Research Article

Psychological Science
24(2) 160–166
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797612451465
http://pss.sagepub.com

Implicit Race Bias Decreases the Similarity of Neural Representations of Black and White Faces

Tobias Brosch[1,2], Eyal Bar-David[1], and Elizabeth A. Phelps[1,3]
[1]Department of Psychology, New York University; [2]Department of Psychology, University of Geneva; and
[3]Nathan Kline Institute, Orangeburg, New York

Abstract

Implicit race bias has been shown to affect decisions and behaviors. It may also change perceptual experience by increasing perceived differences between social groups. We investigated how this phenomenon may be expressed at the neural level by testing whether the distributed blood-oxygenation-level-dependent (BOLD) patterns representing Black and White faces are more dissimilar in participants with higher implicit race bias. We used multivoxel pattern analysis to predict the race of faces participants were viewing. We successfully predicted the race of the faces on the basis of BOLD activation patterns in early occipital visual cortex, occipital face area, and fusiform face area (FFA). Whereas BOLD activation patterns in early visual regions, likely reflecting different perceptual features, allowed successful prediction for all participants, successful prediction on the basis of BOLD activation patterns in FFA, a high-level face-processing region, was restricted to participants with high pro-White bias. These findings suggest that stronger implicit pro-White bias decreases the similarity of neural representations of Black and White faces.

Keywords

social cognition, neuroimaging, face perception, individual differences, stereotyped attitudes

Received 11/15/11; Revision accepted 4/26/12

Understanding the mechanisms that underlie the relations and interactions between members of different social groups is a central topic for social psychology and social neuroscience (Tajfel, 1982). Recently, psychologists, neuroscientists, and lawmakers have begun to explore the extent to which scientific progress can contribute to decision making in the legal system (Gazzaniga, 2008; Lane, Kang, & Banaji, 2007). In this context, the role of implicit social bias—and in particular, race bias—has been the focus of much interest and research activity, as implicit bias is not accessible by introspection, but may nevertheless have significant effects on behavior. For example, people with higher implicit pro-White bias make economic decisions that are more disadvantageous to Black people (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011), prescribe fewer medical treatments to Black people seeking health care (Green et al., 2007), and have less friendly social interactions with Black people (McConnell & Leibold, 2001).

Social stereotypes may not only affect decisions and behaviors, but may even change perceptual experience by increasing perceived differences between social groups (Krueger, 1992). Here, we report a functional MRI (fMRI) study in which we investigated if implicit race bias influences the way the brain represents perceptual information about social group. Previous work has shown that mental templates of out-group faces possess less trustworthy features in people with stronger implicit negative out-group bias (Dotsch, Wigboldus, Langner, & van Knippenberg, 2008). Similarly, the presentation of racial labels changes how participants draw a copy of a face (Eberhardt, Dasgupta, & Banaszynski, 2003). We investigated how this phenomenon is represented at the neural level. Perceptual stimulus information is represented by patterns of distributed neural activation in the ventral visual path (see, e.g., Haxby et al., 2001). It has been shown that in object-specific anterior lateral occipital complex, the degree of similarity between two blood-oxygenation-level-dependent (BOLD) activation patterns directly reflects how similar the observer's perceptual experience of the two stimuli is (Haushofer, Livingstone, & Kanwisher, 2008). We therefore tested the hypothesis that the distributed BOLD patterns representing Black and White

Corresponding Authors:
Tobias Brosch, Department of Psychology, University of Geneva,
40, Boulevard du Pont d'Arve, CH-1205 Geneva, Switzerland
E-mail: tobias.brosch@unige.ch

Elizabeth A. Phelps, Department of Psychology, New York University,
6 Washington Place, New York, NY 10003
E-mail: liz.phelps@nyu.edu

faces are more dissimilar in participants with higher implicit race bias.

We presented pictures of Black and White faces to our participants and used multivoxel pattern analysis (MVPA) to predict the race of the face that a participant perceived at a given moment. In this type of analysis, BOLD activation patterns recorded during the presentation of different types of stimuli are entered into a machine-learning algorithm. If the patterns representing different types of stimuli are sufficiently dissimilar from one another, the algorithm can successfully learn the patterns associated with the stimuli and predict—for new sets of BOLD data that have not been seen by the algorithm before—which stimulus the subject is perceiving at a given moment. Conversely, if a cognitive state can be successfully predicted using MVPA in a particular brain region, one can infer that the cognitive state is represented by a distinguishable pattern of BOLD activation in this brain region (Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006). In this study, we used BOLD data to predict the race of perceived faces. If implicit bias changes the way people perceive others by increasing the differences between social groups, one would expect participants with high implicit race bias to have more dissimilar neural representations of Black and White faces, and thus show better MVPA race prediction, than participants with low levels of implicit race bias.

Face perception is mediated by a distributed neural system. Core brain regions of face processing are located in inferior occipital gyrus (occipital face area, OFA; Puce, Allison, Asgari, Gore, & McCarthy, 1996) and fusiform gyrus (fusiform face area, FFA; Kanwisher, McDermott, & Chun, 1997). Incoming visual information is first encoded structurally, on the basis of the immediate perceptual input, and then transformed into a more abstract, perspective-independent model of the face that can be compared with other faces in memory (Bruce & Young, 1986). The structural-encoding phase has been linked to computations in OFA, whereas the more high-level, identity-based encoding occurs in FFA (Rotshtein, Henson, Treves, Driver, & Dolan, 2005). Other regions involved in face processing are superior temporal sulcus (STS), amygdala, and insula; regions of the reward circuitry, such as caudate and orbitofrontal cortex (OFC); and inferior frontal gyrus (IFG; Haxby, Hoffman, & Gobbini, 2000; Ishai, 2008). We first identified each of these regions in each participant with an independent functional face localizer.

We then tested whether the information contained in the BOLD activation patterns in each region allowed prediction of the race of the face that the participant was perceiving at a given time. This step identified brain regions that represent race-related information. In a second step, we investigated the extent to which individual differences in implicit race bias, as measured by the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), are reflected in race prediction. We tested the hypothesis that MVPA race prediction is better in participants with high implicit race bias—which potentially leads to more dissimilar neural representations of Black and White faces—than in participants with low bias.

Method

Subjects

We recruited 19 right-handed normal volunteers (13 males, 6 females) between 18 and 34 years of age (*M* = 23.6, *SD* = 4.5). Subjects were not selected on the basis of ethnicity (13 Caucasian, 3 Asian, 1 Middle Eastern, 1 Hispanic, 1 African American). To show that our findings were not driven by differences between Caucasians and non-Caucasians, we focus our report on the subgroup of 13 Caucasians (9 males, 4 females; mean age = 23.4 years, *SD* = 4.6). However, we were primarily interested in the relation between implicit bias and BOLD response patterns, so we also conducted and report some analyses including the non-Caucasian participants; including them did not significantly change the pattern of results. The experiment was approved by the New York University Committee on Activities Involving Human Subjects. All subjects gave informed consent and were paid for their participation.

Procedure

During the main task, participants viewed a series of pictures of Black and White faces from the Eberhardt Lab Face Database (Mind, Culture, & Society Laboratory at Stanford University, http://www.stanford.edu/group/mcslab/cgi-bin/wordpress/examine-the-research/). Participants performed a 1-back task, pressing a key whenever the same face was shown twice in a row (which happened in about 5% of the trials). They performed six runs of the main experiment, each run lasting 5 min 20 s and consisting of ten 16-s stimulus epochs interleaved with ten 16-s epochs of fixation. During each stimulus epoch, 20 different pictures (all of the same race) were presented foveally at a rate of 1 every 800 ms (stimulus duration = 550 ms; interstimulus interval = 250 ms). Stimulus epochs randomly alternated between the two different stimulus conditions (Black or White faces).

Participants also underwent an 8-min face localizer task that allowed us to independently identify the regions of the face-processing network. The scan for this task consisted of 15 stimulus epochs interleaved with fixation epochs. The timing parameters were identical to those of the main task, but stimulus epochs alternated between faces (mixed Black and White faces with identities that were not shown during the main experiment), scenes, and objects. Again, participants performed a 1-back task.

At the end of the experiment, participants performed a race IAT following the standard procedure described in Lane, Banaji, Nosek, and Greenwald (2007). The IAT measures the degree to which social groups are automatically associated with positive and negative concepts. Subjects categorize faces as "Black" or "White" while simultaneously categorizing words as "pleasant" (*great*, *fantastic*) or "unpleasant" (*terrible*, *awful*). Strength of association is measured by comparing the response speed under two different sorting conditions: blocks in which items representing "White" and "pleasant" share the same response and blocks in which items representing “Black” and “pleasant” share the same response. Faster responses in the former condition than in the latter condition indicate stronger associative links between “White” and “pleasant” (stronger pro-White bias). Participants’ implicit race bias was measured using the IAT *D* score, which was calculated using the algorithm described in Lane, Banaji, et al. (2007): the difference in average response latency between the two blocks divided by the standard deviation of all latencies. The IAT *D* score can be interpreted as the effect size for an individual’s implicit pro-White bias.

Participants also filled in a series of explicit measures of race attitude, including the Modern Racism Scale (McConahay, 1986), the Symbolic Racism Scale (Henry & Sears, 2002), and the Internal/External Motivation to Avoid Prejudice Surveys (Plant & Devine, 1998), as well as a measure of political leaning (Jost, Glaser, Kruglanski, & Sulloway, 2003).

fMRI acquisition and analysis

A 3-T Siemens Allegra head-only scanner and Siemens standard head coil were used for data acquisition. Anatomical images were acquired using a T1-weighted protocol (256 × 256 matrix, 176 1-mm sagittal slices). Functional images were acquired using a single-shot gradient-echo echo-planar-imaging (EPI) sequence (repetition time = 2.0 s, echo time = 25 ms, field of view = 192 cm, flip angle = 75°). We obtained 39 contiguous oblique-axial slices (3- × 3- × 3-mm voxels) parallel to the anterior commissure–posterior commissure line. Data were preprocessed using SPM8 (Wellcome Trust Center for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm/). All images were realigned, corrected for slice timing, normalized to an EPI template (resampled voxel size of 3 mm), spatially smoothed (8-mm full-width/half-maximum Gaussian kernel), and high-pass-filtered (cutoff = 120 s).

Univariate analysis

Experimental epochs were modeled by a standard synthetic hemodynamic response function (HRF). For the main experiment, two conditions were defined: White faces and Black faces. For the face localizer experiment, three conditions were defined: faces, objects, and scenes. To account for residual movement artifacts after realignment, we entered movement parameters derived from realignment corrections (three translations, three rotations) as additional covariates of no interest. The generalized linear model was then used to generate parameter estimates of activity at each voxel, for each condition and each participant. Statistical parametric maps were generated from linear contrasts between the HRF parameter estimates for the different conditions of interest. We performed random-effects group analyses on the contrast images from the individual analyses, using one-sample *t* tests. Using the functional localizer data, we identified face-sensitive regions (OFA, FFA, amygdala, caudate, STS, insula, IFG, and OFC), as well as an early visual region at the occipital pole (OP) for each individual subject (i.e., faces > others contrast to localize the face network and all stimuli > fixation contrast to localize OP). To define regions of interest (ROIs) for the MVPA analysis, we centered a sphere with a 20-mm radius (covering 1,237 voxels) on the peak coordinates extracted for each face-sensitive region. This procedure resulted in ROIs of equal size (and thus equal numbers of voxels fed into the classifier) for each participant and each brain region (see also Kaul, Rees, & Ishai, 2011).

MVPA

Unsmoothed fMRI data from the six experimental runs were analyzed using the MATLAB routines provided in the Princeton MVPA Toolbox (www.csbmb.princeton.edu/mvpa). The time series from each voxel was de-trended and *z*-scored. Condition onsets were adjusted for the lag in HRF by shifting all block-onset timings by three volumes (6 s). To determine decoding accuracy, we classified activation patterns in each volume using a leave-one-out cross-validation method (Mur, Bandettini, & Kriegeskorte, 2009), training the classifier on five runs and testing it on the sixth run, thus taking into account only classification performance for data that had not been used to train the classifier. We used the sparse logistic regression algorithm (Yamashita, Sato, Yoshioka, Tong, & Kamitani, 2008) for classification. Decoding accuracy was averaged across the six cross-validations for each ROI in each participant. We tested for significant differences from chance performance (50% correct) in all ROIs using Bonferroni-corrected one-sample *t* tests.

Results

Behavioral results

To assess behavioral responses during the main fMRI experiment, we compared participants’ response times and accuracy during the processing of White versus Black faces using paired *t* tests. Responses to White faces (*M* = 532 ms) were faster than responses to Black faces (*M* = 547 ms), *t*(12) = 2.76, *p* = .02. Response accuracy was high and did not differ between White (87% correct) and Black (83% correct) faces, *t*(12) = 1.6, n.s. Participants’ mean IAT *D* score was 0.74 (*SD* = 0.15), and individual *D* scores ranged from 0.55 to 0.99. Thus, all our participants showed an implicit pro-White bias, but there was considerable between-subjects variability in the extent of the bias (for the complete sample, the mean IAT *D* score was 0.68, *SD* = 0.18, and individual *D* scores ranged from 0.19 to 0.99).

MVPA classification results

Above-chance correct classification performance (see Fig. 1) was observed in OP (*M* = 60.0%), *t*(12) = 9.37, *p* < .001; OFA (*M* = 54.7%), *t*(18) = 6.04, *p* < .001; and FFA (*M* = 52.7%), *t*(12) = 3.20, *p* = .005. The same pattern of results was obtained for the complete sample—OP: *M* = 59.1%, *t*(18) = 6.04,

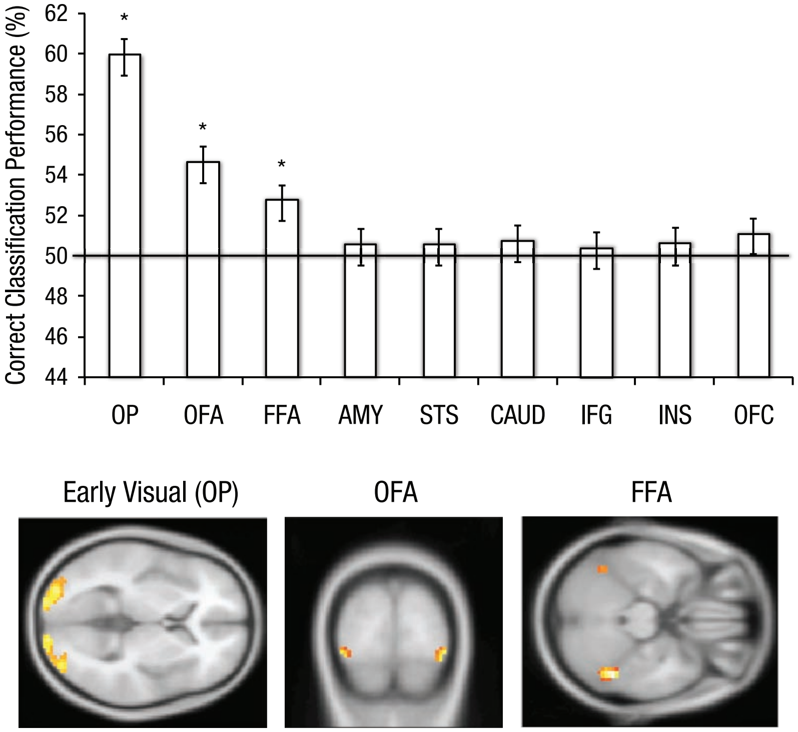

Fig. 1. Classification performance for Black and White faces in the regions of interest. Error bars represent ±1 *SE*. The brain images illustrate the location of the three regions in which classification was significantly better than chance ($p < .05$). OP = occipital pole; OFA = occipital face area; FFA = fusiform face area; AMY = amygdala; STS = superior temporal sulcus; CAUD = caudate; IFG = inferior frontal gyrus; INS = insula; OFC = orbitofrontal cortex.

$p < .001$; OFA: $M = 54.0\%$, $t(18) = 6.04$, $p < .001$; FFA: $M = 52.3\%$, $t(18) = 3.20$, $p = .005$. No other ROI showed statistically significant above-chance classification (amygdala: 50.3%; STS: 50.6%; caudate: 50.7%; IFG: 50.4%; insula: 50.6%; OFC: 51.1%; all *t*s < 1.5, all *p*s > .17).

Uni- and multivariate control analyses

To exclude the possibility that the above-chance classification performance in OP, OFA, or FFA was driven by an overfitting of arbitrary patterns of spatial correlations in the data, we carried out a shuffle-control test (Mur et al., 2009), in which the category labels during training were reshuffled for each round of the cross-validation. Classification performance during the shuffle-control test was not different from chance (OP: 50.0%; OFA: 49.7%; FFA: 50.0%; all *t*s < 1, all *p*s > .70). We also analyzed our data for differences in the univariate mean BOLD signal, but no ROI showed significantly different mean BOLD signals for Black compared with White faces (all *t*s < 1, all *p*s > .37). To further ensure that classification was not driven by mean differences in the ROIs instead of distributed patterns, we ran an additional MVPA analysis using a supervoxel (average of all the voxels within an ROI that showed successful classification). None of the regions showed classification performance different from chance in this analysis (OP: 48.5%; OFA: 49.2%; FFA: 50.0%; all *t*s < 1.4, all *p*s > .19).

MVPA classification and implicit bias

To test whether individual differences in classification performance reflect implicit bias, we then used the classification performance in ROIs that allowed for successful classification to predict participants' IAT scores. Linear regression analysis revealed that classification performance in FFA significantly predicted IAT scores, $\beta = 0.61$, $t(12) = 2.59$, $p = .025$, with higher classification performance predicting higher IAT scores (see Fig. 2a). The same pattern of results was obtained in the complete sample, $\beta = 0.46$, $t(18) = 2.16$, $p = .046$. Classification performance in no other ROI predicted IAT scores (all *p*s > .30), and none of the explicit-bias measures were related to classification performance (all *p*s > .20).

To further explore the link between classification performance and pro-White implicit bias, we divided our participants via median split into a high-bias group (mean IAT *D* score = 0.87) and a low-bias group (mean IAT *D* score = 0.61) and separately calculated classification scores for the two groups (see Fig. 2b). Both the high- and the low-bias groups showed above-chance classification performance in OP (high-bias group: 59.6%, $p = .008$; low-bias group: 60.7%, $p < .001$) and OFA (high-bias group: 53.4, $p = .015$; low-bias group: 56.0%, $p = .009$). Classification performance in OP or OFA was not significantly different between the groups. In FFA, however, the high-bias group showed above-chance classification (54.7%,

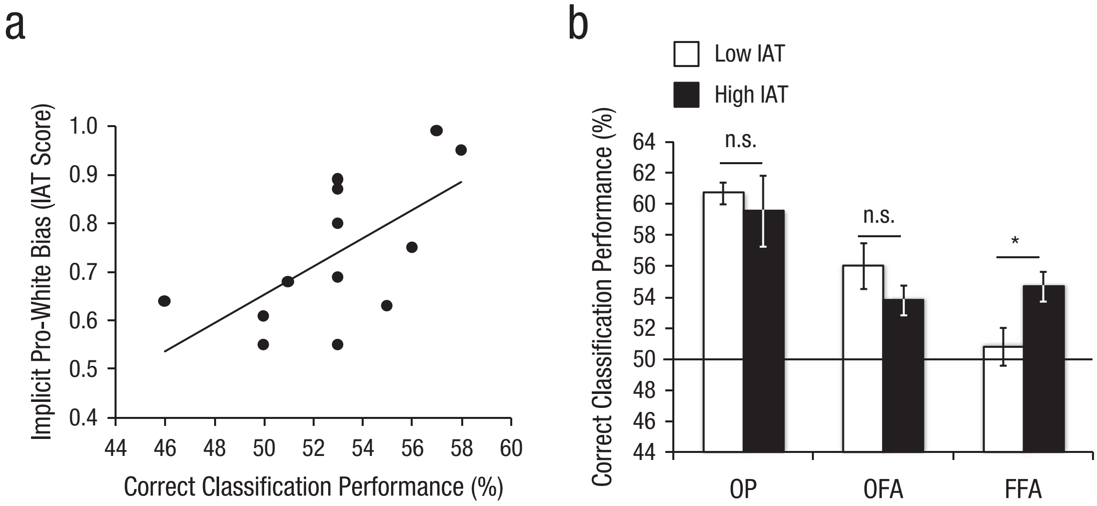

Fig. 2. Link between implicit pro-White bias and ability of the classifier to predict the race of faces participants were viewing. The scatter plot (a) presents implicit bias as a function of correct classification performance in fusiform face area (FFA); the best-fitting regression line is also shown, $\beta = 0.61$, $p = .025$. The bar graph (b) shows correct classification performance in occipital pole (OP), occipital face area (OFA), and FFA separately for participants with low and high scores on the Implicit Association Test (IAT), as determined by a median split. Error bars indicate ±1 *SE*. The asterisk indicates a significant difference between groups ($p < .05$).

$p = .004$), but the low-bias group did not (50.8%, n.s.). When we directly compared FFA classification performance of the two groups, performance was significantly better in the high- than the low-bias group ($p = .032$).

The same pattern of results was observed for the complete sample. In the high-bias group (mean IAT *D* score = 0.81), classification performance was above chance in OP (59.7%), OFA (53.8%), and FFA (54.6%), all *p*s < .003. In the low-bias group (mean IAT *D* score = 0.55), classification performance was above chance in OP (58.4%) and OFA (54.5%), both *p*s < .008, but not in FFA (50.7%), n.s. FFA classification performance was significantly better for the high- than the low-bias group ($p = .006$).

Previous research had revealed a correlation between IAT score and average amygdala BOLD signal during the presentation of Black versus White faces (Cunningham et al., 2004; Phelps et al., 2000). Given these findings, we extracted average BOLD signal from the amygdala (Black faces > White faces contrast, 6-mm spheres) and computed correlations with individual IAT scores. Replicating previous work, we observed a positive correlation between IAT score and average BOLD signal in the right amygdala ($r = .62$, $p = .025$). However, the amygdala ROI did not allow successful classification using MVPA, nor was classification performance in this region associated with IAT scores ($p = .47$).

Discussion

We investigated whether implicit race bias influences the way the brain represents perceptual information about social group. To do this, we used MVPA to predict the race of perceived faces, testing the hypothesis that race prediction is better in individuals with high implicit race bias than in participants with low bias. We successfully predicted the race of a face on the basis of BOLD activation patterns in OP, OFA, and FFA. Race prediction was thus possible both in relatively early visual regions sensitive to low-level physical cues (OP, OFA) and in a higher-level face-processing region (FFA). It is important to note that Black and White faces differ structurally on low-level stimulus features, such as color. As a consequence, successful race prediction based on simple perceptual features is to be expected in early visual regions. Indeed, in this study, successful race prediction was observed in OP and OFA in both low-bias and high-bias participants. In contrast, in FFA, a higher-level face-processing region involved in the encoding of more abstract, identity-based face models (Rotshtein et al., 2005), race prediction was possible only in participants with relatively strong implicit pro-White bias, a result implying that the neural representations of Black and White faces in this region are more different in participants with higher implicit bias. Even though all participants received identical stimulus input, modulating factors may have differentially affected the BOLD patterns representing the information.

As mentioned earlier, implicit racial bias has been shown to powerfully influence perception of a face (Dotsch et al., 2008; Eberhardt et al., 2003). The impact of implicit bias on MVPA race prediction in FFA may be the neural signature of these behaviorally demonstrated biases, suggesting the intriguing possibility that stronger race bias may actually be associated with larger differences in the perceptual experience of Black and White faces (Haushofer et al., 2008).

A recent article provided evidence that race as perceived in faces can be decoded from widely distributed patterns of brain activity in occipital and temporal cortices (Natu, Raboy, &

O'Toole, 2011). In the experiment reported here, we clarified the role of independently identified regions of the face-processing network and investigated how individual differences in implicit race bias are related to neural patterns in these regions. Previous brain-imaging studies on the neural signatures of race and race bias (Cunningham et al., 2004; Golby, Gabrieli, Chiao, & Eberhardt, 2001; Phelps et al., 2000) have relied mainly on univariate fMRI analysis methods, which examine the overall BOLD activation level in a brain region, averaged across multiple voxels and a large number of trials. This technique does not allow prediction of the current cognitive state of the individual using new BOLD data, and thus does not allow the conceptually important inference that a cognitive state is represented by a distinguishable pattern of BOLD activation in a certain brain region.

Furthermore, univariate analysis techniques may be less sensitive than multivariate techniques to race-related information, depending on the nature of the neural representation. For example, in a previous univariate study on memory encoding of Black and White faces (Golby et al., 2001), an increase in average BOLD signal in response to in-group faces was observed in FFA, but not in OP and OFA. The fact that MVPA is able to predict the race of a perceived face on the basis of activation in each of these regions demonstrates the increased sensitivity of this method, which does not require an overall increase in activation levels in response to one stimulus in a given region, but capitalizes on differences in the locally distributed activation patterns. In our study, we did not replicate univariate differences in FFA activation between in-group and out-group stimuli (Golby et al., 2001). However, Golby et al. used an intentional memory-encoding task requiring more effort and deeper encoding than our 1-back task (and indeed, the authors interpreted the differences in FFA activation as reflecting deeper encoding of the in-group faces relative to the out-group faces).

Distributed patterns of BOLD activation in the amygdala did not allow for successful prediction of race of the perceived face or of the observer's race bias, even though previous univariate correlative analyses have linked this region to both race and race bias (Cunningham et al., 2004; Hart et al., 2000; Phelps et al., 2000), and even though we were able to replicate the correlation with IAT scores in our data set. MVPA picks up on distributed activation patterns in cortical regions and may be less successful with small nuclei that are covered by a small number of voxels only. Previous attempts to classify emotional information using MVPA on BOLD patterns in the amygdala have similarly not been successful (Peelen, Atkinson, & Vuilleumier, 2010).

Using MVPA, we were able to predict the race of a perceived face on the basis of activation patterns in several face-processing areas. Race prediction in FFA, a face-processing area involved in the encoding of high-level identity-based face models, was restricted to participants with higher levels of implicit pro-White bias. Our findings suggest that stronger implicit pro-White bias decreases the similarity of neural representations (and potentially the subjective perceptual experience) of Black and White faces. The observation of individual differences in MVPA race prediction suggests possible applications of our findings. For example, it may be possible to predict differences in implicit race bias at the individual level using brain data, which may be of interest given the behavioral and societal implications of race bias (Gazzaniga, 2008; Lane, Kang, & Banaji, 2007).

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This study was funded by a Seaver Foundation grant to the Center for Brain Imaging, a grant from the National Institutes of Health (5R01MH062104) to E. A. P., and a grant from the Swiss National Science Foundation (PA00P1_131435) to T. B.

References

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of Black and White faces. *Psychological Science*, *15*, 806–813.

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*, 978–980.

Eberhardt, J. L., Dasgupta, N., & Banaszynski, T. L. (2003). Believing is seeing: The effects of racial labels and implicit beliefs on face perception. *Personality and Social Psychology Bulletin*, *29*, 360–370.

Gazzaniga, M. S. (2008). The law and neuroscience. *Neuron*, *60*, 412–415.

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*, 845–850.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, *22*, 1231–1238.

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

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*, 2351–2354.

Haushofer, J., Livingstone, M. S., & Kanwisher, N. (2008). Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biology*, *6*, e187. Retrieved from http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.0060187

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*, 223–233.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–534.

Henry, P. J., & Sears, D. O. (2002). The Symbolic Racism 2000 Scale. *Political Psychology*, *23*, 253–283.

Ishai, A. (2008). Let's face it: It's a cortical network. *NeuroImage*, *40*, 415–419.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*, 339–375.

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

Kaul, C., Rees, G., & Ishai, A. (2011). The gender of face stimuli is represented in multiple regions in the human brain. *Frontiers in Human Neuroscience*, *4*, 238. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3026581/

Krueger, J. (1992). On the overestimation of between-group differences. *European Review of Social Psychology*, *3*, 31–56.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: What we know (so far) about the method. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 29–102). New York, NY: Guilford Press.

Lane, K. A., Kang, J., & Banaji, M. R. (2007). Implicit social cognition and law. *Annual Review of Law and Social Science*, *3*, 427–451.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale: Prejudice, discrimination, and racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). San Diego, CA: Academic Press.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*, 435–442.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*, 101–109.

Natu, V., Raboy, D., & O'Toole, A. J. (2011). Neural correlates of own- and other-race face perception: Spatial and temporal response differences. *NeuroImage*, *54*, 2547–2555.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience*, *30*, 10127–10134.

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. *Journal of Cognitive Neuroscience*, *12*, 729–738.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letter-strings, and textures: A functional magnetic resonance imaging study. *Journal of Neuroscience*, *16*, 5205–5215.

Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, *8*, 107–113.

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, USA*, *108*, 7710–7715.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*, 1–39.

Yamashita, O., Sato, M. A., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, *42*, 1414–1429.