomly-selected comments about South Korea, 87.6% expressed negative attitudes toward South Korea, while only 0.7% expressed positive attitudes (11.7% neutral) (Cho, 2017). BBC World Service polls taken between 2010-2014 have also shown that Japanese viewed South Korea more and more negatively over the past years (BBC_World_Service, 2010, 2014). Japanese participants' negative implicit and explicit attitudes toward South Korea were also confirmed by our behavioral data (see below). Thus, the current relationship between the two countries is an ideal intergroup context to explore the neural signatures of prejudice outside of the White vs. Black American context.

Materials and Methods

Participants: Seventy-one right-handed Japanese university students aged 18-22 years with no psychiatric history participated the study. One participant was excluded from the analyses due to excessive error rate during the IAT (> 25%). The final sample consists of n = 70 (27 females, mean age = 18.9 years, SD = 1.11). Participants were recruited from a subject pool of the Kochi University of Technology. All participants gave written informed consent for participation, and ethics approval for the study was granted by the Kochi University of Technology Ethics Board.

Power Analysis: The three previous studies (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000) that reported significant correlation between amygdala activities and race IAT scores reported correlations of r = 0.58 (Phelps et al., 2000) r = 0.71 (Cunningham et al., 2004) and r = 0.62 (Brosch et al., 2013). However, due to a small sample size (n = 12, 13 and 13 respectively) together with likely large measurement error of implicit measures of attitudes (e.g., Nosek et al. (2007) reported that the median test-retest reliability of IAT is 0.56), these correlations are highly likely to be inflated (Loken and Gelman, 2017; Yarkoni, 2009). Thus, we estimated the effect size of r = 0.30 (a correlation between actual IAT scores and predicted scores based on neural activation patterns in the amygdala; see "fMRI Data Analysis (MVPA)" below for more details) for the present study (i.e., about half of the smallest of the three correlations). With this effect size, a sample size of n = 68 should achieve statistical power of 80% (β = 0.2) with α = 0.05 (one-tailed). To account for potential data loss (e.g., due to excessive head motion inside an fMRI scanner), we aimed to recruit 70 participants (and actually recruited 71 participants).

Stimuli: Inside an fMRI scanner, participants were presented with 20 pictures each for the Japan and South Korea conditions. The images depicted Japanese or Korean people and/or national flags (e.g., South Korea fans chanting and cheering their national team on street, South Korea national football team approaching fans after winning a match, South Korea national flag overlaid onto a map of South Korea, etc.) so that it is evident which country each picture represents (based on uniforms they wear and/or national flags). It should be noted that although there are images that depict sports scenes, each image depicted either Japan or South Korea only, and no picture depicted, for example, a scene where the two countries compete with each other. All pictures were taken from the Internet (i.e., Google image search). Unlike the majority of past studies, we avoided using faces as experimental stimuli to minimize the effect of individual differences in facial trustworthiness/attractiveness judgment on the activity of the amygdala (and other brain regions) (Todorov and Engell, 2008). Furthermore, since ethnic prejudice is based on evaluations of an ethnic group, stimuli symbolizing South Korea as a whole (rather than each unfamiliar Korean individual) could evoke clearer neural representations related to evaluations of South Korea. Note that visual stimuli used in the Japan and South Korea conditions were not equated for lower visual features (e.g., contrast and luminance) as the main purpose of the current study is to test whether activation patterns in the amygdala can explain *individual differences* in the implicit evaluations of the outgroup (all participants viewed the same stimuli).

It should be noted that since most of the images we used depicted sports contexts, one might think that it may be possible that the relation between implicit prejudice toward South Korea and neural signals in the amygdala (or any other regions) may be explained by individual differences in a sense of rivalry rather than implicit prejudice. To refute this possibility, we conducted an additional behavioral experiment (n = 49) and confirmed that there is no correlation between implicit prejudice and a sense of rivalry toward South Korea so that any relation found between implicit prejudice and neural signals cannot be attributed to a sense of rivalry felt while viewing South Korea-related images (see the "Additional behavioral results: between implicit prejudice and a sense of rivalry toward South Korea" section below).

Procedure and tasks: During the fMRI scanning, participants viewed 30 blocks comprising 1) Japan blocks, 2) South Korea blocks, and 3) rest (i.e., a fixation cross) blocks (10 blocks each). Each of the Japan and South Korea blocks started with a cue (either "Japan" or "South Korea"; presented for 1 sec) indicating which pictures they were going to see in that trial. After the cue, four different pictures were presented in each block (each picture was presented for 2 sec followed by inter-trial-interval of 0.75 sec; see Figure 1). Each block lasted 12 sec. In each trial, an image was presented slightly to the left or right side of the screen, and participants were asked to press one of two keys to indicate whether an image appeared to the left or right side of the screen. Importantly, to prevent participants from suppressing emotions which are automatically evoked by viewing each picture, before the scanning, they were told that the study was interested in investigating how we perceive various social images (they were never explicitly asked to think about how they feel about each group during the scanning). Participants completed one 6-min fMRI run (12 sec × 30 blocks). All participants also completed another fMRI run for an independent project (where they were presented with food images [the data will not be reported here]). Since we are interested in how individual differences in ethnic prejudice are related to brain activations, the order of blocks (and the order of trials within each block) was fixed across all participants.

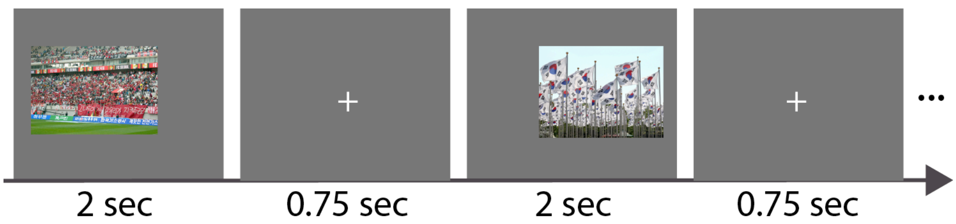

Figure 1. Experimental task. Sample stimuli in the South Korea condition. Within each South Korea block, four South Korea-related images were presented. Participants were asked to indicate whether an image appeared to the left or right side of the screen. Note that due to copyright restrictions, the two South Korea-related images shown here are similar, but not the same as the ones used in the actual experiment.

After fMRI scanning, participants were asked to perform a total of four behavioral tasks; 1) Trust Game, 2) Single-Category Implicit Association Test (SC-IAT) (Karpinski and Steinman, 2006) which measures implicit attitude toward South Korea, 3) SC-IAT which measures implicit attitude toward Japan, and 4) Explicit measure of attitude (semantic differential) toward Japan and South Korea. Since we are interested in whether the amygdala is related to attitudes toward South Korea (not *relative* attitudes toward South Korea compared to Japan), we used the SC-IAT (Karpinski and Steinman, 2006) instead of the conventional IAT (Greenwald et al., 1998). All participants performed the four tasks in this fixed order.

First, in order to measure their trust behavior toward Japan and South Korea in an incentive-compatible manner, we asked each participant to perform a series of single-shot modified Trust Games with 28 Japanese and 28 South Korean male students (Stanley et al., 2011). All participants played a role of a trustor, and in each trial, they were presented with a picture of an East Asian male and a South Korea or Japanese national flag to indicate the nationality of each partner. In reality, all pictures were Japanese individuals (the post-experimental interview confirmed that no participant had any doubt on the nationality of each individual presented during the Trust Game). The two face sets for Japan vs. South Korea were matched on trustworthiness based on an independent pilot study [n = 23, 8 females]). In each trial, participants were asked to decide how much of 500 Japanese yen (~$5) to offer to the partner (trustee; between 0 to 500 yen in increments of 100 yen). Before the Trust Games, participants were instructed that each partner would receive quadruple the amount they offered. They were further told that we had conducted a behavioral experiment in several South Korean and Japanese universities, and each of the 56 individuals had participated the study and already made the decision to return half or keep all of the money they received (in reality, such an experiment was not conducted). During the Trust Games, participants were not informed of each partner's decision, and they were told that one trial would be selected randomly at the end of the experiment, and the outcome of the randomly-selected trial would be implemented. In reality, partner's decision in the randomly selected trial was determined randomly.

Following the Trust Games, each participant was asked to perform an SC-IAT (Karpinski and Steinman, 2006) which measure implicit attitude toward South Korea. The IAT included eight positive (e.g., *Joy, Love, Wonderful*) and eight negative words (e.g., *Agony, Terrible, Nasty*; all words were translated into Japanese). The South Korean category included typical Korean names (e.g., *Han, Kim, Myong*). In addition, each participant performed another SC-IAT that measures implicit attitude toward Japan where typical Japanese names (e.g., *Shima, Nakata, Ono*) were presented instead of the Korean names. All Korean and Japanese names were matched on word length. Each of the two SC-IATs consisted of four blocks; Blocks 1 & 3 (24 trials each) were practice blocks, while Blocks 2 & 4 (72 trials each) were test blocks. During the South Korea SC-IAT, in Blocks 1 & 2, two category labels of Good and

South Korea, were presented in the top left corner, while Bad was presented in the top right corner. Participants were instructed to categorize each target word presented in the center of the screen as soon as they could by pressing either "E" or "I" key on a keyboard. An incorrect response was followed by a red "X" presentation in the center of the screen, and it remained on the screen until a participant pressed the other (correct) key. No feedback was presented after a correct response. Reaction times (RTs) for correct responses were used for data analysis. The inter-trial interval was 300 ms. In Blocks 3 & 4, a category label Good was presented in the top left corner, while Bad and South Korea were presented in the top right corner. Just like Karpinski and Steinman (2006), during Blocks 1 & 2, South Korea words, good words, and bad words were presented in a 7:7:10 ratio so that 42% of correct responses were on the "I" key and 58% of correct responses were on the "E" key. Similarly, during Blocks 3 & 4, South Korea words, good words, and bad words were presented in a 7:10:7 ratio so that 58% of correct responses were on the "I" key and 42% of correct responses were on the "E" key. The Japan SC-IAT was the same as South Korea SC-IAT except that the category South Korea was replaced with Japan, and Japanese names were presented instead of the Korean names as a target word.

After completing the two SC-IATs, each participant was asked to complete a semantic differential scale (Greenwald et al., 1998; Karpinski and Steinman, 2006), which measures explicit attitudes toward Japan and South Korea. Participants rated each of South Korea and Japan on six bipolar dimensions using a 7-point scale; *ugly-beautiful, bad-good, unpleasant-pleasant, honest-dishonest, awful-nice and unfavorable-favorable.* Finally, after completing a demographic questionnaire, they were debriefed, thanked and paid 1,500-2,500 yen for their participation.

Behavioral Data Analysis: For each of the two SC-IATs, a score for each participant was calculated using the D-score algorithm developed by Greenwald et al. (Greenwald et al., 2003). First, after excluding trials whose RT was greater than 10,000 ms, we computed the mean RT for each of the four blocks. Second, we computed a pooled standard deviation (SD) for Blocks 1 & 2 and another pooled SD for Blocks 3 & 4. Third, we computed two differences; 1) mean RT in Block 1 - mean RT in Block 3, and 2) mean RT in Block 2 - mean RT in Block 4. Fourth, each difference was divided by its associated pooled SD from step 2 above. Finally, we computed the mean of the two quotients from Step 4, which is a SC-IAT D score for an individual.

Semantic differential scores for each participant were computed by averaging the six bipolar scales separately for Japan and South Korea (Japan Cronbach's $\alpha = 0.80$, South Korea Cronbach's $\alpha = 0.92$). For both IAT and semantic differential scores, positive numbers indicate more positive evaluation of a target group. To compute correlations among all behavioral variables, in addition to SC-IAT and semantic differentials scores for each of the Japan and South Korea conditions, we computed the disparity in these scores between Japan and South Korea by subtracting scores for the South Korea condition from those for the Japan condition. These disparity indices represent *relative* implicit or explicit attitudes toward Japan relative to South Korea. Similarly, we computed the difference between the average amounts of money transferred to Japan vs. South Korea partners during the Trust Games. Note that the average amount of money transferred in the Japan and South Korea conditions were highly correlated with each other ($r(68) = 0.91$, $p < 0.001$), indicating that decisions made during the Trust Game depended more strongly on the individual differences in personality traits such as general trust or risk-taking than ethnic attitudes. Accordingly, the average amount offered in each of the Japan and South Korea conditions were not included in the correlational analyses (note that the average amount offered in each of the Japan and South Korea conditions were not correlated with SC-IAT or semantic differential scores [$-0.20 < rs < 0.04$, all $ps > 0.10$]). There was one outlier in the Trust Game data (more than 3 SD from the mean), and this participant was excluded when analyzing the Trust Game data. Finally, for paired t tests, reported effect sizes are based on Dunlap et al. (Dunlap et al., 1996)

fMRI data acquisition: All fMRI data were acquired using a Siemens 3.0 Tesla Verio scanner with a 32 channel phased array headcoil. For functional imaging, interleaved T2*- weighted gradient-echo echo-planar imaging (EPI) sequences were used to produce 40 contiguous 3.0-mm-thick trans-axial slices covering nearly the entire cerebrum (repetition time [TR] = 2,500 ms; echo time [TE] = 25 ms; flip angle [FA] = 90°; field of view [FOV] = 192 mm; 64 × 64 matrix; voxel dimensions = 3.0 × 3.0 × 3.0 mm). A high-resolution anatomical T1-weighted image (1 mm isotropic resolution) was also acquired for each participant.

fMRI Data Pre-processing: The fMRI data were analyzed using SPM8 (Wellcome Department of Imaging Neuroscience) implemented in MATLAB (MathWorks). Before data processing and statistical analysis, we discarded the first four volumes to allow for T1 equilibration. After correcting for differences in slice timing within each functional image volume, images were realigned using the mean image as a reference. (note that no participant showed excessive head motion [i.e., 3mm] during the scanning). Following realignment, the volumes were normalized to MNI space using a transformation matrix obtained from the normalization of the first EPI image of each individual participant to the EPI template using an affine transformation (resliced to a voxel size of 3.0 × 3.0 × 3.0 mm). The normalized fMRI data were spatially smoothed with an isotropic Gaussian kernel of 8 mm (full-width at half-maximum). We used the smoothed fMRI data for both univariate as well as MVPA analyses following Op de Beeck et al. (Op de Beeck, 2010) who showed that smoothing can improve decoding performance when large-scale activation patterns are assumed (e.g., (Chang et al., 2015). Note that since the voxel size of the amygdala ROIs are small (left amygdala = 54 voxels, right amygdala = 63 voxels), we also reported decoding results from unsmoothed data only for the amygdala ROIs.

fMRI Data Analysis (Univariate Analysis): We first ran a conventional general linear model (GLM) analysis. In the GLM, each of the Japan and South Korea blocks was separately modeled (duration = 12 sec). Six head motion parameters were also included in the model as nuisance regressors. Three contrast images were created for each participant; 1) a contrast image for the South Korea blocks (vs. implicit rest), 2) a contrast image for the Japan blocks (vs. implicit rest) and 3) a contrast for the South Korea vs. Japan blocks. These contrast images were submitted to the second level analysis. The same three contrast images were also used in subsequent MVPA analyses (see below).

In the second level analysis, for the South Korea contrast, South Korea SC-IAT scores and South Korea semantic differential scores were entered as covariates to test whether implicit or explicit attitudes were linearly related to activations in the brain. Similarly, for the Japan contrast, Japan SC-IAT scores and Japan semantic differential scores were entered as covariates. Finally, to test whether relative attitudes between Japan and South Korea are linearly related to brain activations, the South Korea vs.

Japan contrasts were submitted into the second level analysis, and disparity in SC-IATs scores (Japan SC-IAT minus South Korea SC-IAT), disparity in semantic differential scores (Japan semantic differential minus South Korea semantic differential) and disparity in Trust Game (the average amount offered to Japanese partners minus South Korea partners) were entered as covariates. For the univariate analysis, the statistical threshold was set at $p < 0.001$ voxelwise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

In addition to the whole-brain analysis, we also conducted the ROI analysis. Previous neuroimaging studies have identified a network of brain regions implicated in racial prejudice (the so-called "prejudice network"), which includes the anterior insula, striatum, ventral medial prefrontal cortex, orbitofrontal cortex as well as the amygdala (Amodio, 2014). Among these brain regions, we focused especially on the amygdala as previous studies have reported the link between amygdala activities and race IAT scores (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000). The same left and right amygdala masks were used as the MVPA (see below). We extracted beta values from each of the left and right amygdala ROIs and tested whether the activities were correlated with each of the behavioral indices of prejudice.

fMRI Data Analysis (MVPA): In order to decode implicit and explicit attitudes toward South Korea (and Japan) from neural signals, we employed support vector regression (SVR) (Drucker et al. (1997), as implemented in LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/), with a linear kernel and the default regularization parameter of c = 1 (default; Note that all other parameters were also set to their default values). We previously used the SVR and successfully decoded individuals' attitudes toward familiar celebrities (Izuma et al., 2017) and implicit self-esteem (implicit attitude toward the self) (Izuma et al., 2018).

As stated above, since the amygdala plays a major role in fear learning (e.g., a learned association between an outgroup and negativity) and is one of the brain regions that have been most frequently reported in previous neuroimaging research on racial prejudice (Amodio, 2014; Chekroud et al., 2014; Kubota et al., 2012), we primarily focused on the amygdala in the present study. In addition, we also ran exploratory MVPA using signals from each of 79 anatomical regions (see below). Each ROI was defined by using a WFU pickatlas toolbox for SPM (Maldjian et al., 2003).

To test whether neural signals in each of right and left amygdala ROIs can predict implicit and explicit evaluations of South Korea, we used SC-IAT scores and semantic differential scores for South Korea to decode implicit and explicit attitudes, respectively. Contrast images for the South Korea blocks were used to decode implicit and explicit attitudes toward South Korea. Although our primary interest is Japanese individuals' attitudes toward South Korea, we repeated the same analyses for the Japan condition to test whether neural signals in the prejudice network also represent implicit and explicit attitudes toward Japan. Furthermore, we investigated whether we can decode *relative* attitudes between Japan and South Korea based on relative neural signals (i.e., the South Korea vs. Japan contrast images). For this purpose, we used SC-IAT disparity scores, semantic differential disparity scores and Trust Game disparity scores as labels.

For each MVPA analysis, we computed decoding performance using the 10-fold balanced cross-validation procedure (Cohen et al., 2010; Izuma et al., 2018); we first divided participants into 10 groups (7 individuals in each group) in a way so that when decoding implicit attitude toward South Korea, these 10 groups had roughly the same means and variances of South Korea SC-IAT scores (or semantic differential scores toward South Korea when decoding explicit evaluations). In each cross-validation, one group was left out, and the SVR was performed using the data from participants in all other groups and then tested on the participants in the left-out group. This procedure was repeated for each group (a total of 10 times), and a Pearson's correlation coefficient between actual SC-IAT scores (or semantic differential scores) and predicted scores was computed.

Furthermore, to explore whether brain regions outside of the prejudice network can predict implicit, explicit evaluations, or disparity scores, we repeated the above-mentioned analyses using neural signals from each of a total of 73 regions outside of the prejudice network.

Prediction performance in each ROI was evaluated using a permutation test. We created 5,000 randomly shuffled permutations of IAT or semantic differential scores and ran the SVR using the permutated data in each ROI to obtain a distribution of correlations between predicted and actual scores under the null hypothesis. Note that behavioral scores were shuffled within each of the 10 fold groups so that the averages scores in the 10 fold groups were maintained across the permutations. For the regions outside of the amygdala, false discovery rate (FDR) (Benjamini and Hochberg (1995) correction for multiple comparisons was applied ($q < 0.05$).

Results

Behavioral results: We confirmed that Japanese participants had negative implicit and explicit attitudes toward South Korea, and their negative attitudes were also reflected in trust behavior as measured by the Trust Game. First, not surprisingly, participants had more positive explicit attitudes toward Japan (mean = 5.34, SD =

0.99) compared to South Korea (mean = 3.63, SD = 1.12). Semantic deferential scores were significantly more positive for Japan compared to South Korea ($t(69) = 11.14$, $p < 0.001$, $d = 1.62$, paired t-test; Figure 2a). While semantic differential scores for the Japan were significantly higher than the midpoint (i.e., 4) ($t(69) = 11.36$, $p < 0.001$, Cohen's $d = 1.36$, one-sample t-test), scores for South Korea were significantly lower than the midpoint ($t(69) = 2.79$, $p = 0.007$, Cohen's $d = 0.33$, one-sample t-test), indicating that Japanese participants possess positive and negative explicit attitudes toward Japan and South Korea, respectively.

Similar results were obtained with the implicit attitude measure. Japan SC-IAT scores were significantly higher than South Korea SC-IAT scores (Japan mean = 0.05, SD = 0.32, South Korea mean = -0.30, SD = 0.31; $t(69) = 7.72$, $p < 0.001$, $d = 1.12$, paired t-test; Figure 2b), indicating more positive implicit attitude toward Japan compared to South Korea. While Japan SC-IAT scores were not significantly different from zero ($t(69) = 1.28$, $p = 0.21$, Cohen's $d = 0.15$, one-sample t-test), South Korea SC-IAT scores were significantly negative ($t(69) = 8.13$, $p < 0.001$, Cohen's $d = 0.97$, one-sample t-test), suggesting the presence of negative implicit attitude (i.e., implicit prejudice) toward South Korea and, on average, neutral implicit attitude toward Japan. It should be noted that since all participants completed the two SC-IATs in the same order (South Korea SC-IAT first), the result might be confounded with the effect of task order (e.g., practice effect) (however, see Study 4 of Karpinski and Steinman (2006) for data showing that the order of two SC-IATs has, if anything, a very limited effect).

Participants' differential implicit and explicit evaluations of Japan vs. South Korea were also reflected in trust behavior. The data from the Trust Games revealed that participants transferred significantly more money to Japanese partners compared to South Korean partners (Japan mean = 194.4 yen, SD = 114.7, South Korea mean = 172.2 yen, SD = 110.0; $t(68) = 5.54$, $p < 0.001$, $d = 0.20$, paired t-test; Figure 2c).

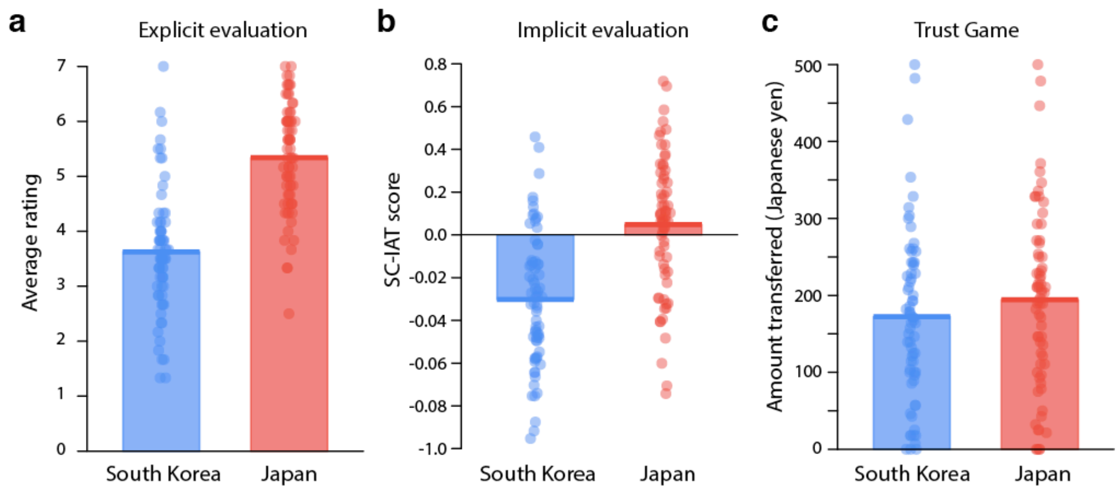

Figure 2. Behavioral results. **a**). Explicit evaluation (average semantic differential scores) of South Korea and Japan. Higher numbers indicate more positive explicit evaluation of each ethnic group. **b**). Implicit evaluation (average SC-IAT scores) of South Korea and Japan. Higher numbers indicate more positive implicit evaluation of each group. **c**). The average amount of money transferred to South Korean and Japanese partners in the Trust Games.

Implicit and explicit evaluations were uncorrelated with each other, which is largely consistent with previous research (Hofmann et al., 2005). All correlations between implicit (SC-IATs) and explicit evaluations (semantic differential) were non-significant ($-0.06 < rs < 0.14$, $ps > 0.23$). Unlike Stanley et al. (2011), our data showed that differential behaviors during the Trust Games with Japan vs. South Korea partners were not correlated with implicit evaluations (disparity in SC-IAT scores) (but see (Oswald et al., 2013) for a meta-analysis demonstrating the low predictive validity of the IAT). Trust Game disparity scores were also not correlated with explicit evaluations (disparity in semantic differential scores). Two SC-IATs were significantly correlated with each other ($r(68) = 0.27$, $p = 0.026$), suggesting that those who possess more positive implicit attitudes toward Japan tend to have more positive implicit attitudes toward South Korea as well. Similarly, two semantic differential scores were significantly correlated with each other ($r(68) = 0.26$, $p = 0.030$). All correlations across the behavioral variables are shown in Table 1.

Inside the fMRI scanner, each participant performed a simple button press task (i.e., indicate whether an image appeared to the left or right side of the screen), and their performance was nearly perfect (Japan block = 98.4%, South Korea block = 97.9%). There was no significant difference in performance between the Japan vs. South Korea blocks ($t(69) = 1.22$, $p = 0.23$, $d = 0.11$, paired t-test). Participants' reaction times (RTs) were significantly slower in the South Korea blocks (mean RT = 670 ms) compared to the Japan blocks (mean RT = 661 ms; $t(69) = 2.06$, $p = 0.043$, $d = 0.06$, paired t-test). Importantly, performance and RTs during the button press task inside the fMRI scanner were not correlated with both implicit (SC-IAT) and explicit evaluations (semantic differential) (Japan $-0.09 < rs(68) < 0.16$, $ps > 0.18$; South Korea $-0.08 < rs(68) < 0.15$, $ps > 0.22$).

fMRI Results (MVPA): We first investigated whether neural signals in the amygdala can predict implicit (and explicit) attitudes toward South Korea. The results revealed that neural signals in the left amygdala significantly predicted the level of implicit evaluations of South Korea ($r(68) = 0.31$, $p_{perm} = 0.021$; Figures 3). Since one data point in the predicted scores was identified as an outlier based on the Grubbs' test (p = 0.028), we also computed Spearman's rank correlation, and the result remains significant ($r_s(52) = 0.28$, $p_{perm} = 0.035$). In contrast, the prediction based on the right amygdala activations was not significant ($p_{perm} = 0.55$; see Table 2). Thus, the result indicates that passive viewing of South Korea-related images automatically evoked similar neural responses in the left (but not right) amygdala across different individuals depending on their level of implicit evaluations. Given the small voxel size of the amygdala ROIs, we also tried the same MVPA analysis on unsmoothed data and found that prediction performance was improved in the left amygdala ($r(68) = 0.40$, $p_{perm} = 0.005$), but that in the right amygdala remain non-significant ($p_{perm} = 0.98$).

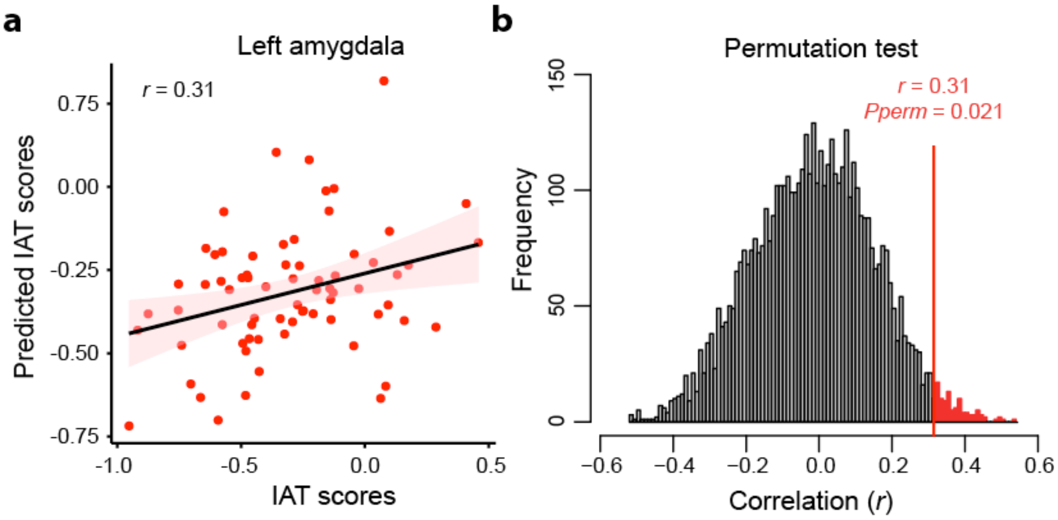

Figure 3. MVPA results. a). A scatter plot showing a correlation between participants' SC-IAT scores for South Korea (i.e., implicit evaluations) and predicted scores based on neural activations in the left amygdala ROI ($r(68) = 0.31$). The predicted scores were cross-validated by using a 10-fold cross-validation procedure. **b**). A histogram showing the distributions of correlation coefficients between actual and predicted SC-IAT scores with randomly permutated data (5,000 times). The correlation with actual data was significant at $p_{perm} = 0.021$.

Since the possibility that different sub-regions of the amygdala may play different roles in prejudice has been suggested previously (Chekroud et al., 2014), we looked at a distribution of weight values in the left amygdala ROI. It revealed that weight values of voxels in the medial part of the amygdala tend to be positive, while weight values of voxels in the lateral part of the amygdala tend to be negative (see Figure 4a). This is confirmed by a significant Spearman's rank correlation between weight values (converted to 1 = positive weight value or -1 = negative weight value) and x-coordinates ($r_s(52) = 0.31$, $p = 0.025$). On the other hand, weight values were not correlated with both y- and z-coordinates (both *ps* > 0.49).

Although the above analysis suggested that the medial and lateral parts of the left amygdala play different roles in implicit evaluations, the classifier weights need to be interpreted with caution as they do not simply reflect the amplitude of a signal in each voxel (see (Haufe et al., 2014; Haynes, 2015)). Thus, to further investigate how different parts within the left amygdala are related to the individual differences in implicit evaluations, we computed a Pearson correlation between South Korea SC-IAT scores and the amplitude of the signals in the South Korea blocks (i.e., univariate activations) in each of 54 voxels within the left amygdala ROI. 54 correlation coefficients ranged from $r = -0.08$ to $r = 0.30$ (average $r = 0.09$, standard deviation = 0.10). Since all of x-, y-, and z-coordinate values were not normally distributed, we computed a Spearman's rank correlation between the correlation coefficients and each of the x-, y-, and z-coordinates. Consistent with the weight map results reported above, the results revealed that there was a significant positive correlation with x-coordinates ($r_s(52) = 0.77$, $p < 0.001$), indicating that the across-subject correlations between the amplitudes of the South Korea univariate contrast and South Korea SC-IAT scores tended to be more positive in more medial parts (i.e., central nuclei) of the left amygdala (see Figure 4b). Note that since x-, y-, z-coordinates are not completely independent (i.e., the left amygdala ROI is not a complete sphere), we also found a significant positive correlation with y-coordinates ($r_s(52) = 0.47$, $p = 0.003$), and z-coordinates also showed a significant trend ($r_s(52) = 0.23$, $p = 0.09$) (see Figure 4).

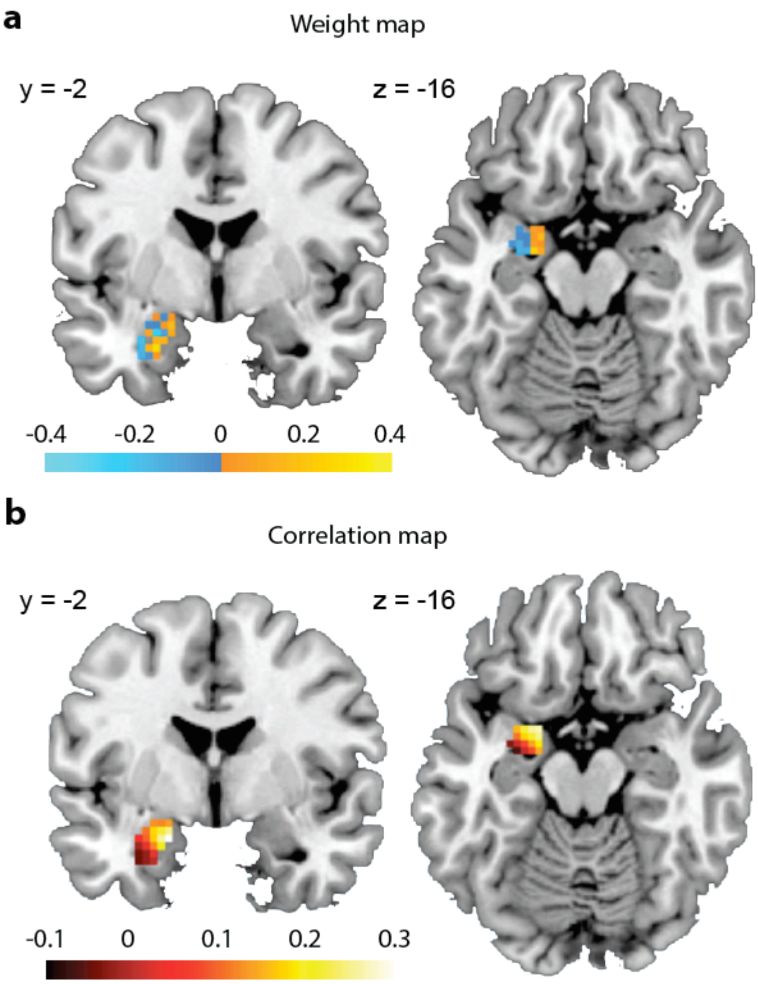

Figure 4. Weight and correlation maps within the left amygdala ROI. a). Weight values of the implicit evaluation decoder. **b**). Across-subject correlational values between univariate activations (i.e., the South Korea contrast) and implicit-evaluations.

We repeated the same MVPA analysis with South Korea SC-IAT scores for each region outside of the amygdala ROIs. Out of a total of 79 regions, only 4 regions (right anterior insula, left middle frontal gyrus, right fusiform gyrus, and right lingual gyrus) could predict implicit evaluations of South Korea at $p_{perm} < 0.05$ (uncorrected), but none of them survived the FDR correction (see Table 3). Although it did not survive the corrected threshold, significant prediction by neural signals in the right anterior insula ($r(68) = 0.28$, $p_{perm} = 0.038$) may be notable as the anterior insula is one of the regions previously implicated in racial prejudice (known as the prejudice network) (Amodio, 2014).

Interestingly, in contrast to implicit evaluations of South Korea, both left and right amygdala could not predict explicit evaluations (i.e., semantic differential scores) of South Korea. Prediction of explicit evaluations was not significant even when some of the 8 ROIs in the prejudice network were combined (the highest decoding performance among a total of 247 combinations was $r(68) = 0.24$, $p_{perm} = 0.071$ [based on neural signals from the left and right caudate nucleus ROIs]).

Neural signals in each of the left and right amygdala ROIs were also unrelated to both implicit and explicit evaluations of Japan (Table 2). Neural signals in regions outside of the amygdala were also largely unrelated to explicit evaluations of South Korea as well as both implicit and explicit evaluations of Japan (see Table 3). Furthermore, neural signals both inside and outside of the amygdala ROIs were unrelated to *relative* implicit and explicit evaluations (i.e., SC-IAT disparity scores and semantic differential disparity scores) and Trust Game disparity scores.

fMRI Results (Univariate Analyses): The contrast between the South Korea vs. Japan blocks revealed a significant activation only in the visual cortex (right: x = 12, y = -97, z = -2, left: x = -18, y = -106, z = -5; 1,849 voxels). Even when a statistical threshold was lowered to $p < 0.01$, we didn't find any activation in any regions in the prejudice network. We tested whether activity in each of the left and right amygdala ROIs was significantly correlated with SC-IAT disparity scores (i.e., relative implicit prejudice), as reported previously (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000). The results revealed that amygdala activities in both hemispheres were not correlated with relative implicit prejudice, and if anything, the correlations were in the opposite direction (the more positive their attitudes were toward South Korea relative to Japan, the higher the activity in the amygdala in response to South

Korea-related images relative to Japan-related images; left amygdala $r(68) = 0.19$, $p = 0.11$, and right amygdala $r(68) = 0.09$, $p = 0.44$). We further conducted the whole-brain analysis and investigated whether SC-IAT disparity scores were significantly correlated with activations in regions other than the amygdala. However, SC-IAT disparity scores were not significantly related to activations in any of the brain regions (neither positively nor negatively). Even when the threshold was lowered to $p < 0.01$, we did not find any significant activation in the amygdala ROIs or any of the regions in the prejudice network.

Similarly, activities in each of the left and right amygdala ROIs were not correlated with disparity in semantic differential between Japan vs. South Korea (left amygdala $r(68) = -0.01$, $p = 0.95$, and right amygdala $r(68) = 0.07$, $p = 0.59$). The whole brain analysis showed that the explicit disparity scores were not correlated with activation in any region except for right intraparietal sulcus (IPS; x = 45, y = -28, z = 61; 199 voxels). Those who have more positive explicit evaluations of South Korea relative to Japan showed higher IPS activations in response to South Korea images compared to Japan images.

We further tested whether activations during each of the Japan and South Korea blocks (vs. implicit rest) were correlated with corresponding SC-IAT scores. South Korea SC-IAT scores were not significantly related to activations in any of the brain regions in response to South Korea-related images (neither positively nor negatively) (see Figure 4b for the same analysis only within the left amygdala ROI). Similarly, Japan SC-IAT scores were not significantly related to activations in any of the brain regions in response to Japan-related images. The same analyses were repeated using semantic differential scores instead of SC-IAT scores, but it revealed no significant correlation between explicit attitudes and brain activations (for both Japan and South

Korea). Thus, our univariate analyses failed to show any reliable association between implicit (and explicit) evaluations and neural activities. We also didn't find any significant association between Trust Game behaviors and neural activities.

Additional behavioral results: between implicit prejudice and a sense of rivalry toward South Korea

As stated above, we selected South Korea- and Japan-related images in a way so that no image depicted a direct competition between Japan and South Korea. Nonetheless, since most of the images we used depicted sports contexts, it may be possible that the relation between implicit prejudice toward South Korea and neural signals in the amygdala (or any other regions) may be explained by a sense of rivalry rather than implicit prejudice toward South Korea. To refute this possibility, we conducted an additional behavioral experiment and tested whether implicit prejudice as measured by the Single-Category Implicit Association Test (SC-IAT) is related to a sense of rivalry toward South Korea.

We recruited an additional independent sample of 49 university students (23 females, 18-24 years old, mean age = 19.8 years, SD = 1.36) for a behavioral experiment (without neuroimaging). Data from one additional participant was excluded from the analyses because his data was not correctly saved due to a malfunction of the task presentation program. Like the main fMRI experiment, all participants were recruited from a subject pool of the Kochi University of Technology.

Participants in this behavioral experiment completed the following three tasks; 1) SC-IAT which measures implicit attitude toward South Korea, 2) Explicit measure of attitude (semantic differential) toward South Korea, and 3) the Sport Rivalry Fan Perception Scale (SRFPS) (Havard et al. 2013). The SRFPS consists of four subscales measuring Outgroup Competition against others (Indirect) (OIC; e.g., "I want my favorite team's rival to win all games except when they play my favorite team"), Outgroup Academic Prestige (OAP; e.g., "The academic prestige of my favorite team's rival is poor"), Outgroup Sportsmanship (OS; e.g., "Fans of my favorite team's rival demonstrate poor sportsmanship at games"), and Sense of Satisfaction when the favorite team defeats the rival team in direct competition (SoS; e.g., "I feel a sense of belonging when my favorite team beats my favorite team's rival') (3 items for each of the four subscales). In each item, we replaced "my favorite team" with Japan or the Japanese national team and "my favorite team’s rival" with South Korea or the South Korean national team. Furthermore, we removed the OAP subscale (all 3 items) and the following 2 items from the OIC subscale ("I would support my favorite team’s rival in a championship game," and "I would support my favorite team’s rival in out-of-conference play") because they are specific to an American college sports context. Accordingly, participants answered a total of 7 items, and three following scores were computed for each participants; 1) OIC score (1 item), 2) OS score (average of 3 items; Cronbach's $\alpha = 0.81$), and 3) SoS score (average of 3 items; Cronbach's $\alpha = 0.64$).

Although we found a significant positive correlation between the South Korea SC-IAT scores and the OIC score at a $p < 0.05$ level ($r(47) = 0.30$, $p = 0.017$; one-tailed, no correction for multiple comparison), the same OIC scores were more strongly related to explicit attitudes toward South Korea ($r(47) = 0.52$, $p < 0.001$). The other two subscales were not related to the SC-IAT scores (*rs* < 0.03), while the SoS subscale was related to explicit attitudes toward South Korea ($r(47) = 0.37$, $p = 0.005$; but not the OS subscale $r(47) = 0.01$). Thus, our results showed that in general, the more positive explicit attitude a Japanese individual has toward South Korea, the higher the sense of rivalry they have toward South Korea. These results indicate that it is highly unlikely that the link between implicit prejudice toward South Korea (i.e., the SC-IAT scores) and neural signals in the left amygdala we found (Figure 3) can be explained by the sense of rivalry toward South Korea.

Although our additional data indicate that a sense of rivalry is an unlikely explanation, one may still argue that there was an important difference in experimental procedures between the original fMRI study and this behavioral study. While participants in this behavioral study (n = 49) completed only behavioral measures (e.g., the SRFPS scale, SC-IAT, and semantic differential scales), participants in the original fMRI study (n= 70) had been exposed to the pictures depicting South Korea (some of which depicted sports related scenes) inside an fMRI scanner before they completed the behavioral measures. Thus, it might be possible that feelings of competitiveness evoked by viewing these pictures might have affected both their SC-IAT scores as well as the amygdala activation. However, it should be stressed that given the stronger link between a sense of rivalry and explicit evaluations found in the behavioral study, if a sense of rivalry evoked by the pictures were a major factor, we should have found a stronger correlation between the amygdala activations and *explicit* evaluations.

Discussion

The present study showed that using MVPA, neural activation patterns in the left amygdala could predict Japanese participants' level of implicit negative evaluations of South Korea. With the much larger sample size (n = 70) than the previous studies (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000), the present study provides reliable evidence that the left amygdala plays a key role in representing negative implicit evaluations of an ethnic outgroup. Our results also suggest that despite that ethnic prejudice is more variable across different cultures compared to stereotypes of sex and age (Fiske, 2017), the link between the amygdala and implicit negative evaluations of an ethnic or racial outgroup might be generalizable across different intergroup contexts beyond the intergroup relation between White vs. Black Americans. Even more generally, given the well-known role of the amygdala in fear learning (learning of the associations between initially neutral stimuli and aversive events) (Fendt and Fanselow, 1999; Pape and Pare, 2010), it seems likely that the amygdala plays a key role in implicit attitudes toward not only social groups but also non-social objects, although this idea should be formally tested in future research. Implicit measures of attitudes such as IAT are thought to measure the strength of automatically activated evaluative associations (e.g., the association between an outgroup and negatively-valence words) stored in memory (i.e., the brain) (Greenwald et al., 1998). Thus, the present results suggest that the associations between South Korea and negativity possessed by Japanese individuals are stored in the left amygdala. In contrast, explicit evaluations of South Korea were not robustly related to neural signals. Thus, neural activation automatically evoked by the passive viewing of South Korea-related images predicted implicit (automatic) evaluations of South Korea, but not explicit (controlled) evaluations.

The result further showed that activations in the medial part of the amygdala contribute positively to the prediction of implicit evaluations, whereas those in the lateral part contribute negatively to the prediction (Figure 4a). Similarly, the mass-univariate correlational analyses revealed that univariate signals were more positively related to implicit evaluations in the more medial and anterior part of the left amygdala (Figure 4b). This medial-lateral distinction is largely consistent with the anatomical organization of the amygdala (the medial part = centromedial nuclei of the amygdala, the lateral part = basolateral nuclei of the amygdala) (Sah et al., 2003). The results may suggest that medial vs. lateral regions of the amygdala play different roles in implicit evaluations, and this is consistent with past research showing the functional distinction between basolateral nuclei and central nuclei of the amygdala (Balleine and Killcross, 2006). As the lateral nuclei of the amygdala play a key role in the formation of memories during fear conditioning (Rodrigues et al., 2004), our results may suggest that neural signals in the centromedial nuclei of the amygdala reflecting other factors such as negative affect or motivational salience of the stimuli (some of which are reflected in the IAT scores) might contribute to the prediction of implicit evaluation.

While neural signals in the left amygdala significantly predicted implicit evaluations of South Korea, the right amygdala was not associated with implicit evaluations. The findings from past neuroimaging studies on racial prejudice were inconsistent as to the lateralization of the amygdala responses (see Chekroud et al. (2014) for review). While some studies reported activations in bilateral amygdala in response to black faces vs. white faces (e.g., (Hart et al., 2000); Phelps et al. (2000)), other studies observed activations only in left (e.g., (Wheeler and Fiske, 2005)) or right amygdala (e.g., (Cunningham et al., 2004; McCutcheon et al., 2018)). Furthermore, Phelps et al. (2000) found that activations in the amygdala in both hemispheres were correlated with IAT scores, while activations only in the left amygdala were correlated with the startle eyeblink potentiation bias (a physiological measure of indirect racial bias). In contrast, both Cunningham et al. (2004) and Brosch et al. (2013) found the correlation between amygdala activities and IAT scores only in the right amygdala. Past meta-analyses found that left amygdala activation is more consistently observed than right amygdala activation during the processing of affective stimuli (Baas et al., 2004) and that the left amygdala is more likely to be activated when stimuli contain language whereas the right amygdala is more likely to be activated when stimuli were masked to prevent conscious awareness (Costafreda et al., 2008). Although these explanations could account for some of the past findings (e.g., in Cunningham et al. (2004), faces were briefly presented [30 ms] and masked, and they found activations only in the right amygdala), it is unlikely that differences in the use of language or affective stimuli can explain the lateralization of amygdala activations found in other previous studies as well as the present study. Future studies should systematically manipulate these factors and test whether the left and right amygdala play different roles in racial and ethnic prejudice.

In contrast to evaluations of South Korea (outgroup), we found that both implicit and explicit evaluations of Japan (ingroup) were not robustly associated with neural signals in any of the brain regions. There are at least two possible interpretations for this dissociation between Japan and South Korea. One idea is that as the amygdala plays a key role in fear learning, neural signals in the left amygdala are related to negative implicit evaluations, but not neutral evaluations. Our behavioral data showed that participants' implicit evaluations of Japan were not clearly negative or positive, while their implicit evaluations of South Korea were largely negative (Figure 2b). Note that it might have been possible to decode individuals' level of implicit evaluations of Japan if their attitudes toward Japan were clearly positive, as the amygdala has been implicated in processing positive as well as negative values of stimuli (Murray, 2007). The other idea is that neural representations evoked by Japan-related images are more complex than those evoked by South Korea-images. While all South Korea-related images may be automatically evaluated in a similar manner across Japanese individuals (i.e., outgroup homogeneity effect), each Japan-related image is likely to evoke a variety of different psychological and emotional reactions in each Japanese individual, which in turn made across-subject decoding of attitudes toward Japan more difficult. Thus, it is important to further investigate the role of the amygdala and other brain regions in negative (and positive) implicit evaluations across a variety of different intergroup contexts in future research.

While past social neuroscience studies reported the involvements of the anterior cingulate cortex (ACC) and dorsolateral prefrontal cortex (DLPFC) in racial prejudice (Amodio, 2014; Kubota et al., 2012), and one study (Richeson et al., 2003) reported significant correlations between implicit evaluations and activities in these regions, our results showed that both of these regions were not related to implicit and explicit evaluations. In past studies, ACC and DLPFC activations in response to outgroup faces were often interpreted as reflecting a conflict between automatic affective responses and intention to respond fairly to outgroup faces and cognitive regulation of evoked negative affective responses to the outgroup, respectively. Thus, although speculative, our results might suggest that the level of implicit evaluations people possess is not related to how much conflict they feel and how much they try to suppress their prejudiced responses.

Our findings also showed that despite Japanese participants' clear negative implicit as well as explicit evaluations of South Korea (Figure 2), conventional univariate fMRI data analyses failed to find differences in activations in any of the prejudice network between the South Korea vs. Japan blocks. Thus, our results are in line with previous studies that did not find such amygdala activations (Brosch et al., 2013; Cassidy and Krendl, 2016; Gilbert et al., 2012; Golby et al., 2001; Li et al., 2016; Mattan et al., 2018; Phelps et al., 2000; Richeson et al., 2003; Stanley et al., 2012; Terbeck et al., 2015). However, it should be noted that the experimental design of the present study was optimized for individual difference analyses (i.e., across-subject correlation). For example, we fixed the block order for all participants (so that the order effect, if there is any, should affect all participants in a similar manner). In addition, we did not match lower visual features of the stimuli across the two conditions. These differences might explain the lack of significant activations in the South Korea vs. Japan contrast. Nonetheless, the order effect and the differences in visual features of the stimuli are unlikely to explain the lack of a significant across-subject correlation between univariate activations in the amygdala and implicit evaluations. Contrary to the three small studies (Brosch et al., 2013; Cunningham et al., 2004; Phelps et al., 2000), our univariate analysis revealed no correlation between amygdala activities and implicit evaluations of South Korea, which is in agreement with the six studies with larger sample sizes (Cassidy and Krendl, 2016; Cassidy et al., 2016; Gilbert et al., 2012; Li et al., 2016; Richeson et al., 2003; Terbeck et al., 2015).

Thus, in contrast to MVPA, conventional univariate fMRI data analyses may not be sensitive enough to detect the neural signatures of implicit evaluations, and it may be a reason for the inconsistent findings in previous research (see (Amodio, 2014; Chekroud et al., 2014; Kubota et al., 2012). Thus, the present study demonstrates the utility of MVPA, which may be able to refine the past findings and provide important insights into the role played by each region within the prejudice network in future research. For example, previous studies demonstrated, using univariate fMRI data analysis, that amygdala activity in response to a racial outgroup is modulated by perceiver's goals (Lieberman et al., 2005; Van Bavel et al., 2008; Wheeler and Fiske, 2005), and one study found no difference in the amygdala activity when participants performed a simple dot detection task while ingroup or outgroup images were presented on the screen (Wheeler and Fiske, 2005). Since MVPA is capable of detecting differences between conditions even when there is no overall difference in the average amplitude of fMRI signals (e.g., (Harrison and Tong, 2009; Kohler et al., 2013), it is interesting to test in future research whether MVPA can decode implicit and/or explicit evaluations of an outgroup regardless of perceiver's goals, which can provide an important insight into the automaticity of stereotyping and prejudice (Bargh, 1999).

While the present study provides evidence for the link between the amygdala and implicit prejudice in the intergroup context of Japan vs. South Korea, it is important to investigate the same link in other intergroup contexts including White vs. Black Americans in future research (i.e., using MVPA and with a larger sample size). Although it is conceivable to think that the amygdala plays a major role in implicit prejudice (i.e., association between an outgroup and negativity) in all intergroup contexts given its well-known role in fear learning (Fendt and Fanselow, 1999; Pape and Pare, 2010), other brain regions may play additional roles in a different intergroup context. Given that prejudice is a world-wide problem (Landis and Albert, 2012; Noor and Montiel, 2009), future research should investigate similarities and differences in neural signatures of implicit prejudice across different intergroup contexts to have a comprehensive understanding of its neural mechanisms.

It is also important to use various stimuli in future research (e.g., pictures unrelated to sports scenes). Although our additional behavioral study found only a very weak link between a sense of rivalry or competitiveness and implicit evaluations toward the outgroup, there was an important difference in experimental procedures between the original fMRI study and this behavioral study. While participants in this additional behavioral study (n = 49) completed only behavioral measures (e.g., the SRFPS scale, SC-IAT, and semantic differential scales), participants in the original fMRI study (n= 70) had been exposed to the pictures depicting South Korea (some of which depicted sports related scenes) inside an fMRI scanner before they completed the behavioral measures. Thus, it is possible that feelings of competitiveness evoked by viewing these pictures might have affected their SC-IAT scores as well as the amygdala activation. Nonetheless, it should be stressed that given the stronger link between a sense of rivalry and *explicit* evaluations, it is highly unlikely that our main amygdala findings (Figure 3) can be explained by the sense of rivalry.

Finally, an important implication of the present finding is that neural signals in the amygdala could be used as an independent neural index of implicit attitudes toward an outgroup. Past social psychological studies have reported that a variety of simple behavioral interventions could reduce implicit attitudes toward an outgroup as measured by IAT (Lai et al., 2014). However, it has been debated whether such interventions actually reduced implicit attitudes or just IAT scores (i.e., implicit attitudes remain unchanged) (Han et al., 2010). An independent neural index has a potential to provide a unique insight into this debate, and thus, the present finding is the important step toward formulating effective interventions to regulate and reduce various prejudice in societies.

Acknowledgements

We thank Kosuke Motoki and Daniel Pledger for help with behavioral data collection and Stephanie J. Wake for her comments on an earlier version of the manuscript. This work was supported by the JSPS London JBUK Japan Award (to K.I.), the JSPS KAKENHI Grant Number 17H04789 (to K.S.), JSPS KAKENHI Grant Number 15K12777 and 17H00891 (to K.N.) and University Grant for Research Facility and Project, Kochi University of Technology (to R.A. and K.N.).

References

Amodio, D.M., 2014. The neuroscience of prejudice and stereotyping. Nature Reviews Neuroscience 15, 670-682.
Baas, D., Aleman, A., Kahn, R.S., 2004. Lateralization of amygdala activation: a systematic review of functional neuroimaging studies. Brain Research Reviews 45, 96-103.
Balleine, B.W., Killcross, S., 2006. Parallel incentive processing: an integrated view of amygdala function. Trends in Neurosciences 29, 272-279.
Bargh, J.A., 1999. The cognitive monster: The case against controllability of automatic stereotype effects. In: Chaiken, S., Trope, Y. (Eds.), Dual-process theories in social psychology. Guilford Press, New York, pp. 361-382.
BBC_World_Service, 2010. Global views of United States improve while other countries decline.
BBC_World_Service, 2014. Negative views of Russia on the rise: Global poll.
Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57, 289-300.
Brosch, T., Bar-David, E., Phelps, E.A., 2013. Implicit race bias decreases the similarity of neural representations of black and white faces. Psychol Sci 24, 160-166.
Bruneau, E.G., Saxe, R., 2010. Attitudes towards the outgroup are predicted by activity in the precuneus in Arabs and Israelis. Neuroimage 52, 1704-1711.
Cassidy, B.S., Krendl, A.C., 2016. Dynamic neural mechanisms underlie race disparities in social cognition. Neuroimage 132, 238-246.
Cassidy, B.S., Lee, E.J., Krendl, A.C., 2016. Age and executive ability impact the neural correlates of race perception. Soc Cogn Affect Neurosci 11, 1752-1761.
Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., Wager, T.D., 2015. A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol 13, e1002180.
Chekroud, A.M., Everett, J.A., Bridge, H., Hewstone, M., 2014. A review of neuroimaging studies of race-related prejudice: does amygdala response reflect threat? Frontiers in Human Neuroscience 8, 179.
Cho, K., 2017. Quantitative Text Analysis of "Yahoo! News" : Focusing on Comments on Koreans The Journal of Applied Sociology 59, 113 - 127.
Cloutier, J., Li, T., Correll, J., 2014. The impact of childhood experience on amygdala response to perceptually familiar black and white faces. J Cogn Neurosci 26, 1992-2004.
Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., Poldrack, R.A., 2010. Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals. Frontiers in Human Neuroscience 4, 47.
Costafreda, S.G., Brammer, M.J., David, A.S., Fu, C.H.Y., 2008. Predictors of amygdala activation during the processing of emotional stimuli: A meta-analysis of 385 PET and fMRI studies. Brain Research Reviews 58, 57-70.
Cunningham, W.A., Johnson, M.K., Raye, C.L., Chris Gatenby, J., Gore, J.C., Banaji, M.R., 2004. Separable neural components in the processing of black and white faces. Psychol Sci 15, 806-813.
Demos, K.E., Kelley, W.M., Ryan, S.L., Davis, F.C., Whalen, P.J., 2008. Human amygdala sensitivity to the pupil size of others. Cereb Cortex 18, 2729-2734.
Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9 9,

155-161.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., Burke, M.J., 1996. Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods 1, 170-177.

Fendt, M., Fanselow, M.S., 1999. The neuroanatomical and neurochemical basis of conditioned fear. Neuroscience and Biobehavioral Reviews 23, 743-760.

Fiske, S.T., 2017. Prejudices in Cultural Contexts: Shared Stereotypes (Gender, Age) Versus Variable Stereotypes (Race, Ethnicity, Religion). Perspectives on Psychological Science 12, 791-799.

Gilbert, S.J., Swencionis, J.K., Amodio, D.M., 2012. Evaluative vs. trait representation in intergroup social judgments: distinct roles of anterior temporal lobe and prefrontal cortex. Neuropsychologia 50, 3600-3611.

Golby, A.J., Gabrieli, J.D., Chiao, J.Y., Eberhardt, J.L., 2001. Differential responses in the fusiform region to same-race and other-race faces. Nat Neurosci 4, 845-850.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology 74, 1464-1480.

Greenwald, A.G., 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, 197-216.

Han, H.A., Czellar, S., Olson, M.A., Fazio, R.H., 2010. Malleability of attitudes or malleability of the IAT? Journal of Experimental Social Psychology 46, 286-298.

Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632-635.

Hart, A.J., Whalen, P.J., Shin, L.M., McInerney, S.C., Fischer, H., Rauch, S.L., 2000. Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. Neuroreport 11, 2351-2355.

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.D., Blankertz, B., Biessgmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96-110.

Haynes, J.D., 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87, 257-270.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., Schmitt, M., 2005. A meta-analysis on the correlation between the implicit association test and explicit self-report measures. Pers Soc Psychol Bull 31, 1369-1385.

Izuma, K., Kennedy, K., Fitzjohn, A., Sedikides, C., Shibata, K., 2018. Neural activity in the reward-related brain regions predicts implicit self-esteem: A novel validity test of psychological measures using neuroimaging. J Pers Soc Psychol 114, 343-357.

Izuma, K., Shibata, K., Matsumoto, K., Adolphs, R., 2017. Neural predictors of evaluative attitudes towards celebrities. Social Cognitive and Affective Neuroscience 12, 382-390.

Jimura, K., Poldrack, R.A., 2012. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia 50, 544-552.

Karpinski, A., Steinman, R.B., 2006. The single category implicit association test as a measure of implicit social cognition. J Pers Soc Psychol 91, 16-32.

Kohler, P.J., Fogelson, S.V., Reavis, E.A., Meng, M., Guntupalli, J.S., Hanke, M., Halchenko, Y.O., Connolly, A.C., Haxby, J.V., Tse, P.U., 2013. Pattern classification precedes region-average hemodynamic response in early visual cortex. Neuroimage

78, 249-260.

Kubota, J.T., Banaji, M.R., Phelps, E.A., 2012. The neuroscience of race. Nat Neurosci 15, 940-948.

Lai, C.K., Marini, M., Lehr, S.A., Cerruti, C., Shin, J.E.L., Joy-Gaba, J.A., Ho, A.K., Teachman, B.A., Wojcik, S.P., Koleva, S.P., Frazier, R.S., Heiphetz, L., Chen, E.E., Turner, R.N., Haidt, J., Kesebir, S., Hawkins, C.B., Schaefer, H.S., Rubichi, S., Sartori, G., Dial, C.M., Sriram, N., Banaji, M.R., Nosek, B.A., 2014. Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions. Journal of Experimental Psychology-General 143, 1765-1785.

Landis, D., Albert, R.D. (Eds.), 2012. Handbook of Ethnic Conflict: International Perspectives. Springer.

Li, T., Cardenas-Iniguez, C., Correll, J., Cloutier, J., 2016. The impact of motivation on race-based impression formation. Neuroimage 124, 1-7.

Lieberman, M.D., Hariri, A., Jarcho, J.M., Eisenberger, N.I., Bookheimer, S.Y., 2005. An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. Nat Neurosci 8, 720-722.

Loken, E., Gelman, A., 2017. Measurement error and the replication crisis. Science 355, 584-585.

Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H., 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19, 1233-1239.

Mattan, B.D., Kubota, J.T., Dang, T.P., Cloutier, J., 2018. External motivation to avoid prejudice alters neural responses to targets varying in race and status. Soc Cogn Affect Neurosci 13, 22-31.

McCutcheon, R., Bloomfield, M.A.P., Dahoun, T., Quinlan, M., Terbeck, S., Mehta, M., Howes, O., 2018. Amygdala reactivity in ethnic minorities and its relationship to the social environment: an fMRI study. Psychol Med, 1-8.

Murray, E.A., 2007. The amygdala, reward and emotion. Trends Cogn Sci 11, 489-497.

Noor, N.M., Montiel, C.J. (Eds.), 2009. Peace Psychology in Asia. Springer.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10, 424-430.

Nosek, B.A., Greenwald, A.G., Banaji, M., 2007. The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In: Bargh, J.A. (Ed.), Automatic processes in social thinking and behavior. Psychology Press.

Op de Beeck, H.P., 2010. Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage 49, 1943-1948.

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. J Pers Soc Psychol 105, 171-192.

Pape, H.C., Pare, D., 2010. Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiol Rev 90, 419-463.

Phelps, E.A., O'Connor, K.J., Cunningham, W.A., Funayama, E.S., Gatenby, J.C., Gore, J.C., Banaji, M.R., 2000. Performance on indirect measures of race evaluation predicts amygdala activation. J Cogn Neurosci 12, 729-738.

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. Nat Neurosci 6, 1323-1328.

Richeson, J.A., Todd, A.R., Trawalter, S., Baird, A.A., 2008. Eye-gaze direction modulates race-related amygdala activity. Group Processes & Intergroup Relations 11, 233-246.

Rodrigues, S.M., Schafe, G.E., LeDoux, J.E., 2004. Molecular mechanisms underlying emotional learning and memory in the lateral amygdala. Neuron 44, 75-91.

Ronquillo, J., Denson, T.F., Lickel, B., Lu, Z.L., Nandy, A., Maddox, K.B., 2007. The effects of skin tone on race-related amygdala activity: an fMRI investigation. Soc Cogn Affect Neurosci 2, 39-44.

Sah, P., Faber, E.S., Lopez De Armentia, M., Power, J., 2003. The amygdaloid complex: anatomy and physiology. Physiol Rev 83, 803-834.

Said, C.P., Baron, S.G., Todorov, A., 2009. Nonlinear Amygdala Response to Face Trustworthiness: Contributions of High and Low Spatial Frequency Information. Journal of Cognitive Neuroscience 21, 519-528.

Sapountzis, P., Schluppeck, D., Bowtell, R., Peirce, J.W., 2010. A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. Neuroimage 49, 1632-1640.

Stanley, D.A., Sokol-Hessner, P., Banaji, M.R., Phelps, E.A., 2011. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. Proceedings of the National Academy of Sciences of the United States of America 108, 7710-7715.

Stanley, D.A., Sokol-Hessner, P., Fareri, D.S., Perino, M.T., Delgado, M.R., Banaji, M.R., Phelps, E.A., 2012. Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. Philos Trans R Soc Lond B Biol Sci 367, 744-753.

Telzer, E.H., Flannery, J., Shapiro, M., Humphreys, K.L., Goff, B., Gabard-Durman, L., Gee, D.D., Tottenham, N., 2013a. Early experience shapes amygdala sensitivity to race: an international adoption design. J Neurosci 33, 13484-13488.

Telzer, E.H., Humphreys, K.L., Shapiro, M., Tottenham, N., 2013b. Amygdala sensitivity to race is not present in childhood but emerges over adolescence. J Cogn Neurosci 25, 234-244.

Terbeck, S., Kahane, G., McTavish, S., McCutcheon, R., Hewstone, M., Savulescu, J., Chesterman, L.P., Cowen, P.J., Norbury, R., 2015. beta-Adrenoceptor blockade modulates fusiform gyrus activity to black versus white faces. Psychopharmacology (Berl) 232, 2951-2958.

Todorov, A., Engell, A.D., 2008. The role of the amygdala in implicit evaluation of emotionally neutral faces. Social Cognitive and Affective Neuroscience 3, 303-312.

Van Bavel, J.J., Packer, D.J., Cunningham, W.A., 2008. The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. Psychol Sci 19, 1131-1139.

Wheeler, M.E., Fiske, S.T., 2005. Controlling racial prejudice: social-cognitive goals affect amygdala and stereotype activation. Psychol Sci 16, 56-63.

Winston, J.S., Strange, B.A., O'Doherty, J., Dolan, R.J., 2002. Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nat Neurosci 5, 277-283.

Yarkoni, T., 2009. Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). Perspectives on Psychological Science 4, 294-298.

Table 1. Correlations across behavioral measures

Measure	1	2	3	4	5	6	7

1. Japan SC-IAT	1						
2. South Korea SC-IAT	0.27*	1					
3. Disparity in SC-IAT scores (Japan - South Korea)	N/A	N/A	1				
4. Japan Semantic differential	0.12	0.12	0.00	1			
5. South Korea Semantic differential	-0.06	-0.02	-0.03	0.26*	1		
6. Disparity in semantic differential ratings (Japan - South Korea)	0.14	0.12	0.02	N/A	N/A	1	
7. Trust Game score†	0.02	-0.07	0.07	0.03	-0.08	0.10	1

SC-IAT: Single-Category Implicit Association Test. Note that the Trust Game score is the difference between the average amount offered to Japanese partners minus South Korean partners so that higher numbers indicate that they are more trusting toward Japanese relative to South Korean individuals. N/A: The four correlations are omitted from the table because they are correlations between two non-independent variables (and not surprisingly, they are all highly correlated at $p < 0.001$ level). * $p < 0.05$. † One outlier (more than 3 SD from the mean) was excluded when analyzing the Trust Game data.

Table 2: Decoding performance in the amygdala

ROI name	The number of voxels	Decoding performance (r):			
		South Korea		Japan	
		Implicit attitude	Explicit attitude	Implicit attitude	Explicit attitude
L amygdala	54	0.31* (0.09, 0.51)	-0.13 (-0.37, 0.11)	-0.03 (-0.27, 0.20)	-0.24 (-0.45, -0.00)
R amygdala	63	-0.04 (-0.27, 0.20)	0.11 (-0.12, 0.34)	0.07 (-0.17, 0.30)	-0.30 (-0.50, -0.07)

Implicit attitudes were measured by SC-IAT, while explicit attitude was measured by the semantic differential. Amygdala masks were taken from the Anatomical Automatic Labeling (AAL) masks implemented in the WFU pickatlas toolbox. Voxel size = 3 × 3 × 3 mm. * $p_{perm} < 0.05$ (p-value based on permutation test [5,000 times]). Numbers in parentheses are 95% confidence interval.

Table 3: Decoding of implicit and explicit attitudes toward each of South Korea and Japan in each of a total of 79 regions outside of the prejudice network.

	Name of mask in the WFU pickatlas toolbox	The number of voxels	Decoding performance (r):			
			South Korea		Japan	
			Implicit attitude	Explicit attitude	Implicit attitude	Explicit attitude
Frontal lobe	*vmPFC*	364	-0.46	0.02	0.08	-0.35
	mPFC	1683	-0.17	-0.25	0.01	0.03
	Rectus	314	-0.20	0.16	0.10	0.04
	Frontal_Mid_L	1277	0.30*	0.05	-0.03	-0.04
	Frontal_Mid_R	1379	-0.03	-0.27	0.09	-0.07
	Frontal_Inf_Oper_L	274	0.09	0.19	-0.15	0.11
	Frontal_Inf_Oper_R	367	-0.01	-0.14	0.22	-0.06
	Frontal_Inf_Orb_L	358	0.19	-0.01	0.09	0.14
	Frontal_Inf_Orb_R	351	-0.05	-0.26	-0.04	0.12
	Frontal_Inf_Tri_L	608	-0.19	0.17	0.03	0.01
	Frontal_Inf_Tri_R	475	-0.20	0.19	-0.07	0.02
	Frontal_Sup_L	930	0.08	0.07	-0.21	0.21
	Frontal_Sup_R	1067	0.15	-0.08	-0.01	0.28*
	Frontal_Sup_Orb_L	58	0.21	-0.22	0.08	-0.11
	Frontal_Sup_Orb_R	51	0.00	0.20	0.02	-0.14
	Olfactory_L	42	0.16	0.01	0.05	-0.29
	Olfactory_R	49	0.08	0.13	0.21	0.14
	Precentral_L	874	-0.26	0.01	-0.23	0.32*
	Precentral_R	845	-0.05	-0.17	-0.15	-0.11
	Supp_Motor_Area	1617	-0.05	-0.28	-0.14	0.10
Parietal lobe	Angular_L	329	0.06	0.09	0.02	0.11
	Angular_R	422	0.06	0.15	-0.12	0.48*
	Parietal_Inf_L	669	-0.08	-0.06	0.11	0.21
	Parietal_Inf_R	384	-0.04	-0.41	0.04	0.14
	Parietal_Sup_L	499	0.02	-0.15	-0.09	0.25
	Parietal_Sup_R	426	0.01	-0.19	-0.20	0.06
	Postcentral_L	979	-0.16	-0.01	0.03	0.06
	Postcentral_R	898	0.08	0.09	-0.22	0.00
	Rolandic_Oper_L	271	0.00	0.09	-0.22	0.41*
	Rolandic_Oper_R	358	0.23	0.15	0.24	0.00
	SupraMarginal_L	317	-0.13	0.14	0.26	0.15
	SupraMarginal_R	471	-0.09	0.23	-0.12	0.09
	Paracentral_Lobule	726	0.00	0.06	0.01	0.03
	Precuneus	2291	0.07	0.03	-0.14	-0.06
Temporal lobe	Fusiform_L	445	0.12	-0.28	0.27	-0.08
	Fusiform_R	424	0.31*	-0.21	-0.14	0.01

	Heschl_L	69	-0.28	0.02	0.02	0.17
	Heschl_R	70	0.02	-0.12	0.15	0.10
	Temporal_Inf_L	596	-0.02	-0.01	0.21	0.02
	Temporal_Inf_R	414	0.14	-0.20	-0.02	-0.24
	Temporal_Mid_L	1121	0.00	-0.16	-0.05	0.06
	Temporal_Mid_R	964	-0.34	0.13	-0.32	0.11
	Temporal_Pole_Mid_L	89	-0.05	0.08	-0.18	0.01
	Temporal_Pole_Mid_R	138	-0.09	0.10	0.13	-0.15
	Temporal_Pole_Sup_L	197	0.08	-0.14	-0.24	0.12
	Temporal_Pole_Sup_R	160	0.12	0.12	0.08	0.14
	Temporal_Sup_L	560	-0.12	0.01	-0.05	0.19
	Temporal_Sup_R	730	-0.03	-0.29	-0.07	0.19
Occipital lobe	Calcarine_L	494	0.08	-0.07	0.18	-0.16
	Calcarine_R	504	-0.04	0.26	-0.21	-0.16
	Cuneus_L	365	0.14	-0.24	0.09	0.04
	Cuneus_R	402	0.07	-0.16	-0.46	-0.09
	Lingual_L	509	-0.12	0.07	-0.01	-0.22
	Lingual_R	453	0.30*	-0.11	-0.23	-0.05
	Occipital_Inf_L	211	0.15	-0.21	-0.07	-0.23
	Occipital_Inf_R	204	-0.09	0.05	0.01	-0.08
	Occipital_Mid_L	916	0.15	-0.11	-0.07	0.13
	Occipital_Mid_R	498	0.06	-0.33	0.17	0.10
	Occipital_Sup_L	372	-0.02	-0.08	0.10	0.18
	Occipital_Sup_R	359	0.25	-0.05	-0.28	0.15
Insular lobe	*Ant_Insula_L*	391	0.02	-0.06	-0.01	0.12
	Ant_Insula_R	353	0.28*	-0.03	0.05	-0.03
	Post_Insula_L	210	-0.47	-0.02	-0.17	0.26*
	Post_Insula_R	212	0.00	0.03	-0.07	0.04
Limbic lobe /Subcortical structures	*L Caudate nucleus*	270	0.26	0.23	-0.15	0.06
	R Caudate nucleus	283	0.08	0.13	-0.03	0.15
	Cingulum_Ant	1151	-0.15	0.08	0.16	-0.05
	Cingulum_Mid	1540	-0.01	-0.22	0.07	-0.05
	Cingulum_Post	309	-0.12	-0.24	0.06	-0.17
	Hippocampus_L	240	0.09	0.14	-0.05	-0.01
	Hippocampus_R	251	0.02	0.10	0.08	-0.05
	ParaHippocampal_L	190	-0.12	-0.11	-0.04	-0.02
	ParaHippocampal_R	248	0.14	-0.19	0.13	0.10
	Pallidum_L	67	-0.04	-0.05	-0.16	-0.04
	Pallidum_R	64	0.07	-0.14	-0.02	0.06

Putamen_L	317	-0.16	-0.01	-0.05	-0.05
Putamen_R	321	-0.01	0.02	-0.15	-0.04
Thalamus_L	313	0.13	-0.18	-0.05	-0.29
Thalamus_R	291	0.03	0.23	-0.01	-0.13

Implicit attitudes were measured by SC-IAT, while explicit attitude was measured by the semantic differential. All masks were taken from the Anatomical Automatic Labeling (AAL) masks implemented in the WFU pickatlas toolbox. For each of the midline regions (e.g., vmPFC, mPFC, Rectus, Supp_Motor_Area, Paracentral_Lobule, Precuneus, etc.), mask images in both hemispheres are combined to create a single mask image. Voxel size = 3 × 3 × 3 mm. * $p < 0.05$ (based on permutation test [5,000 times], uncorrected for multiple comparisons). *Italics* indicates regions included in the prejudice network reported previously (Amodio, 2014). Note that since the anterior part of the insula, rather than its posterior part, has been more frequently implicated in prejudice (Amodio, 2014), we separated each insula mask into anterior (y coordinate ≥ 0) and posterior (y coordinate < 0) parts.